Object Pose Estimation using Patch-Duplet/SIFT Hybrids

Fredrik Viksten
Information Coding Group
Linköping University, Sweden
viksten@isy.liu.se

Abstract

Recent years have seen a lot of work on local descriptors. In all published comparisons or evaluations, the now quite well-known SIFT-descriptor has been one of the top performers. For the application of object pose estimation, one comparison showed a local descriptor, called the Patch-Duplet, of equal or better performance than SIFT. This paper examines different properties of those two descriptors by forming hybrids between them and extending the object pose tests of the original Patch-Duplet paper. All tests use real images.

1 Introduction

Object pose estimation, or estimation of the 6 degree-of-freedom *geometrical state* of one or more objects from a single 2D image is an important problem that has received considerable attention over the years [20, 16, 19, 9]. Applications include industrial automation such as bin picking (see figure 1), support systems for augmented reality as well as a whole range of consumer products including toys and house-hold appliances.

The local features typically used in view-based pose estimation have previously been evaluated for the purposes of view matching, and object recognition. In such evaluations, computation can be divided into three steps: *detection* of interest points, *descriptor construction*, and descriptor *matching* [5].

Figure 1: Bin picking.

A pose estimation system, by necessity, has to contain two additional steps: *pose hypothesis generation*, and *pose clustering* [20, 9]. This justifies the need for specific evaluation of local features in the pose estimation framework.

1.1 Related Research

Evaluation of interest point detectors and local descriptors have previously been done on the wide-baseline stereo task [13, 5, 14], and in the setting of recognition of objects or object class [12, 15]. The object pose estimation problem is however sufficiently different from wide baseline stereo and general object recognition to require a separate feature evaluation. In object recognition and wide baseline stereo, view invariance for features is a good thing. In the object pose estimation application it is on the other hand important that a descriptor can be distinguished within a large database of descriptors, many of which were generated from visually similar image patches. In other words, the features need to be view specific if they are to tell one view from another. For this reason it is not obvious that a pose estimation evaluation will rank local descriptors in the same way as wide-baseline and object recognition tests.

All published comparisons or evaluations show the now quite well-known SIFT-descriptor to be one of the top performers. In the object pose estimation test published in [8] results were slightly different, showing SIFT [11] being outperformed by the local descriptor introduced in that paper called the Patch-Duplet. We will examine that claim once again and also try to evaluate what the two descriptors' strengths and weaknesses are.

1.2 Contributions

The contribution of this work is an extention of the tests found in [8] with an additional sequence of pose estimation under light setting changes on hybrids between the two local descriptors in [8] to work out how we might improve SIFT or the Patch-Duplet in the setting of object pose estimation.

2 Pose Estimation Framework

This presentation uses a match–vote–cluster scheme for performing view-based pose estimation. The approach is common in the literature [9, 8, 6].

When estimating an object pose from local image features, it is convenient to use this coordinate representation, $\mathbf{E} = (x\ y\ \Delta\alpha\ \Delta s\ \phi\ \theta)^T$, which we refer to as *estimation coordinates*. Two degrees-of-freedoms (DOF) can be determined from the image plane location of the object (x, y). Another two DOFs are given by the relative image plane rotation $(\Delta\alpha)$, and the relative scale change (Δs), both in relation to a reference

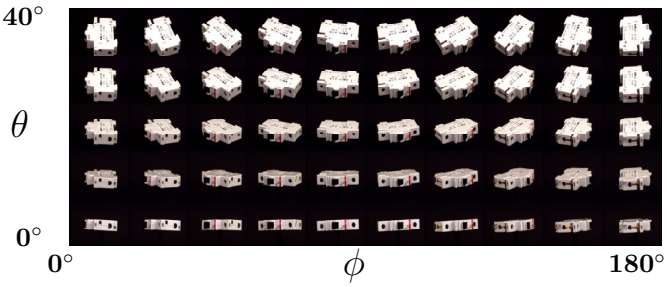

Figure 2: An object sampled over the two pose angles.

view. The two remaining DOFs are represented by the two object rotation angles (ϕ, θ), see figure 2. We will in this presentation refer to these angles as *pose angles*.

The system is trained using a set of real images of an object sampled from pose angles. It can be argued that the more physical state attributes the method/system is invariant to, the fewer samples are needed. The features we use are invariant to position, image plane rotation and to some extent scale. We therefore only need to sample images in 10° steps of the two pose angles ϕ and θ, see figure 2.

Collecting and storing data in this manner for later use in e.g. interpolation, is known as *lazy learning* or *memory based learning* [1]. Interestingly, the human vision system also appears to work as if it used database look-up functions when recognizing objects [4].

