[DEMO] Tracking Texture-less, Shiny Objects with Descriptor Fields

Alberto Crivellaro, Yannick Verdie, Kwang Moo Yi, Pascal Fua
Computer Vision Laboratory
École Polytechnique Fédérale de Lausanne (EPFL)
alberto.crivellaro@epfl.ch

Vincent Lepetit
Institute for Computer Graphics and Vision
Graz University of Technology
lepetit@icg.tugraz.at

Abstract

Our demo demonstrates the method we published at CVPR this year [3] for tracking specular and poorly textured objects. Instead of detecting and matching local features, we retrieve the pose in the input images by aligning them with a reference image exploiting dense optimization techniques. Our main contribution is an efficient novel local descriptor that can be used in place of the intensities to make the alignment much more robust.

Our approach, which requires only a standard, monocular camera (no need for a depth sensor), is of great interest for all Augmented Reality applications involving shiny, texture-less objects, such as those typically encountered in industrial environments.

Index Terms: Robust tracking; Dense Descriptors; Specular objects;

1 Introduction

Despite a long history of research in 3D tracking [4, 6], it is still very challenging to reliably register poorly textured, specular objects. This is a clear obstacle to the development of Robotics and Augmented Reality (AR) applications in industrial environments, where such objects can typically be found.

Because of the typical presence of metallic and non-Lambertian masses in such environments, 3D sensors and depth cameras do not perform well, and markers are at best highly impractical. Because of the absence of textures and presence of specularities, standard Computer Vision techniques, which rely mostly on feature points or texture, also completely fail.

We use an an approach that we refer to as "Descriptor Fields", introduced in [3], that resolves this issue. Instead of detecting and matching local features, we rely on a dense image alignment framework [7, 2, 5, 1, 9, 10]. Dense alignment is attractive because it globally exploits most of the image information, even when local image features such as interest points or edges are ambiguous. However it typically relies directly on image intensities, which is prone to fail in presence of non-Lambertian effects such as specularities, or when the objects do not exhibit convenient textures. Moreover, a multi-scale approach is typically required for robust alignment, where low-pass filters are applied to the signals to align. When the signals are the image intensities, or a linear combination of them, low-pass filtering rapidly deteriorates information.

We therefore employ a more robust local descriptor in place of the pixel intensities. Our descriptor allows us to handle challenging imaging artifacts on a highly specular, poorly textured object. It is computed from a small set of convolutional filters applied to the images, which makes it suitable for real-time applications. However, instead of relying on the simple linear transformation of the intensity signal issued by the convolutions, we apply a non-linear operation that separates the descriptors' positive values from the negative ones. This step is crucial for obtaining the best tracking performances.

This can be explained by the fact that, thanks to our non-linear operation, our Descriptor Fields remain discriminant even after low-pass filtering. As a result, large Gaussian kernels can be used to significantly broaden the region of convergence of the alignment optimization algorithms, which is an important factor for robustness.

2 Dense Alignment for Camera Tracking

For retrieving an estimation of the pose $\widehat{\mathbf{p}}$ of a captured image J, we align it against a reference image T. More exactly, optimize the following cost function:

$$F(\mathbf{p}) = \sum_{\mathbf{x}} \|d(J, W(\mathbf{x}, \mathbf{p})) - d(T, \mathbf{x})\|^2 \,, \tag{1}$$

where $d(I, \mathbf{x})$ is a function that returns a descriptor for location $\mathbf{x}$ in a generic image I, $W : \mathbb{R}^2 \times \mathbb{R}^n \to \mathbb{R}^2$ is a warp function depending on an array $\mathbf{p}$ of n parameters, and the sum is extended to a dense subset of the pixels $\mathbf{x}$ on the template. Finally, we set:

$$\widehat{\mathbf{p}} = \mathrm{argmin}_{\mathbf{p}}\, F(\mathbf{p}). \tag{2}$$

We employ an homography ($n = 8$) for tracking planar objects, or the transfer function used in [3] ($n = 6$) for tracking a general 3D scene. In previous dense alignment works, $d(I, \mathbf{x})$ is almost always taken as $I(\mathbf{x})$, the intensity in image I at location $\mathbf{x}$.

