TILDE: A Temporally Invariant Learned DEtector

Yannick Verdie[1]* Kwang Moo Yi[1]* Pascal Fua[1] Vincent Lepetit[2]
[1]Computer Vision Laboratory, École Polytechnique Fédérale de Lausanne (EPFL)
[2]Institute for Computer Graphics and Vision, Graz University of Technology

We introduce a learning-based approach to detect repeatable keypoints under drastic imaging changes of weather and lighting conditions to which state-of-the-art keypoint detectors are surprisingly sensitive. We first identify good keypoint candidates in multiple training images taken from the same viewpoint. We then train a regressor to predict a score map whose maxima are those points so that they can be found by simple non-maximum suppression. As there are no standard datasets to test the influence of these kinds of changes, we created our own, which we will make publicly available. We will show that our method significantly outperforms the state-of-the-art methods in such challenging conditions, while still achieving state-of-the-art performance on untrained standard datasets.

Overview of our Approach We start by collecting a set of training images of the same scene captured from the same point of view but at different seasons and different times of the day. With these images, we identify a set of locations that we think can be found consistently over the different imaging conditions. To learn to find these locations in a new input image, we propose to train a regressor to return a value for each patch of a given size of the input image. These values should have a peaked shape similar to the one proposed in [5] on the positive samples, and we also encourage the regressor to produce a score that is as small as possible for the negative samples. We can then extract keypoints by looking for local maxima of the values returned by the regressor, and discard the image locations with low values by simple thresholding. Moreover, our regressor is also trained to return similar values for the same locations over the stack of images. This way, the regressor returns consistent values even when the illumination conditions vary. An image matching example is shown in Fig. 1.

A Piece-wise Linear Regressor As we want to evaluate the regressor on the whole image, it is important that the regressor is very efficient. In our work, we propose a *piece-wise linear function* expressed using Generalized Hinging Hyperplanes (GHH) [6]:

$$\mathbf{F}(\mathbf{x};\omega) = \sum_{n=1}^{N} \delta_n \max_{m=1}^{M} \mathbf{w}_{nm}^{\top}\mathbf{x} \ , \qquad (1)$$

where $\mathbf{x}$ is a vector made of image features extracted from an image patch, ω is the vector of parameters of the regressor and can be decomposed into $\omega = \left[\mathbf{w}_{11}^{\top},\ldots,\mathbf{w}_{MN}^{\top},\delta_1,\ldots,\delta_N\right]^{\top}$. The $\mathbf{w}_{nm}$ vectors can be seen as linear filters. The parameters δ_n are constrained to be either -1 or +1. N and M are meta-parameters which control the complexity of the GHH.

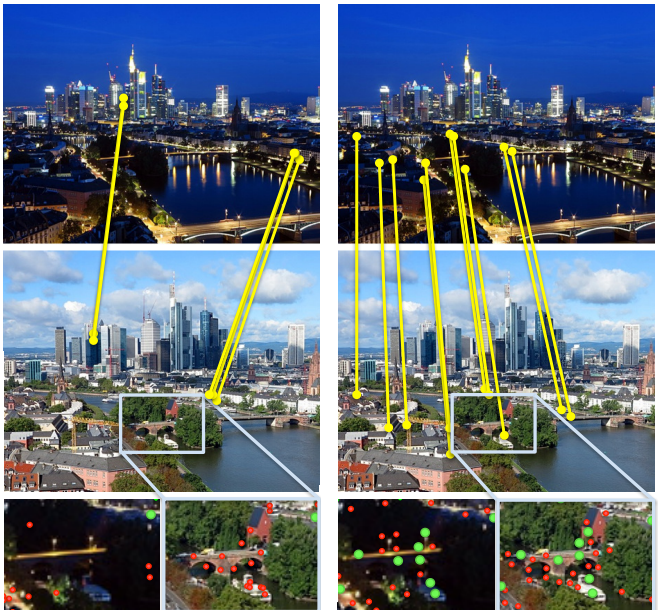
Objective Function The objective function $\mathcal{L}$ we minimize over the parameters ω of our regressor can be written as the sum of three terms, where each terms serves a different purpose:

$$\underset{\omega}{\text{minimize}} \quad \mathcal{L}_c(\omega) + \mathcal{L}_s(\omega) + \mathcal{L}_t(\omega) \ . \qquad (2)$$

The *Classification-Like Loss* $\mathcal{L}_c$ is useful to separate well the image locations that are close to keypoints from the ones that are far away.

The *Shape Regularizer Loss* $\mathcal{L}_s$ makes the regressor have local maxima at the keypoint locations, by enforcing the response of the regressor to have a specific shape [5] at these locations.

The *Temporal Regularizer Loss* $\mathcal{L}_t$ enforces the repeatability of the regressor over time, by making the regressor to have similar responses at the same locations over the stack of training images.



(a) With SURF [1] keypoints (b) With our keypoints

Figure 1: Image matching example using Speeded-Up Robust Features (SURF) [1] and our method. The *same number of keypoints* and descriptor [3] was used for both keypoint detectors. Detected keypoints are shown in the third row, with the repeated ones in green. Only one SURF keypoint detected in the daytime image was repeated in the nighttime image while our method returns many repeated keypoints regardless of the lighting change.

Dataset and Experiments As there is currently no standard benchmark dataset designed to test the robustness of keypoint detectors to various kinds of temporal changes, we created our own from images from the Archive of Many Outdoor Scenes [2] and our own panoramic images to validate our approach, which we refer to as the *Webcam* dataset. We train our keypoint detectors with our *Webcam* dataset and evaluate our method using not only the *Webcam* dataset, but also the standard *Oxford* [4] and *EF* [7] datasets.

Conclusion Detailed experimental results are available in the paper, but our conclusions is three fold; we significantly outperform the state-of-the-art when we apply our keypoint detector on a learned scene, we also outperform the state-of-the-art when applied to different unlearned sequences in the same dataset, and we still achieve state-of-the-art performance even on standard datasets which are completely different from the training dataset.

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. *Computer Vision and Image Understanding*, 10(3):346–359, 2008.

[2] N. Jacobs, N. Roman, and R. Pless. Consistent Temporal Variations in Many Outdoor Scenes. In *Conference on Computer Vision and Pattern Recognition*, 2007.

[3] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 20(2), 2004.

[4] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A Comparison of Affine Region Detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2005.

[5] A. Sironi, V. Lepetit, and P. Fua. Multiscale Centerline Detection by Learning a Scale-Space Distance Transform. In *Conference on Computer Vision and Pattern Recognition*, 2014.

[6] S. Wang and X. Sun. Generalization of Hinging Hyperplanes. *IEEE Transactions on Information Theory*, 51(12):4425–4431, 2005.

[7] C.L. Zitnick and K. Ramnath. Edge Foci Interest Points. In *International Conference on Computer Vision*, 2011.

*First two authors contributed equally