During training, the system does the following for each training image:
1) Detect interest points (IP). 2) Extract local descriptors. 3) Store each descriptor together with *auxiliary information*.

In the case of the SIFT descriptor the auxiliary information consists of the pose angles (ϕ, θ) the position in the image where the IP was found (x, y), the scale where it was found (s), and the reference direction specified by the SIFT descriptor (α).

Once the system has been trained, we want to use it to estimate the geometrical state of an object (represented by the estimation coordinates). The whole estimation procedure is illustrated in figure 3, and can briefly be described according to:
1) Detect IPs. 2) Extract descriptors. 3) Find the k most similar features in the database. 4) Retrieve the pose angles from the auxiliary information. 5) Compute the rest of the pose estimate using the feature location and scale, and the auxiliary information. 6) Cluster in the 6 dimensional *vote space* (where the estimation coordinates live) to find the most likely pose estimate.

The last step is the same for all methods in this presentation. To find local density peaks in this space and estimate a mean of such a peak, or cluster, mean-shift clustering [3] is used. Mean shift clustering outputs a cluster density value D_i for each cluster, with $D_i \geq D_{i+1}$ and from these we compute a certainty measure $c \in [0, 1]$, as $c = 1 - D_2/D_1$. A high c value signifies that the highest peak D_1 in the pose estimate density is well above the second highest D_2, and is thus most likely the correct one. This approach is quite common in the literature, see e.g. [9, 8, 19, 6]. Please refer

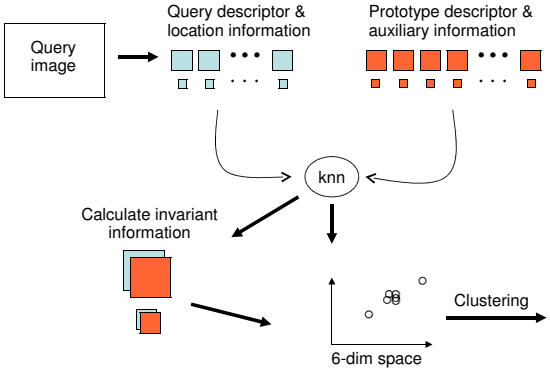

Figure 3: Overview of the query mode.

to [8] for more details on the pose estimation.

3 Local Descriptors

3.1 Patch-Duplets (PD)

The Patch-Duplet [8], referred to as PD, uses a sub-pixel Harris detector for IP detection. It extracts IPs and descriptors at two resolutions of the input image. This method forms pairs between each IP and its four closest IPs.

The patch-duplet uses a descriptor computed from the *double angle* (DA) representation [7] of the local orientation in box-shaped area around each IP. The idea behind this representation is that intensity steps from dark to light or light to dark will produce the same orientation vector. Ordinary vector fields are converted to DA representation by multiplying vector direction angles by 2 and clamping to $[0, 360]$.

The connection of two IPs gives both an orientation for the boxes as well as a size for the area which depends on the distance between the IPs. Duplets use the distance between its two points to recover scale. Rotation and position of the object uses the center point on the line connecting the two IPs in the duplet.

3.2 SIFT

A very good and detailed presentation of the scale invariant feature transform can be found in [11]. For details on how SIFT is used in pose estimation, see section 2. We used the implementation provided at [10].

3.3 Hybrids

The two descriptors above are different in a number of ways so we decided to implement some hybrid versions to find if any specific difference is of specific importance. In this work we focus on the descriptors, so we decided to use the same IP detectors for all our new hybrids. The versions are based upon code found at [17]. In table 1 we show the different versions, including our new ones (bold text) and the previous published descriptors, that use DoG for IP detection.

Converting SIFT to the DA representation is done by forming the orientation histogram of the descriptor from an underlaying orientation in DA representation.

Table 1: Descriptors using DoG as detector.

#IP	360° angle	Double-angle representation
1	SIFT [11]	Patch-singlet (**PS**), DA-SIFT (**DAS**)
2	SIFT-duplets (**SIFTD**)	Patch-duplet (**PDDoG**), DA-SIFT-duplets (**DASD**)

To keep the angular resolution of the orientation bins in the SIFT histogram we halved the number of orientation bins since one interpretation is that the DA representation only uses angles between 0 and 180 degrees.

Converting SIFT to a duplet is not entirely novel, in [2] local invariant frames where formed by a number of points before the extraction of SIFT descriptors. In our application we do with-out affine transformations since the information needed is actually in the shape of the detected region [18], it is what descriminates one pose angle from an other.

