PROJECT REPORT - EE604A
DISPARATE IMAGE MATCHING USING DUALITY DESCRIPTOR (DUDE)

Akshat Agarwal, Ayush Shakya, Siddharth Tanwar

ABSTRACT

We have implemented a new feature descriptor and a new feature detector for the purpose of establishing image correspondences between highly disparate images. We follow [1] which presents a novel descriptor algorithm, Duality Descriptor (DUDE) using line/point duality and a randomization strategy that provides simple but robust, consistent feature extraction and correspondence. By using the dual representation we are able to effectively capture the line segments, and by the customized detector we are able to generate more repeatable and suitable features between two images. We have used a dataset of disparate images given by Hauagge and Snavely [2], mostly architectural scenes that include symmetric shapes, exhibiting dramatic variations in lighting, time period, modality, etc. The DUDE descriptor implemented by us gives comparable results to state-of-the-art methods with much less computational expense.

1. INTRODUCTION

The primary objective here is to find the similarities between disparate images and find a match. Currently there are many methods for feature matching like SIFT[3], SURF[4], SYM-G[2] etc. which perform very well for the normal images. However the results are not very good when there is dramatic appearance change such as images from different modalities sensor, age, lighting etc. The motivation behind the project was its scope in different areas of vision including 3D reconstruction, object tracking, object recognition, depth estimation and others.

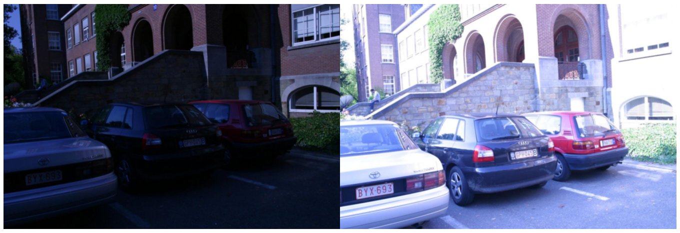


Fig. 1: An example of a disparate image pair taken from [2]

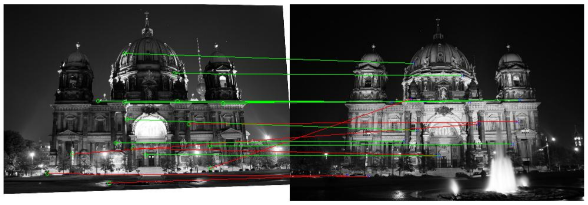


Fig. 2: Results of our implementation of DUDE over a highly disparate image pair. Green and red indicate correct and incorrect matches respectively.

2. RELATED WORKS

2.1. SIFT (Scale-Invariant Feature Transform) [3]

It is used to detect and describe local feature in images wherein the method transforms an image into large collection of feature vectors. Each of these vectors is invariant to image translation, scaling, rotation, illumination changes and local geometric changes.

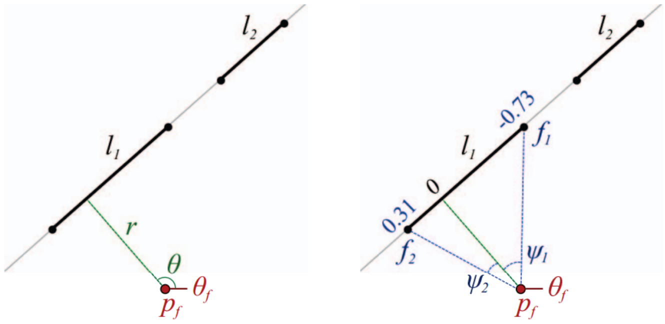
First an internal representation of original space is created to ensure scale invariance. The Laplacian of Gaussian is approximated using the representation created earlier for finding interesting points. These are mainly maxima and minima in Difference of Gaussian image. Edges and low contrast points are eliminated for efficiency and robustness. Orientation is calculated for each point, further calculations are done relative to this orientation which makes them rotation invariant.

2.2. SURF (Speeded Up Robust Features)[4]

To describe each feature point SURF summarizes pixel information within a local neighbourhood. First it determines orientation for each feature by convolving pixels in its neighbourhood with the horizontal and vertical Haar wavelet filters. By using intensity changes to characterise orientation, the descriptor is able to describe features in the same manner regardless of orientation of object or camera.

