Learning Visual Feature Descriptors for Dynamic Lighting Conditions
(Extended Abstract)

Nicholas Carlevaris-Bianco and Ryan M. Eustice

I. Introduction

Standard hand-designed visual features such as scale invariant feature transform (SIFT) [1] and speeded up robust features (SURF) [2] detect key-points in an image and then describe the local visual appearance of these key-points as a vector. Image registration can then be performed by matching the key-points in multiple images by comparing the $\mathcal{L}_2$ distance between the descriptors. In order for matching to be successful, the key-point detector and descriptors must be at least partially robust to common image variations such as scale, rotation, view-point, and lighting changes. Invariance with respect to scale and rotation are usually accounted for at the feature detection stage, where key-points will be detected at a canonical scale and orientation. The description stage then focuses on representing the appearance of the local region around the key-point such that the descriptor is discriminative while being robust to view-point and illumination changes.

In this abstract, we overview a method to increase the illumination robustness of feature point description to lighting changes. Hand-designed descriptors such as SIFT and SURF have limited lighting invariance—often allowing for affine transformations in image intensity by considering the gradient of intensity, and through other mechanisms such as mean subtraction and normalization. However, in general, the change in appearance caused by lighting affects the image intensity in a complex, nonlinear way.

In many robotic applications, the success or failure of feature-based image registration is largely determined by changes in lighting. This is especially true for medium to long-term outdoor applications, where the scene structure has not changed dramatically, but images separated by even a few hours may be unmatchable due to cyclical changes in lighting. This phenomenon is illustrated in Fig. 1, which shows example imagery from three different locations in our experimental dataset. In this dataset, only a small fraction of the possible matches are successfully registered using standard features, largely because of cyclical changes in lighting.

In this work, we seek to learn a feature descriptor that is more robust to changes in local image appearance caused by lighting. To observe how the local appearance of image patches changes under dynamic lighting conditions, we first track key-points and their associated image patches through time-lapse video using a representative training dataset. We then train a feature descriptor using matching and non-matching pairs of image patches sampled from these patch tracks. A contrastive cost function is used so that matching patches are mapped close together (in terms of Euclidean distance in feature space) while separating non-matching patches. The resulting descriptor is more robust to the types of lighting variation observed in the training data.

Fig. 1: Sample images from 3 of the 500 locations in the North Campus dataset. Imagery was collected in 27 sessions over the course of 15 months with lighting conditions ranging from early morning to just after dusk. The success or failure of feature-based image registration in this dataset is largely determined by the similarity of lighting conditions.

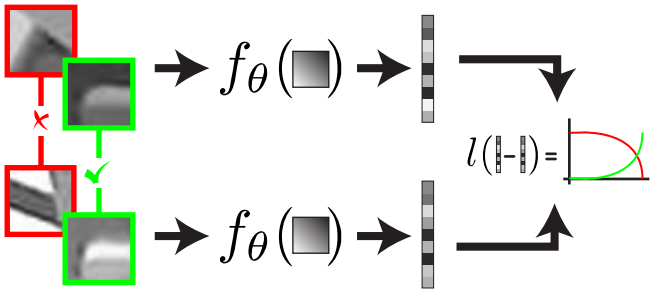

Fig. 2: Illustration of the "Siamese" network training scheme.

*This work was supported in part by the National Science Foundation under award IIS-0746455, the Office of Naval Research under award N00014-12-1-0092, and Ford Motor Company via the Ford-UM Alliance under award N015392.

N. Carlevaris-Bianco is with the Department of Electrical Engineering & Computer Science, University of Michigan, Ann Arbor, MI 48109, USA carlevar@umich.edu.

R. Eustice is with the Department of Naval Architecture & Marine Engineering, University of Michigan, Ann Arbor, MI 48109, USA eustice@umich.edu.

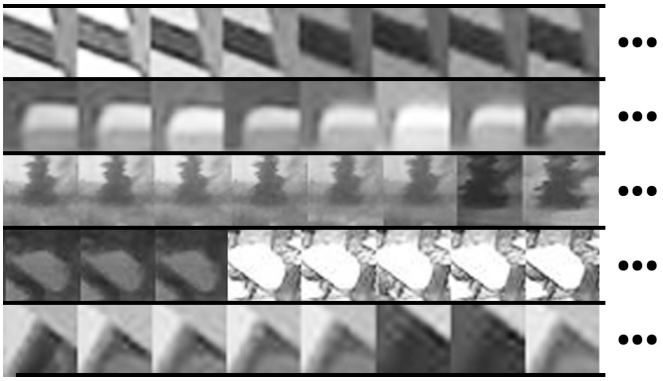

Fig. 3: Sample patch tracks extracted from the webcam dataset. Note that the tracks will be of varying lengths, and only a very small subset is shown here.

