Probabilistic Cue Integration for Real-time Object Pose Tracking*

Johann Prankl, Thomas Mörwald, Michael Zillich, Markus Vincze

Automation and Control Institute
Vienna University of Technology, Austria
{*prankl,moerwald,zillich,vincze*}*@acin.tuwien.ac.at*

Abstract. Robust real time object pose tracking is an essential component for robotic applications as well as for the growing field of augmented reality. Currently available systems are typically either optimized for textured objects or for uniformly colored objects. The proposed approach combines complementary interest points in a common tracking framework which allows to handle a broad variety of objects regardless of their appearance and shape. A thorough evaluation of state of the art interest points shows that a multi scale FAST detector in combination with our own image descriptor outperforms all other combinations. Additionally, we show that a combination of complementary features improves the tracking performance slightly further.

1 Introduction

Vision systems for detection and tracking of objects are dominated by approaches based on interest points and local descriptors. While successful for textured objects the performance decreases if objects are uniformly colored. In industrial applications the variety of objects is limited and the environment can be adapted, but home and service robotic scenarios require systems which are able to handle different kinds of objects and a cluttered environment in a common framework. This can be achieved by designing a framework with multiple components, each optimized for the detection of one object category. In contrast, we propose a single detection and tracking approach which is able to handle a broad variety of objects by integrating complementary features. We evaluate state of the art interest point detectors and descriptors and develop a strategy to combine different feature types by learning a probabilistic reliability model. Note that in the remaining text we use the term interest point to indicate a salient image point plus the descriptor computed from the surrounding patch.

Starting with the *scale invariant feature transform (SIFT)* developed by Lowe [1] numerous successful interest point types have been proposed. A comparison of affine region detectors and descriptors can be found in [2] and [3]. Most

* The work described in this article has been funded by the European project under the contract no. 600623, as well as by the Austrian Science Fund under the grant agreements I 513-N23 and TRP 139-N23 and the Austrian Research Promotion Agency under the grant agreement 836490.

of them are optimized for object recognition and are able to handle a limited number of objects with a fair amount of texture. More recently, interest points which are faster to compute and thus are more appropriate for object tracking have been proposed (FAST [4], MSER [5], SURF [6], ORB [7]. We focus on these interest point types and compare them to our own SIFT-like descriptor which we simply call *image gradient histogram descriptor (ImGD)*.

Our work is placed in the context of robotics applications where an accurate object pose is necessary for path planning or visual servoing. Hence, we evaluate the interest points within a complete tracking system. The tracker uses a monocular image sequence to compute the object pose and it is based on an iterative particle filter framework similar to the approach proposed by Mörwald [8]. In contrast to Mörwald who renders complete textures of 3D models into the image in order to compute particle quality, our object model consists of a sparse set of interest points of possibly different types and their 3D location on the object surface. A problem when trying to integrate interest points of different types is how to compare their matching quality. Therefore, we propose to learn a probabilistic confidence measure for each interest point detector/descriptor pair using Bayes theorem, which is used during tracking to compare and rank matched point pairs.

In this paper we propose a probabilistic framework for tracking objects by combining complementary interest point types. Concretely, our contributions are:

- A probabilistic tracking approach using Bayes theorem to combine complementary interest points.
- A framework for learning and evaluation of the probabilistic model.
- An evaluation of state of the art interest points including Harris [9] FAST [4], MSER [5], SIFT [1], SURF [6] and our own ImGD.

The paper proceeds with a discussion of the related work in Section 2. After that, the method, including learning of the probabilistic model and object tracking is described in Section 3. Then the interest point detectors and descriptors are reviewed in Section 4 which are evaluated in Section 5.

2 Related Work

State of the art interest points are reviewed in Sec. 4. In what follows we summarize related work for object pose tracking.

Robustness has always been a concern for visual tracking [10–12]. Especially the introduction of particle filtering to visual tracking [13] boosted robustness [14–16], aided by the increased use of GPU-optimised algorithms [17, 18]. Also the combination of complementary cues proved to boost robustness and broaden the range of objects that could be handled. [19] extended earlier work on tracking based on planar patches to take into account contour information, but remain limited to planar objects. [20] extended their earlier feature-point based 3D tracker with the ability to integrate edge information, handling the problem of ambiguities resulting from spurious background edges.[21] start with a 3D wireframe model based tracker and augment this with point features (small image patches around Harris corners [9]) collected online from front-facing surfaces and fuse measurements using an Iterated Extended Kalman Filter (IEKF). [22] combine edges and texture (again image patches around Harris corners) in a single non-linear minimization scheme, which they apply to 2D tracking (homography estimation) of planar objects delineated by straight edges or NURBS, as well as full 3D tracking of objects such as boxes and balls. [23] combine depth data, appearance features, silhouettes and even tactile data within an Unscented Kalman Filter (UKF) and achieve high robustness tracking an articulated robotic manipulator and complexly shaped work piece.