We can exploit several algorithms to efficiently optimize functions in the form of Equation (1), including the Lucas-Kanade (LK) algorithm [7, 1], the Inverse Compositional Algorithm (ICA) [1], and the Efficient Second Order Method (ESM) [8].

In practice, a multi-scale approach is used to optimize Eq. (1) by taking and considering the intermediate objective function:

$$F(\mathbf{p}; \sigma) = \sum_{\mathbf{x}} \|D_\sigma(J, W(\mathbf{x}, \mathbf{p}_T, \mathbf{p})) - D_\sigma(T, \mathbf{x})\|^2 \,, \tag{3}$$

where $D_\sigma(\mathbf{x})$ is a low-pass version of $d(\mathbf{x})$:

$$D_\sigma(\mathbf{x}) = (G^\sigma * d)(\mathbf{x}) \tag{4}$$

with G^σ a Gaussian kernel of standard deviation σ. The optimization scheme starts with a large value for σ, optimizes $F(\mathbf{p}; \sigma)$ to obtain a first estimate $\tilde{\mathbf{p}}$ of the actual pose, decreases σ, optimizes $F(\mathbf{p}; \sigma)$ again starting from $\tilde{\mathbf{p}}$, and iterates for a fixed number of iterations.

This multiscale optimization scheme is important in practice as low-pass filtering increases the basin of convergence, but it also degrades the localization of the minimum of the original function in Eq. (1). In our implementation, the optimization is initialized with the camera pose for the template T. We use 4 scales, with σ initialized to a fixed parameter $\sigma_{\max}$ for the coarsest scale, and divided by 2 between each scale level.

The next section discusses how we compute the d function to improve the convergence when the images exhibit challenging artifacts.

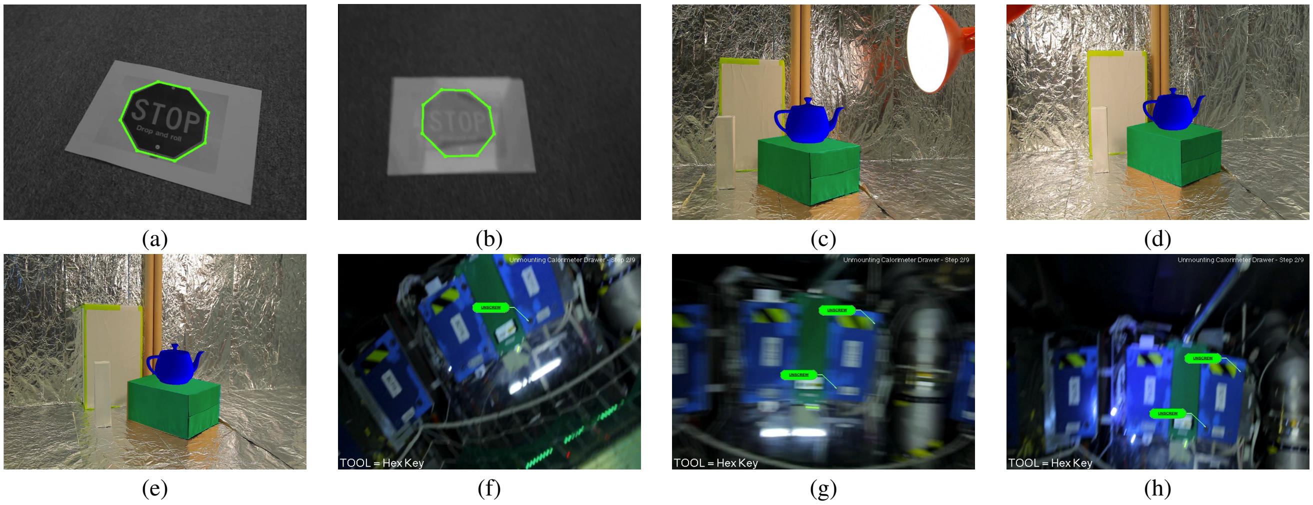

Figure 1: Some examples of objects our method can track: (a) a flat, lambertian object; (b) a flat, shiny object with specular artifacts; (c), (d), (e) a 3D scene with shiny and poorly specular objects; (f), (g), (h) an indstrial machinery in a cluttered environment.