Connection of the two DoG IPs in the new SIFT-duplets and the new Patch-Duplet is done only from finer to coarser scale to make sure that the second IP in the descriptor has a high probablility of existing if the first one does. Besides this ordering of IPs, the DoG detected scale is discarded and scale is detected as for the Patch-Duplets, i.e. by changes in the distance between the two IPs. Orientation for each of the two descriptors in the duplet is the same and is set by the line connecting the two IPs and thus there is no need to calculate the descriptor orienation as in the original SIFT formulation. For an easy comparison between descriptors, we shortened the orientation histogram to 4 bins for SIFTD, thus halving the orientation histogram resolution. This gives the same descriptor vector length as for SIFT. Sampling of the DoG-based Patch-Duplet descriptor is done from different layers in the scale pyramid depending on descriptor area, which in turn depends on the distance between IPs of the duplet.

The Patch-Singlet extracts a single part, i.e. half, of the Patch-Duplet descriptor at the orientation and scale found by the DoG detector. The area size of the extracted descriptor is given by

$$R = 1.4 n_v \sigma_0 2^{1+s/S} \tag{1}$$

where n_v is the number of vectors in each spatial direction, σ_0 and S is set in the DoG detector and s is the scale at which the IP is detected, see [11] and [8].

3.4 Descriptor size

Besides the performance, it is also of interest to compare the number of elements in each descriptor, see table 2. A larger descriptor means more storage requirements, and thus, at equal performance, a smaller descriptor is usually preferred.

Table 2: Number of elements in each descriptor.

32	PS
64	DAS, PD, PDDoG
128	DASD, SIFT, SIFTD

4 Pose Estimation Experiment

The new test in this work uses a view sampling as in figure 2, with three light settings of the object seen in the example images (one from each light setting) in figure 4. In this test we trained for each light setting and then evaluated on the two other light settings. The evaluation is done using real images of the object in figure 4 with both black background and with cluttered background (middle image in figure 4).

All learning used 10° intervals for both the pose angles. The evaluation is then done at the sample positions in-between, yielding a worst-case in regards to geometric image distortions from the training images to the evaluation images. This gives 95 training views and 72 evaluation views per light setting. The test is performed both without a background, which is similar to having objects on a conveyor belt in a factory, and with a heavily cluttered background. Thus each background setting uses 432 evaluation poses/images.

Each evaluation view is subjected to a random scaling between 1.3 and 0.7 as well as a random image plane rotation between $-180°$ and 180°. For all scale values, downsampling from the original high resolution images is used. All interpolations are done by bi-cubic interpolation. The vector of random scaling and rotation for each view was created once and then reused for all descriptors. This ensures a fair comparison.

The error values presented in the experiment results are distances between estimates and ground-truth. The two pose angles are combined into a vector on the unit sphere so only one value is presented. The error value is given in degrees by the space angle between the ground-truth vector and the measured pose angles vector. Scale error is linear. In the same plot as the object state values, we show the measure of certainty.

4.1 Results

The results from experiments on black background can be seen in figure 5. They are somewhat different than those reported in [8], SIFT slightly outperforms PD in this test. We see that the hybrid versions all improve on those results, with the DA version of the SIFT duplet being the strongest performer if looking at the boxplots. However taking the number of outliers found on the right of each plot into account, PDDoG

Figure 4: Ambient, left, and right illumination.

should be considered an equal performer. We also see that PS is not doing well compared to the other ones.

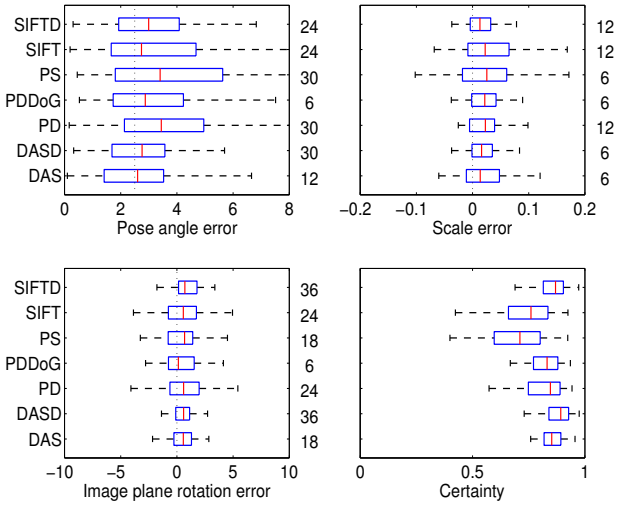

Figure 5: Error values for black background.