2.3. MSER (Maximally Stable Extremal Regions)[5]

MSER is a method for blob detection in images,the algorithm extracts a number of co-variant regions from the image, called



(a) r and θ

(b) f

Fig. 3: Dual representation r, , f_1, f_2 of a line segment [1]

MSERs. MSER is a stable connected component of some gray-level sets of the image. It is based on the idea of taking regions which stay nearly the same through a wide range of thresholds, the algorithm is as follows:

First we look for the anchor points which are local intensity extremum points. Then explore the image around, going along every ray starting from anchor point until an extremum of function f is reached. All points create some irregularly shaped region, which are then approximated. As function f is the characteristic function of the region, moments upto 2nd order allow us to approximate the region with an ellipse. Optionally, elliptical frames are attached to the MSERs by fitting ellipses to the regions. These region descriptors are kept as features.

3. DUAL REPRESENTATION

Disparate meaning fundamentally different or distinct in quality or kind are the images taken in different conditions such as picture vs painting, different age, lighting conditions etc. Generally for these kind of application line detectors are used for finding the feature points and defining a descriptor. But line detection is parameter-sensitive, and does not guarantee unique endpoints. Occlusion and appearance change can cause line segments to become disconnected. To overcome this issue we implemented the descriptor proposed as DUDE (DUality DEscriptor) that uses a 3D cylindrical histogram based on a transformation of line segments to a dual space of points. By exploiting line-point duality, DUDE is designed so as to be less affected by line segment disconnection.

The DUDE descriptor takes advantage of dual representation by transforming line segments into points in dual space. We represent each line segment as $[r, \theta, f_1, f_2]$, instead of $[x_1, y_1, x_2, y_2]$, where r and θ are defined by the infinite line containing the segment, and the two f values, calculated from sin(ψ), represent how far each endpoint is from the orthogonal projection.

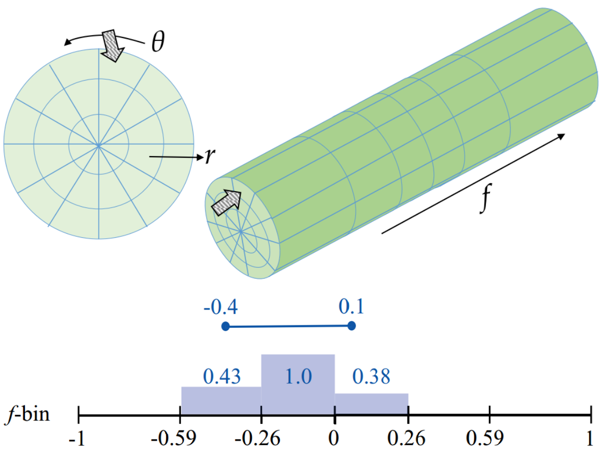


Fig. 4: An example of binning in θ, r and f [1]

4. DUALITY DESCRIPTOR (DUDE)

Normally for image matching for a given set of feature points **F**, a descriptor is assigned to each feature point F_i. Almost in every feature point detection method a feature point is identified by four parameters $[x_i, y_i, s_i, \theta_i]$ representing location, scale and orientation respectively. Our descriptor is designed to work as follows, for each feature point F_i the descriptor goes through following steps:

1. A set of line **S** is extracted such that each line segment S_i is within the circle centered at $[x_i, y_i]$ and of radius qs_i, where q is a parameter for our descriptor
2. We then calculate $[r, \theta, f_1, f_2]$ values as described above for each line segment S_i relative to the F_i with origin at $[x_i, y_i]$ and orientation θ_i
3. For each line segment S_i we create a cylindrical coordinate (r, θ, f) and accumulate them to form a 3D histogram for point F_i
4. We then perform the binning for (r, θ, f) into n_r, n_θ and n_f bins forming a $(n_r \times n_\theta \times n_f)$ dimension descriptor

By the above implementation we addressed the underlying difficulties that line-based approaches must overcome. First, for the cases where a same line segment may be detected as multiple line segments in the counterpart, we decreased the problem as the collinear lines share same r and θ, and their f ranges are accumulated. Second, slight changes in the end points can cause variation in r and θ values, we overcome this by blurring the end points within 3 pixels in x and y.

5. MULTI-MODAL IMAGE DETECTOR (MMID)

The detector is specially designed to generate features across disparate images that are more suited for the DUDE descriptor. Since many feature detectors return feature points in the

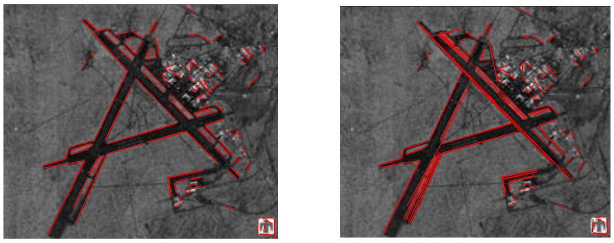


Fig. 5: (left) Initial Line segments from LSD [6] and (right) Proposed Merged Line Segments from MMID[1]