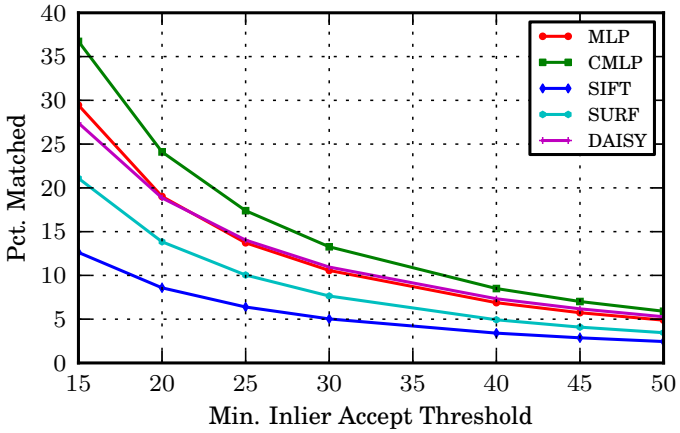

Fig. 4: Average matching results over 500 locations in the North Campus dataset.

II. Learning Visual Feature Descriptors

To learn an illumination robust feature descriptor we employ a training scheme referred to as a "Siamese" network [3–7], with the goal of minimizing a contrastive cost function [4, 5, 7] that encourages a nonlinear mapping to a lower-dimensional space where matching features are close together and non-matching features are far apart in Euclidean distance. This goal is often referred to as embedding learning, manifold learning, or distance metric learning.

The training scheme is illustrated in Fig. 2. Pairs of image patches labeled either as matching or non-matching are supplied as input to a feature descriptor function, $f_{\boldsymbol{\theta}}(\cdot)$, parameterized by $\boldsymbol{\theta}$, that maps the input patch to a feature vector. A contrastive cost function, $l(\cdot)$, based on the Euclidean distance between the feature vectors, encourages matching feature vectors to be close together in feature space while encouraging non-matching features to be far apart. By learning parameters $\boldsymbol{\theta}$ that minimize this cost function, we produce a mapping to a feature space where Euclidean distance captures the similarity and differences amongst the training pairs. By training with data that includes variation due to changes in lighting, the feature descriptor learns to be robust to lighting variation.

In our experiments we consider two standard model classes for the learned feature descriptor; a multi-layer perceptron (MLP) and a convolutional multi-layer perceptron (CMLP) [8]. Batch stochastic gradient descent was employed in order to optimize the model parameters.

III. Generating Training Data

In order to capture the changes in appearance caused as the lighting changes with time, we generated training patches using time-lapse videos. The videos were created by downloading imagery from webcams at fixed locations every 20 minutes over the course of several days. Given a sequence of webcam frames, we track image patches through time by detecting interest points in each frame and then associating interest points between frames. This results in image patches that have been tracked through time as illustrated in Fig. 3.

Given a set of patch tracks, it is easy to generate a very large set of matching and non-matching pairs for training, by sampling matching pairs from within tracks and non-matching pairs from between tracks.

IV. Experimental Results

We evaluate the proposed feature descriptors using imagery collected in 27 sessions over the course of 15 months on University of Michigan's North Campus (Fig. 1). This dataset contains a wide variety of lighting conditions ranging from early morning to just after dusk. Additionally, this data includes viewpoint variance and additional challenges caused by moving objects, seasonal changes, and even construction projects. Given known image robot pose, the dataset is split up into 500 locations with an average of 37 images per location. At each location we match all pairs of images and reject outliers by fitting an Essential matrix using random sample consensus.

In addition to the proposed descriptors, we compare against SIFT [1], SURF [2], and DAISY [9]. For the DAISY descriptor we use learned parameters provided by [10], specifically the "T1-4-2r8s" version as it has an output dimension of 68 and computation time comparable with our learned descriptors.

In Fig. 4 we show the percentage of image pairs successfully matched as a function of the minimum number of inliers (the higher the min. inlier threshold the more confident we are of the matches). Here, we see that the CMLP learned feature provides the best results, matching around 37% of possible pairs at the lowest threshold. The MLP learned feature and DAISY have similar performance, while SIFT and SURF are significantly less successful.

V. Conclusions

We have presented a method to learn visual feature point descriptors that are more robust to changes in scene lighting than standard hand-designed features. Our preliminary experimental results are very promising, showing that the learned features provide better image registration performance on a challenging robotic dataset than standard hand-designed features including SIFT and SURF.

REFERENCES

[1] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[3] J. Bromley, I. Guyon, Y. LeCun, E. Sackinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," *Int. J. Pattern Recog. and Artifical Intell.*, vol. 7, no. 4, pp. 669–688, 1993.

[4] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verication," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, San Diego, CA, Jun. 2005, pp. 539–546.

[5] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, New York, NY, Jun. 2006, pp. 1735–1742.

[6] H. Mobahi, R. Collobert, and J. Weston, "Deep learning from temporal coherence in video," in *Proc. Int. Conf. on Machine Learn.*, Montreal, Canada, Jun. 2009, pp. 737–744.

[7] G. W. Taylor, I. Spiro, C. Bregler, and R. Fergus, "Learning invariance through imitation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Providence, RI, Jun. 2011, pp. 2729–2736.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[9] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.

[10] S. Winder, G. Hua, and M. Brown, "Picking the best DAISY," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Miami, FL, Jun. 2009, pp. 178–185.