In our work we are interested not in a particular choice of feature combination, but in the methodology to evaluate and combine different features. This is then applied to full 3D tracking of arbitrarily shaped 3D objects.

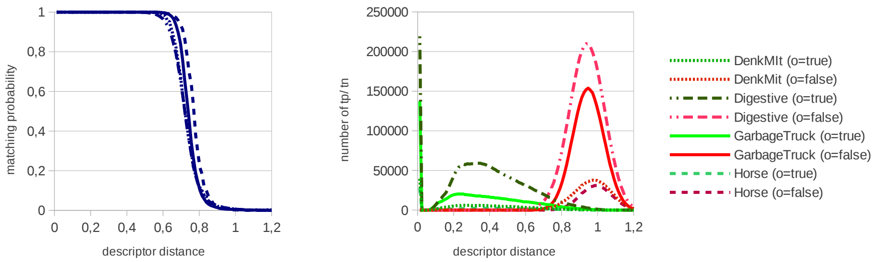
3 Method

Algorithms for object pose estimation need three or more corresponding 2D/3D point pairs [24]. To tolerate inaccurate and false correspondences these methods are typically embedded in a robust estimation schemes (e.g. RANSAC). In order to ensure convergence a set of "good" point correspondences with a high probability of being correct is necessary. In case of using a single feature type this is often implemented by comparing interest point descriptors and applying a threshold. If different types are combined a solution would be to provide individual thresholds for each descriptor. Instead of using *heuristic* thresholds, we propose to learn a mapping of descriptor distances to a probabilistic confidence measure using Bayes theorem. In detail, given an object model consisting of interest points, their 3D location on the object surface and interest points detected in a query image, the goal is to compare model descriptors with descriptors of query points and compute the probability of being a correct match. *Good* matches are then used to estimate the 3D location of the object with respect to the camera.

3.1 Training of the Probabilistic Model

Object models are reconstructed by collecting a sequence of RGB-D images and using a standard RGB-D SLAM approach [25, 26] to compute the camera pose.[1] For training of the probabilistic model we need positive training examples, i.e., correct matches (tp) and negative examples (tn). Positive and negative examples are directly acquired from the image sequence used for reconstruction. In each frame interest points are detected and reprojected to the object surface. Then neighboring reprojected points from all frames are clustered and marked as positive training example if the size of the corresponding image patches for descriptor calculation is similar. Patch similarity is given if the difference of the semi-axes of circumscribed ellipses is smaller than a threshold t_e and if the orientation is similar ($\Delta\theta_{i,j} < t_\theta$). Negative training examples are sampled from matches of

[1] There is no restriction to RGB-D images. The reconstruction pipeline could easily be substituted by a SfM approach using a monocular image sequence.



(a) matching probability

(b) probability density

Fig. 1. Example of a learned matching probability (multi scale FAST detector and ImGD).

one descriptor of a cluster and a randomly selected interest point outside of the cluster or of an interest point detected in a dedicated "false positive" image set.

With $d = ||D(i) - D(j)||_2$ being the descriptor distance of matched interest point pairs $m(i, j)$ the training set is used to compute the prior probability $p(m = true)$ of correct matches, the prior probability $p(d)$ of the occurrence of each distance d in the training set and the conditional probability $p(d|m = true)$ of a descriptor distance d being a correct match $(m = true)$. During tracking the posterior probability $p(m = true|d)$ of being a correct match can then be computed using Bayes rule:

$$p(m = true|d) = \frac{p(d|m = true)p(m = true)}{p(d)}. \tag{1}$$

Probability density functions are approximated with histograms of the descriptor distances. In our implementation we pre-compute the posterior probabilities and use a lookup table during tracking, resulting in probabilities as shown in Fig. 1.

3.2 Object Tracking

Our tracking framework is based on a Sequential Importance Resampling (SIR) particle filter proposed by Doucet et al. [27] and adapted by Mörwald et al. [8] for tracking the 3D pose of objects. In contrast to Mörwald, who renders textured 3D models to the image and counts consistent edgels in order to evaluate the pose hypotheses of particles we project the 3D location of matched interest points to the image and use the distance of these point pairs to compute a confidence value.