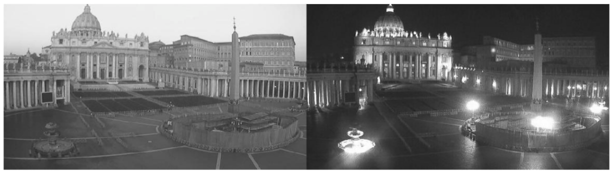


Fig. 6: An example disparate image of the Vatican taken from the dataset [2]

form $[x_i, y_i, s_i, \theta_i]$. MMID describes feature points for each line, (x_i, y_i) as its mid point, scale (s_i) as half of its length and slope of the segment as orientation (θ_i). To increase the consistency it generates multiple groupings of line segments obtained from Line Segment Detector (LSD) [6] by randomly merging them incrementally(inspired by [7]).

The merging is carried out as follows:

1. For given line segment set create a graph with line segments as its nodes and each edge connects two neighbouring line segments with some weight
2. Edge weight w_{ij} between line segments i and j contains three terms: shortest distance $\delta_{1,i,j}$, perpendicular distance $\delta_{2,i,j}$, and angle $\delta_{3,i,j}$ between them $w_{ij} = (1 - \delta_{1,i,j}/\alpha)(1 - \delta_{1,i,j}/\beta)(1 - \delta_{1,i,j}/\gamma)$
3. Collect the edges with weight greater than the predefined threshold $\delta_w(0.5)$, sort them in descending order
4. Finally incrementally merge the lines

6. RESULTS

The performance of the descriptor and detector proposed in the paper were evaluated both separately and along with each other. Multiple combinations of detectors and descriptors have been studied as well. The challenging dataset[2] containing 46 image pairs has been used for this study.

To evaluate the efficacy of the detector, repeatability is used as a measure. For feature sets $F1$ and $F2$, from an image pair, repeatability is the fraction of the number of repeatedly detected features over the total number of features. To avoid the possibility of saturation of repeatability, the metric was computed and compared only for "random-k" detections.

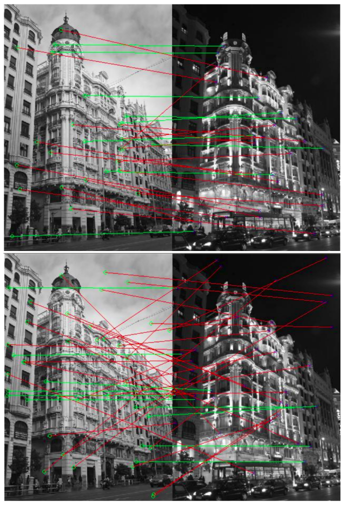


Fig. 7: Matching using DUDE descriptor (top) and SIFT descriptor (bottom)

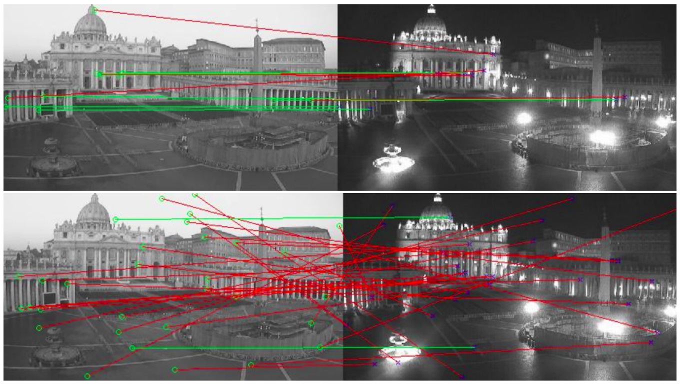


Fig. 8: Matching using DUDE descriptor (top) and SIFT descriptor (bottom)

VLfeat toolbox [8] and the methodology of [9] has been used to calculate this metric. The results were derived for SIFT detector as well to draw comparison. Table 1 shows our results on an image pair (see fig. 6.) where we see that MMID (ours) outperforms SIFT and produces good repeatability.

To evaluate the descriptor, we followed the author and calculated NNDR (Nearest Neighbor Distance Ratio) score [3] for each of the matches paired by descriptor similarity. By varying the threshold on the NNDR score, and identifying which matches are correct (with a known ground truth transformation), mean average precision (mAP) over the range of recall is calculated. Again VLfeat toolbox [8] was used to find the NNDR score. The precision and recall calculations were coded by us as they weren't available in the toolbox.

