Learning Descriptors for Object Recognition and 3D Pose Estimation

Paul Wohlhart, Vincent Lepetit
Institute for Computer Vision and Graphics, Graz University of Technology, Austria

Detecting and recognizing poorly textured objects and estimating their 3D pose reliably is still a very challenging problem. Although impressive results have been achieved in 3D pose estimation of objects from images during the last decade, current approaches cannot scale to large-scale problems because they rely on one classifier per object, or multi-class classifiers such as Random Forests, whose complexity grows with the number of objects.

So far the only recognition approaches that have been demonstrated to work on large scale problems are based on **Nearest Neighbor (NN) classification**, because extremely efficient methods for NN search exist with an average complexity of $O(1)$. Moreover, Nearest Neighbor classification also offers the possibility to trivially add new objects, or remove old ones. which is not directly possible with other classifiers, such as neural networks, for example. However, to the best of our knowledge, such an approach has not been applied to the 3D pose estimation problem, while it can potentially scale to many objects seen under large ranges of poses.

For NN approaches to perform well, a **compact** and **discriminative description vector** is required. Such representations that can capture the appearance of an object under a certain pose have already been proposed, however they have been handcrafted. We, thus, introduce a simple but powerful approach to computing descriptors for object views that efficiently capture both the object's identity and 3D pose.

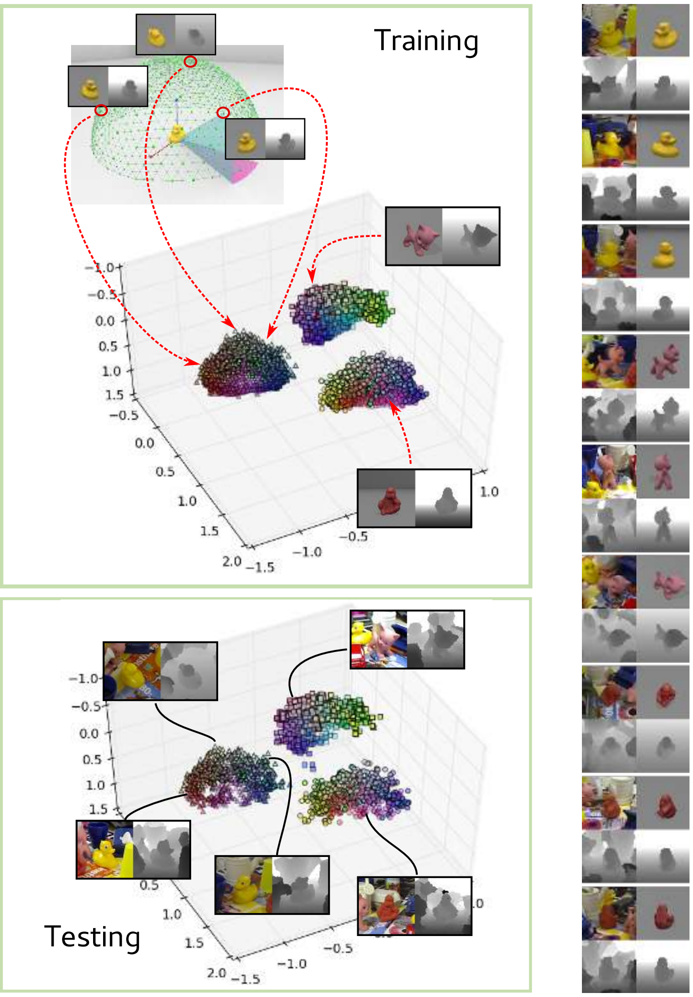

Figure 1: Exemplary 3-dim. descriptors for several objects under many different views computed by our method on RGB-D data. **Top-Left:** The training views of different objects are mapped to well-separated descriptors, and the views of the same object are mapped to descriptors that capture the geometry of the corresponding poses, even in this low dimensional space. **Bottom-Left:** New images are mapped to locations corresponding to the object and 3D poses, even in the presence of clutter. **Right:** Test RGB-D views and the RGB-D data corresponding to the closest template descriptor.

Method

What we seek is a **function that maps input images to descriptors** with the two following properties: a) The Euclidean distance between descriptors from two different objects should be large; b) The Euclidean distance between descriptors from the same object should be representative of the similarity between their poses. This way, given a new object view, we can recognize the object and get an estimate of its pose by matching its descriptor against a database of registered descriptors.

To achieve this, we train a **Convolutional Neural Network** to compute these descriptors by enforcing simple similarity and dissimilarity constraints between the descriptors. In particular, we train the network by minimizing a **cost function based on pairs and triplets** of samples. The pair-wise cost enforces the descriptors of different images of the same object from the same viewpoint to be the same, making the representation robust to noise and different imaging conditions. Triplets are formed by associating with each training sample a more similar and a more dissimilar sample (templates) that will later on be used in the NN search. The similar one is set to be the template with the most similar pose. Dissimilar samples are either other templates from the same object but different pose, or a template from a different object under any pose. The triplet-cost is then defined to be positive if the descriptors of the dissimilar samples are closer than those of the similar samples, forcing the descriptors of different objects to be far apart and those of each object to arrange on individual manifolds.

Results

We train the network on a **mixture of real and synthetic views** of objects from the LineMOD dataset. Real images recorded with Kinect are provided. We additionally render synthetic views of the available 3D models against clean background to create templates and additional training data samples from further refined poses and with added noise.

We show that **our constraints nicely untangle the images** from different objects and different views into clusters that are not only **well-separated** but also **structured as the corresponding sets of poses**, as demonstrated in Figure 1 for a toy example with three objects and only three-dimensional descriptors. Additionally, results on a the LineMOD benchmark data show that our descriptors outperform state-of-the-art object views representations for this task. We show that for the given use case an **only 16-dimensional descriptor** computed with our method is enough to achieve excellent recognition and pose estimation performance at low angle error when only taking the first nearest neighbor in the lookup step as final output.

In a further experiment we also show the potential of the method to **generalize to unseen objects**. We take the pre-trained network to compute descriptors for templates of an object that was not seen during training and use them in the NN classification. The results show that, while the classification score is slightly reduced since the network was not trained to differentiate views under which some of the objects appear to be very similar, the general tendency of correlating differences in pose and object identity with descriptor distance is well preserved.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.