In detail, first complementary interest points are detected in a query image (e.g. SURF and FAST+ImGD) and matched to descriptors stored in the object database. For each match we use the descriptors D and the pre-calculated lookup table presented in Section 3.1 to derive the posterior probability $p(m = true|d)$. Matches of different interest point types are then combined and sorted in decreasing $p(m = true|d)$ and the best N matches are passed on to the pose tracking system. Pose tracking is the problem of finding the transformations T_t of an object with respect to a camera given a sequence of images (observations). This results in an estimation of 6 parameters for each observed image. For sampling pose hypotheses directly in a 6 DOF space an untractable number of trials would be necessary. Hence, we bootstrap our system with pose hypotheses computed with the three point pose (P3P) algorithm [24] and a robust RANSAC scheme.

The quality measure

$$c(\mathrm{T}) = \sum_{i=1}^{N} \max(0, t_{inl} - ||\mathbf{p}_{im,i} - \mathrm{C}\,\mathrm{T}\,\mathbf{p}_{model,i}||_2^2) \qquad (2)$$

for P3P-RANSAC and for particles, is computed from the best N interest point matches. Where t_{inl} stands for an inlier threshold, $\mathbf{p}_{im,i}$ is a matched query point, $\mathbf{p}_{model,i}$ is the corresponding 3D model point in homogeneous coordinates and C is the intrinsic camera matrix. In comparison to the tracking framework proposed in [8] which needs a separate recognizer to reinitialize, our framework continuously adds poses from the P3P-RANSAC algorithm and thus automatically reinitializes if the object is lost.

4 Interest Point Detectors and Descriptors

In the following paragraphs we review the interest point detectors and descriptors evaluated with our system. For all detectors and descriptors we use implementations available in OpenCV[2], except SIFT, where we use the implementation of Wu[3], and MSER and ImGD where we use our own implementations.

4.1 Interest Point Detectors

SIFT detector (DoG) is proposed by Lowe [1]. The idea is to approximate the Laplacian of Gaussian with the difference of adjacent Gaussian images which is faster to compute. Additionally Lowe eliminates minima and maxima detected at edges and computes the dominant image gradient orientation. Hence, SIFT detects blob-like structures and it is scale and rotation invariant.

SURF detector (Hes) is developed by Bay et al. [6]. They propose to use an Hessian matrix approximation. The implementation is based on integral images which reduces the computation time drastically. SURF also detects blob-like structures and it is scale and rotation invariant.

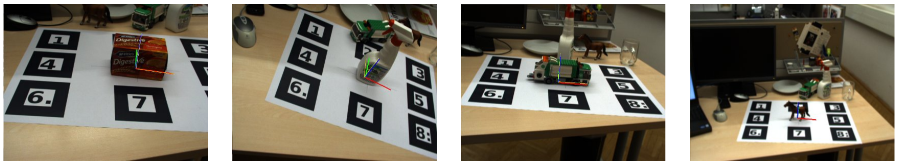
Maximally Stable Extremal Regions (MSER), developed by Matas et al. [5] is an affinely-invariant region detector. It detects the extremal property of the intensity function of regions by increasing/decreasing a brightness threshold and reporting stable parts of the image.

Harris detector is the classical corner detector developed by Harris et al. [9]. To achieve rotation invariance we compute the dominant gradient orientation.

Features from Accelerated Segment Test (FAST), proposed by Rosten et al. [4] is a heuristic for feature detection which uses machine learning

[2] http://www.opencv.org

[3] http://cs.unc.edu/ ccwu/siftgpu/



(a) Digestive (b) DenkMit (c) GarbageTruck (d) Horse

Fig. 2. Example images of our evaluation sequences including the ground truth coordinate system (yellow) detected with the ARToolKit [29] and the tracking coordinate system (red-green-blue).

to classify corner candidates by considering a circle of 16 pixels around a point. We use an implementation proposed in [7] where FAST corners are detected in an image pyramid to cover the scale space and the intensity centroid is used to detect the dominant orientation.