We did an extensive study of the precision metric and as done by the author, calculated the mAPs for different combinations of detectors and descriptors. Detectors used were SIFT[3], SYM-I[2], SYM-G[2], MMID(ours) and

k	**SIFT**	**MMID**
50	0.0327	0.0979
100	0.0524	0.1264
200	0.11	0.18
300	0.14	0.1825

Table 1: Mean repeatability for fig. 6

Descriptors	**Detectors**				
	SIFT	**SYM-I**	**SYM-G**	**JSPEC**	**MMID**
SIFT	**0.1878**	0.28	0.25	0.25	**0.1**
SYMD	0.22	0.20	0.25	-	0.26
SIFT-SYMD	0.28	0.35	0.36	-	-
DUDE	**0.1121**	**0.1741**	**0.2009**	-	**0.5343**

Table 2: Mean average precision (mAP) for different combinations of detector and descriptors (implementation taken from [8]), and DUDE. The data in bold has been computed by us, rest are taken from [1]

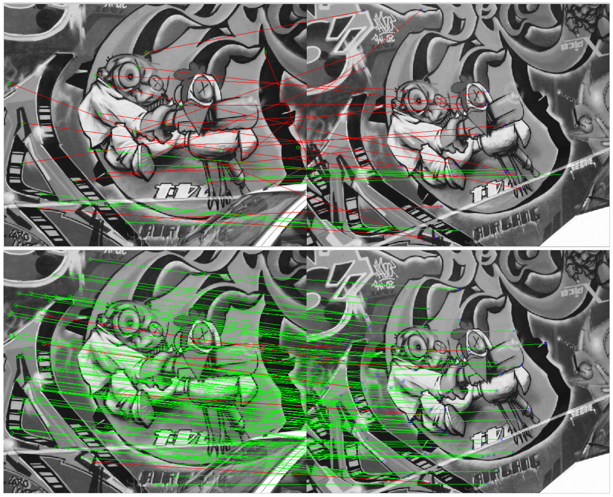


Fig. 9: Matching using DUDE descriptor (top) and SIFT descriptor (bottom)

JSPEC[10]. Descriptors used were SIFT, SYMD, SIFT-SYMD[2] and DUDE (ours). VLfeat toolbox and other open source implementaions of the above descriptors and detectors have been used. Since, we could not find an open implementation of JSPEC, results for the same have been taken from the paper. Table 2 shows our results over an image instance from the dataset (see fig. 6). Our descriptor and detector pair outperforms all other combinations except JSPEC+SIFT which is state-of-the-art. But our algorithm does feature matching in under a minute for the shown image (see fig. 6) whereas JSPEC takes orders of magnitude more time to do the same.

Fig. 7, 8 and 9 show qualitative results of our detection and compare them to SIFT on the same images. As is evident, in highly disparate images, our implementation of DUDE descriptors produces better results with larger number of correct matches. In fig. 7, although SIFT detects more matches most of them are wrong whereas at the same threshold DUDE detects higher number of positive matches. However, it is to be noted that our implementation does poorly in cases when there is no dramatic change in images, where SIFT outperforms our implementation by a huge margin (see fig. 9). This is so because SIFT is able to match more points correctly our of the larger number of detections contrary to earlier where most matches were wrong. Our implementation's matches, on the other hand, do not improve in quantity or precision leading to poor results.

7. CONCLUSION

We implemented a novel feature detection and description system for disparate image matching. The descriptor-detector pair shows promise in disparate image matching wherein it outperforms most famous and common descriptors on the challenging 46 image-pair dataset [2]. The performance that we achieved is close to the state-of-the-art (JSPEC[10]) with significantly more efficient computation. Not only this, the descriptor and detector are general enough to be integrable with other detectors and descriptors respectively. The descriptor successfully captures the uneven and unstably detected distribution of line segments over different modal images and other disparate images in general.

8. REFERENCES

[1] Youngwook P Kwon, Hyojin Kim, Goran Konjevod, and Sara McMains, "Dude (duality descriptor): A robust descriptor for disparate images using line segment duality," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 310–314.

[2] Daniel Cabrini Hauagge and Noah Snavely, "Image matching using local symmetry features," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 206–213.

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

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[5] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.

[6] Rafael Grompone von Gioi, Jérémie Jakubowicz, Jean-Michel Morel, and Gregory Randall, "Lsd: a line segment detector," *Image Processing On Line*, vol. 2, pp. 35–55, 2012.

[7] Hyojin Kim, Jayaraman J Thiagarajan, and Peer-Timo Bremer, "Image segmentation using consensus from hierarchical segmentation ensembles," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 3272–3276.

[8] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[9] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, and Luc Van Gool, "A comparison of affine region detectors," *International journal of computer vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[10] Mayank Bansal and Kostas Daniilidis, "Joint spectral correspondence for disparate image matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2802–2809.