Results for the test with cluttered background can be found in figure 6. Again, the results are not quite what was shown in [8], even though the difference between SIFT and PD is quite small.The hybrid versions all show better performance than the original descriptors. We see that in this case, PDDoG is the best performer both in accuracy and in the number of outliers which is very important for the robustness of a fully automatic system.

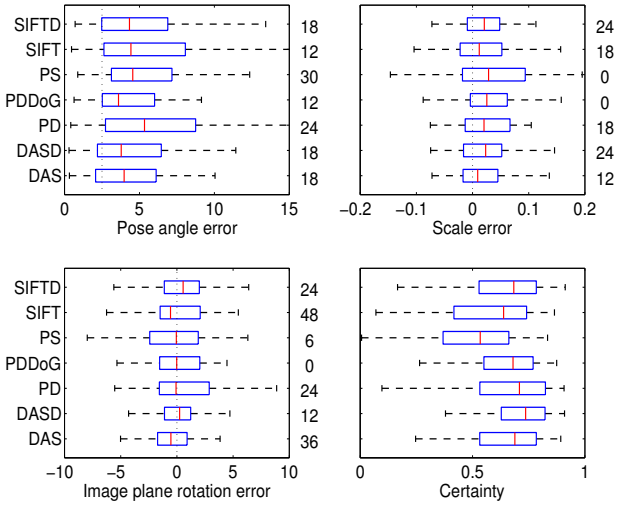

Figure 6: Error values for cluttered background.

The certainty measures show us that all the hybrids form a tighter cluster in the voting space.

5 Concluding remarks

DoG-based duplets were found to produce by far the most descriptors followed by PD and then the DoG singlets. The ordering was the same for black background as for cluttered background. Also measuring the number of descriptors in the winning cluster and normalizing by number of descriptors found on the object (i.e. from the black background test) showed the highest percentage for PD, followed by SIFT/DAS and then the DoG-based duplets and last PS for both black and cluttered background. This seems to suggest that most of the performance gain for the DoG-based duplets comes from the more numerous descriptors for those methods.

We have seen that the new hybrids can improve the performance of a pose estimation system. Robustness can also be increased by reducing the number of outliers in pose estimates which is very important in a bin-picking application since it will reduce damage on the system and unneccessary stops.

References

[1] C.G. Atkeson. Using locally weighted regression for robot learning. In *ICRA*, pages 958–963, 1991.

[2] M. Brown and D. Lowe. Invariant features from interest point groups, 2002.

[3] Yizong Cheng. Mean shift, mode seeking, and clustering. *PAMI*, 17(8):790–799, August 1995.

[4] S. Edelman and H. Bülthoff. Modeling human visual object recognition. In *Proc. IJCNN*, Sept. 1992.

[5] K. Mikolajczyk et al. A comparison of affine region detectors. *IJCV*, 65:43–72, 2005.

[6] P-E Forssén and A. Moe. Autonomous learning of object appearances using colour contour frames. In *CRV*, June 2006.

[7] G. Granlund and H. Knutsson. *Signal Processing for Computer Vision*. Kluwer Academic Publishers, 1995.

[8] B. Johansson and A. Moe. Patch-duplets for object recognition and pose estimation. In *CRV*, May 2005.

[9] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC*, 2003.

[10] David Lowe. Demo software: SIFT keypoint detector, 2005. http://www.cs.ubc.ca/∼lowe/keypoints/.

[11] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[12] K. Mikolajczyk, B. Leibe, and B. Schiele. Local features for object class recognition. In *ICCV*, pages 1792–1799, 2005.

[13] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60:63–86, 2004.

[14] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27:1615–1630, 2005.

[15] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3D objects. *IJCV*, 73:263–284, July 2007.

[16] R. Söderberg, K. Nordberg, and G. Granlund. An invariant and compact representation for unrestricted pose estimation. In *Iberian Conf. on PRIA*, June 2005.

[17] A. Vedaldi. SIFT - an open implementation of SIFT, 2006. http://vision.ucla.edu/∼vedaldi/code/sift/sift.html.

[18] F. Vikstén, P-E Forssén, Björn Johansson, and A. Moe. Comparison of local image descriptors for full 6 degree-of-freedom pose estimation. In *ICRA*, May 2009.

[19] F. Vikstén and A. Moe. Local single-patch features for pose estimation using the log-polar transform. In *Iberian Conf. on PRIA*, June 2005.

[20] F. Vikstén, R. Söderberg, K. Nordberg, and C. Perwass. Increasing Pose Estimation Performance using Multi-cue Integration. In *ICRA*, May 2006.