4.2 Interest Point Descriptors

SIFT descriptor, proposed by Lowe [1] describes a patch with histograms of gradient orientations sampled from 4×4 subregions. Each orientation sample is weighted with its magnitude and a Gaussian weight.
SURF descriptor, developed by Bay et al. [6] uses the first order Haar wavelet responses in x and y direction, exploiting integral images for speed, to describe the patch around interest points.
Image gradient descriptor (ImGD), our own descriptor is similar to the original SIFT descriptor proposed by Lowe. We use the same grid layout to compute 4×4 orientation histograms. To speed up computation we skip the interpolation used to distribute the value of each gradient sample into adjacent histogram bins. Instead, we compute an element wise square root of the L1 normalized descriptor proposed by Arandjelović et al. [28]. This transformation is equivalent to using the Hellinger kernel for comparing descriptors instead of the Euclidean distance.

5 Evaluation

To evaluate the interest points we propose two methods. First the *meaningfulness* of the descriptor is compared by computing the probability density function and the corresponding matching probability $p(m = true|d)$. Then we evaluate different interest points and combinations of them with our complete tracking system. In all experiments the Euclidean distance is used to compare descriptors.

5.1 Comparison of Interest Point Detectors and Descriptors

For learning the probabilistic model in order to evaluate the meaningfulness of descriptors we use the reconstruction and training pipeline described in Section 3.1. We select four objects ranging from highly textured surfaces with a simple repetitive shape 2(a) to a single colored surface with a more complex shape 2(d). For each object we used about 800 RGBD-frames, compute the camera poses and reconstruct the upper hemisphere of the objects. The interest points detected in each frame and reprojected to the object surface are used to generate true and false training examples. Depending on the interest point type and the object surface this results in up to $500k$ positive and $1.5M$ negative examples.

In general, it can be seen in Figs. 1 and 3, that different interest point detectors result in different matching probabilities even if the same descriptors are used. In every case ImGD leads to a better separation of true and false matches, i.e., a steeper slope of the posterior matching probability, no matter which interest point detector is used. It can also be seen, that the density functions have a similar behavior for each object which results in almost identical posterior curves. Hence, in the following tracking evaluation we use the posterior curve for *GarbageTruck* which is approximately the mean in every case.

5.2 Evaluation of the Object Pose Tracking system

For evaluation of tracking we place each object on a ground truth pattern and capture a trajectory covering different orientations and scales (see Fig. 2). Each sequence is annotated with the ARToolKit [29] by detecting the camera pose of about 600 frames per object. Then different interest point detectors, descriptors and combinations of them are used to track the object pose. To compare the results we compute the precision

$$p_{pr} = \frac{n_{tp}}{n_{tp} + n_{fp}}, \tag{3}$$

where n_{tp} is the number of true object detections and respectively n_{fp} is the number of false detected objects. To decide which object pose is correct and which is false we compare the tracked pose with the pose computed with the ARToolKit and use an inlier threshold $t_{detection} = 40mm$ which is about half of the object size. In addition to the precision we record the pose accuracy in x, y and z direction (camera coordinate system) and the overall computation time per frame.

Results are shown in Table 1. It can be seen, that the MultiFAST detector in combination with our ImGD achieves the highest precision as well as the highest frame rat (for the Horse up to 20fps). An interesting insight is, that SIFT has almost the worst matching probability curve (see Fig. 3) but it achieves the second best tracking result. As proposed by Lowe [1], this motivates using the second nearest neighbor ratio to prune weak matches instead of using a fixed threshold. The results in Table 1 also indicate, that the performance can (slightly) be increased if different interest point types are combined. No improvement has been achieved for combinations with *MultiFAST/ImGD* where the precision is already at a very high level when using it alone.

6 Conclusion

We presented a system for real time object pose tracking. By learning probabilistic confidence measures for interest points the proposed system is able to

Table 1. Tracking evaluation using different detector and descriptor combinations.

detector / descriptor	precision	x [mm]	y [mm]	z [mm]	time [ms]
Harris/ImGD	0.59	4.6	5.6	18.9	83
MSER/ImGD	0.38	3.2	4.0	10.4	106
MultiFAST/ImGD	0.98	2.1	2.2	13.8	62
SIFT/SIFT	0.89	2.2	3.2	12.6	163
SURF/SURF	0.47	3.1	5.6	19.2	100
SURF/ImGD	0.81	3.3	3.8	14.4	77
Harris/ImGD + MSER/ImGD	0.60	3.9	7.5	18.5	172
Harris/ImGD + SURF/SURF	0.62	4.7	6.7	17.3	159
Harris/ImGD + SURF/ImGD	0.82	3.2	4.0	15.1	130
MultiFAST/ImGD + MSER/ImGD	0.98	2.0	2.2	13.4	149
MultiFAST/ImGD + SURF/SURF	0.98	2.2	2.3	13.4	134
MultiFAST/ImGD + SURF/ImGD	0.98	2.0	2.1	12.2	108

