Boosted Edge Orientation Histograms for Grasping Point Detection*

Leonidas Lefakis†
IDIAP Research Center
Martigny, Switzerland
leonidas.lefakis@idiap.ch

Horst Wildenauer
Institute for Computer Aided Automation
Vienna University of Technology
wilde@prip.tuwien.ac.at

Manuel Pascual García-Tubío
Automation & Control Institute
Vienna University of Technology
mp@acin.tuwien.ac.at

Lech Szumilas
Ind. Research Inst. for Automation & Measurement
Al. Jerozolimskie 202, 02-486 Warsaw, Poland
lszumilas@piap.pl

Abstract

In this paper, we describe a novel algorithm for the detection of grasping points in images of previously unseen objects. A basic building block of our approach is the use of a newly devised descriptor, representing semi-local grasping point shape by the use edge orientation histograms. Combined with boosting, our method learns discriminative grasp point models for new objects from a set of annotated real-world images. The method has been extensively evaluated on challenging images of real scenes, exhibiting largely varying characteristics concerning illumination conditions, scene complexity, and viewpoint. Our experiments show that the method works in a stable manner and that its performance compares favorably to the state-of-the-art.

1. Introduction

In this work, we focus on mining monocular vision input to detect potential points for robotic grasping of previously unseen objects. Grasping of novel objects using vision input is among the most challenging and difficult problems in robotic research. In the past, approaches assumed a-priori knowledge about objects, or in the case of previously unseen objects, relied on the extraction of sufficiently complete 3-d models eg. by using stereopsis, see [14] for an overview. However, in realistic scenarios, where objects are occluded and only partially visible, or do not exhibit enough texture for stereo-based reconstruction, the latter are likely to fail.

Only recently, Saxena et al. [13] presented a promising approach capable of grasping previously unseen objects (classes) purely based on vision. Their local, image-based, grasp point representation is learned from artificially created images of object examples and are separately searched for in pairs of stereo images. Then, only image locations with a high confidence of being a grasp point are triangulated to infer the 3D-position were the object can be grasped - thus avoiding the need of reconstructing the object’s 3D shape. In [4] the authors find grasping points by describing the global object shape using shape context [3]. However, as shape context is known to perform poorly in cluttered scenes [16] the work relies on high quality figure-
Figure 2. Illustration of the descriptor. Probes (circles) are radially arranged around the center (dot). Each probe pools the edge density separately from underlying orientation channels and stacks them in a histogram. Here, only four out of eight edge orientation channels are shown.

ground segmentation, achieved with an active stereo setup, and knowledge about the workspace in which objects are placed.

Our approach is motivated by the existence of similar semi-local object parts in objects that themselves have rather dissimilar shapes. A typical example is the presence of handles in a large variety of objects ranging from scissors to jugs. In that sense, our method is similar in spirit to [13]. However, by encoding shape information of semi-local structures around grasp points, we arrive at discriminative representations which are able to ignore image clutter to a larger extent.

The contribution of our work is twofold: 1) We devised a novel image descriptor based on radially configured edge orientation histograms. The descriptor is simple to implement, efficient, and can be easily extended to include more cues such as color or texture. 2) In contrast to preceding work utilizing artificially created data, we demonstrate that discriminative grasp point representations can be learned from images of real scenes. Experiments on a challenging data set show that our method is able to significantly outperform the state of the art.

2 Method

The approach consists of two stages: 1) Discriminative grasp points models are learned from annotated grasp points in real images. For this, our novel descriptor is employed, which is able to efficiently encode the grasp point’s shape and its semi-local context. 2) In the detection phase, an input image is scanned densely over a range of scales using the learned model. On the resulting scale-space response maps, mean-shift mode seeking is employed to find the position and scale of potential grasp points. A typical result obtained with our method is depicted in Fig. 1.

2.1 Grasp Point Representation

Our grasp point representation is an extension to Carmichael’s [5] shape descriptor using a circular arrangement of edge probes. Each of these probes captures the density of the underlying edge image by weighted integration in a gaussian-shaped receptive field. Borrowing the idea from [17], we extend the descriptor to operate on channel images obtained from any orientation selective feature detector or filter. Specifically, having an input image \( I \), we compute a number \( C \) of blurred orientation channels \( G_{\sigma_p}^o \text{ } = \text{ } G_{\sigma_p} \ast C_o, \quad o = 1 \ldots C \), one for each discretized orientation. Here, the channel image \( C_o \) is the component of the feature detector’s output for direction \( o \). \( G_{\sigma_p} \) denotes a Gaussian kernel with standard deviation \( \sigma_p \) and \( \ast \) stands for convolution. Probe values at image location \( (x, y) \) for orientation \( o \) can be now efficiently obtained by simply accessing \( G_{\sigma_p}^o(x, y) \) which equals the pooled oriented response at that position. By stacking all channel-values for one probe location into a vector, a \( C \)-dimensional orientation histogram \( p \) is obtained.

Surrounding a probe at the query position, additional probes are located on \( K \) concentric circles with radii \( r_k = k\sigma_p, k = 1..K \). Each circle is populated with an increasing number of \( 6k \) evenly spaced probes, see Fig. 2.