3 Descriptor Fields

As mentioned in the previous section, a very common choice for the function $d(I,\mathbf{x})$, which appears in Eq. (1) and on which image alignment is based, is simply

$$d(I,\mathbf{x}) = I(\mathbf{x})\,, \tag{5}$$

that is, the pixel intensity in image I at location $\mathbf{x}$. However, this option is very sensitive to complex light changes, especially in the absence of texture. For improving the tracking robustness, we employ the Descriptor Fields proposed in [3].

They consist in the following dense descriptor:

$$d(I,\mathbf{x}) = \begin{bmatrix} [(\mathbf{f}_1 * I)(\mathbf{x})]^+, [(\mathbf{f}_1 * I)(\mathbf{x})]^-, \dots, \\ [(\mathbf{f}_n * I)(\mathbf{x})]^+, [(\mathbf{f}_n * I)(\mathbf{x})]^- \end{bmatrix}^\top, \tag{6}$$

where the $\mathbf{f}_i$ filters are typically Gaussian derivatives kernels, and the $[\cdot]^+$ and $[\cdot]^-$ operations respectively keep the positive and negative values of a signal:

$$[x]^+ = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}, \text{ and} \qquad [x]^- = [-x]^+ .$$

These operations are simple but fundamental, and make the descriptor of Eq. (6) stay discriminant after strong Gaussian smoothing. This yields an objective function with a large basin of attraction and a well localized minimum, which is key for robustness of the alignment.

4 Demo

Our demo aims at highlighting the advantages of our 3D tracking approach based on Descriptor Fields for shiny, texture-less objects, such as a whiteboard, a plastified sheet of paper or an industrial machinery.

Users will be able to experiment the robustness of our approach by comparing it with state-of-the-art methods, such as those based on local features or dense image alignment and intensities.

We believe that our approach will contribute to expand the domain of application of Augmented Reality to environments where previous visual tracking techniques do not provide enough robustness. A demo video of our tracking framework is available on the project page at `http://cvlab.epfl.ch/research`.

5 Conclusion

We present a visual tracking framework based on a novel local descriptor, the Descriptor Fields, that makes tracking much more robust to various imaging artefacts. Since Descriptor Fields are efficient and very simple to implement, they are suited for AR applications in difficult environments.

Acknowledgements

This work was supported in part by the EU project EDUSAFE.

References

[1] S. Baker and I. Matthews. Lucas-Kanade 20 Years On: A Unifying Framework. *IJCV*, pages 221–255, March 2004.

[2] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active Appearance Models. *PAMI*, 23(6), June 2001.

[3] A. Crivellaro and V. Lepetit. Robust 3D Tracking with Descriptor Fields. In *CVPR*, 2014.

[4] C. Harris and M. Stephens. A Combined Corner and Edge Detector. In *Fourth Alvey Vision Conference*, 1988.

[5] F. Jurie and M. Dhome. Hyperplane Approximation for Template Matching. *PAMI*, 24(7):996–100, July 2002.

[6] D. G. Lowe. Robust Model-Based Motion Tracking through the Integration of Search and Estimation. *IJCV*, 8(2):113–122, 1992.

[7] B. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *IJCAI*, pages 674–679, 1981.

[8] E. Malis. Improving Vision-Based Control Using Efficient Second-Order Minimization Techniques. In *ICRA*, pages 1843–1848, 2004.

[9] C. Mei, S. Benhimane, E. Malis, and P. Rives. Efficient Homography-Based Tracking and 3D Reconstruction for Single-Viewpoint Sensors. *IEEE Transactions on Robotics*, 24(6):1352–1364, 2008.

[10] R. Newcombe, S. Lovegrove, and A. Davison. DTAM: Dense Tracking and Mapping in Real-Time. In *ICCV*, 2011.