integrate complementary features in a common tracking framework. This allows to handle a broad variety of objects regardless of their appearance and shape. We evaluate state of the art interest points and compare them to our own image gradient histogram descriptor (ImGD). Results show that a multi scale FAST detector in combination with our ImGD outperforms all other detector/descriptor combinations, which could be only slightly improved by combining it with another feature type. Future work will explore the possibly more pronounced improvements using more complementary feature types beyond the point-like features above (e.g. object contours), in handling even broader object classes such as partly transparent or shiny objects.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2) (2004) 91–110
2. Comparison of Affine-Invariant Local Detectors and Descriptors. In: European Signal Processing Conference. (2004)
3. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. IJCV **65**(1) (2005) 43–72
4. Rosten, E., Porter, R., Drummond, T.: Faster and better: A machine learning approach to corner detection. PAMI **32** (2010) 105–119
5. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing **22**(10) (2004) 761 – 767
6. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). CVIU **110**(3) (2008) 346–359
7. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: ICCV. (2011) 2564 –2571

8. Mörwald, T., Zillich, M., Prankl, J., Vincze, M.: Self-monitoring to improve robustness of 3d object tracking for robotics. In: IEEE International Conference on Robotics and Biomimetics (ROBIO). (2011) 2830 –2837
9. Harris, C., Stephens, M.: A combined corner and edge detector. In: In Proc. of Fourth Alvey Vision Conference. (1988) 147–151
10. Drummond, T., Cipolla, R.: Real-Time Visual Tracking of Complex Structures. PAMI **24**(7) (2002) 932–946
11. Comport, A.I., Kragic, D., Marchand, E., Chaumette, F.: Robust Real-Time Visual Tracking: Comparison, Theoretical Analysis and Performance Evaluation. In: ICRA. (2005)
12. Babenko, B., Member, S., Yang, M.h., Member, S.: Robust Object Tracking with Online Multiple Instance Learning. PAMI **33**(8) (2011) 1619–1632
13. Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. IJCV **29**(1) (1998) 5–28
14. Klein, G., Murray, D.: Full-3D Edge Tacking with a Particle Filter. In: BMVC. Volume 3. (2006) 1119–1128
15. Cai, Y., Freitas, N.D., Little, J.J.: Robust Visual Tracking for Multiple Targets. ECCV (2006) 107–118
16. Choi, C., Christensen, H.I.: 3D textureless object detection and tracking: An edge-based approach. In: IROS. (2012) 3877–3884
17. Chestnutt, J., Kagami, S., Nishiwaki, K., Kuffner, J., Kanade, T.: GPU-Accelerated Real-Time 3D Tracking for Humanoid Locomotion. In: IROS. (2007)
18. Sánchez, J.R., Álvarez, H., Borro, D.: Towards Real Time 3D Tracking and Reconstruction on a GPU using Monte Carlo Simulations. In: ISMAR. (2010) 185–192
19. Masson, L., Jurie, F., Dhome, M.: Contour/texture approach for visual tracking. In: Scandinavian Conference on Image Analysis (SCIA). (2003) 661–668
20. Vacchetti, L., Lepetit, V., Fua, P.: Combining Edge and Texture Information for Real-Time Accurate 3D Camera Tracking. In: ISMAR. (2004)
21. Kyrki, V., Kragic, D.: Integration of model-based and model-free cues for visual object tracking in 3D. In: ICRA. (2005) 1566–1572
22. Pressigout, M., Marchand, E.: Real-time Hybrid Tracking using Edge and Texture Information. IJRR **26**(7) 689–713
23. Hebert, P., Hudson, N., Ma, J., Howard, T., Fuchs, T., Bajracharya, M., Burdick, J.: Combined shape, appearance and silhouette for simultaneous manipulator and object tracking. In: ICRA. (2012) 2405–2412
24. Haralick, R., Joo, H., Lee, C., Zhuang, X., Vaidya, V., Kim, M.: Pose estimation from corresponding point data. IEEE Transactions on Systems, Man and Cybernetics **19**(6) (1989) 1426–1446
25. Zillich, M., Prankl, J., Mörwald, T., Vincze, M.: Knowing your limits - self-evaluation and prediction in object recognition. In: IROS. (2011) 813 –820
26. Engelhard, N., Endres, F., Hess, J., Sturm, J., Burgard, W.: Real-time 3d visual slam with a hand-held camera. In: Proc. of the RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum, Vasteras, Sweden (2011)
27. Doucet, A., de Freitas, N., Gordon, N.: Sequential Monte Carlo Methods in Practice. Statistics for Engineering and Information Science Series. Springer (2001)
28. Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: CVPR. (2012)
29. Kato, H., Billinghurst, M.: Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In: IEEE/ACM International Workshop on Augmented Reality (IWAR). 85–94

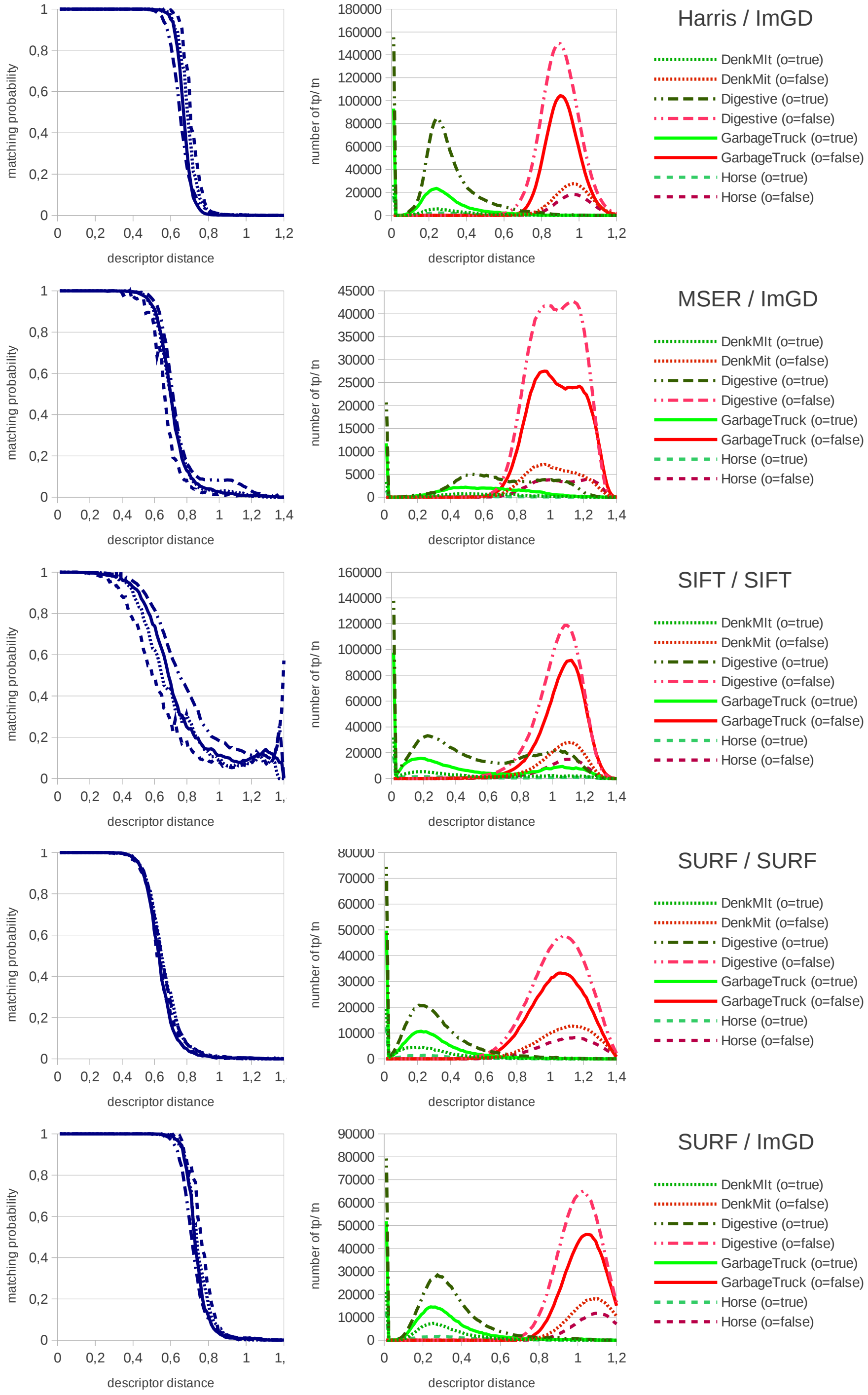


Fig. 3. Matching probability (left column) and probability density (right column) for different detector/descriptor combinations.