2.2 Learning

Here, we utilize the MCBoost algorithm [11] to build \( L \) separate so-called strong classifiers, each com-
bining the outputs of weak classifiers. The responses $H_l$ of the multiple strong classifiers are then aggregated using a Noisy-OR framework:

$$P(x) = 1 - \prod_l (1 - P_l(x)), \quad (1)$$

where the probability $P_l$ of the classifier $l$ is given by

$$P_l(x) = \frac{1}{1 + e^{-H_l(x)}}, \quad (2)$$

The clustering nature of this approach, is especially relevant in our application as our positive training examples belong to diverse object classes. Given the weight update rule for each boosted classifier $l$:

$$w_{li} = \frac{y_i - P(x_i)}{P(x_i)} \frac{P_l(x_i)}{P_l(x_i)}, \quad (3)$$

the learning algorithm effectively co-clusters the data, dividing the positive examples into subsets, one for each classifier $H_l$. The algorithm is thus allowed to concentrate for each of these subsets on the most informative features, ignoring the remaining positive examples.

The weak learners beeing the base of the multiple strong classifiers, have the form of regression stumps [10] built from individual probe-based orientation histograms. For each strong classifier, at each boosting round, we run weighted Linear Discriminant Analysis (wLDA) [12] on the vectors formed by the bins of the orientation histograms for each probe position in the descriptor. The histogram-vectors are then projected onto the normal $w$ of the discriminant and regression stumps are fitted to the resulting scalars.

After $M$ rounds of boosting, each final classifier $l$ has the form of:

$$H_l = \sum_{m=1}^M a_{ml}(w_{ml}^T \mathbf{D} > th_{ml}) + b_{ml}, \quad (4)$$

where $a_{ml}$, $b_{ml}$, $th_{ml}$ are the parameters of the best weak classifier, $a_{ml}$ is calculated using line search, and $w_{ml}$ is returned by wLDA - all at round $m$ and classifier $l$. $\mathbf{p}$ is the histogram described in Sec. 2.1.

At training time, positive examples are obtained by extracting the descriptor at the center of grasping regions normalized to a canonical radius of 7 pixels. To increase the positive sample set, we augment it with randomly (in small ranges) translated, scaled, and rotated descriptor versions [12].

To obtain negative examples, descriptors are taken at random from the background of training images. For positions close to the grasping region the classifier is often not able to construct adequate discriminative models based on the randomly chosen negative examples. To counter this, we provide additionally negative examples near the grasping region [15]. In particular, we use positions located on circles centered at the grasp points, with a radius 1.5 times of that of the grasping region.

Once the initial detector is learned, one can bootstrap the gathering of hard examples [6]. We scan the training set (see Sec. 2.3) for misclassifications and inject them into the training set for full retraining.

2.3 Detection

Grasp points are found by a sliding window approach commonly used in object detection frameworks [18]. We scan images in a range of predefined scales $\{s^k\}, k = 1 \ldots K$. Specifically, for an image at scale $s^k$, one proceeds as follows: 1) Canny edges are computed and the components are distributed over the $C$ channel images according to their orientation. The resulting maps are then smoothed by a Gaussian kernel to obtain blurred channel images $G^{x}_{s^k}$, see Sec. 2.1. 2) At each scale $s^k$ and position $(x, y)$ the classifiers are evaluated and we obtain the $L$ classifiers’ confidences
\( H_l(x, y, s^k) \) which we convert to the posterior probabilities of a grasp point presence using the logistic transform (cf. [9]):

\[
P(grasp\_point(x, y, s^k)) = 1 - \prod_l \left( \frac{1}{1 + e^{-H_l(x, y, s^k)}} \right)
\]

(5)

For a confidence map computed in such way, we refer the reader to Fig. 1. To find the location and scale of grasp points, mean-shift mode estimation is adopted as described by Shotton et al. [15].

3. Experimental setup

We compiled a challenging dataset containing images of 3 object categories. The collection consists of 630 images, of which 210 show mugs, 210 bottles, and 210 Martini glasses, 30 of the mug images and 30 bottle images were taken from the database of Ferrari et al. [8], the remainder was found by Google image search. The images exhibit viewpoint changes, considerable background clutter and often more than one object instance and class are present, see Fig. 3. The number of annotated objects totaled 720. Grasp points are represented by circular regions giving position and approximate scale of the relevant structure. Two grasp points were selected for each mug - one at the top of the handle and one in the middle. Martini glass grasp points are located at the upmost part of the shaft, bottles were annotated by the top of the neck. Overall, 956 grasp points have been annotated. In addition, each object instance is provided with a bounding box, designating the class of associated grasp points. Fig. 4 shows examples of annotated object instances and grasp points.

The dataset is split into two equally sized sets for training and testing. Grasp points in test images exhibit a scale range of roughly \( 3 \times \) from smallest to largest. Given a minimum confidence threshold, detections are regarded as correct if the circular region of the inferred grasp point \( r_{inf} \) agrees sufficiently with the ground truth \( r_{gt} \), checked by the symmetric overlap criterion \( \frac{\text{Area}(r_{inf} \cap r_{gt})}{\text{Area}(r_{inf} \cup r_{gt})} > 0.25 \) similar to [1].

4. Results

For all tests reported, the descriptor parameters were set to \( \sigma_p = 5 \) and \( K = 5 \), determined by cross-validation. We also experimented with orientation quantization levels and found that \( C = 4 \) channel images and ignoring direction by mapping the orientation into the range \([0, \pi)\) worked best.

We compared our method with the approach suggested in [13], based on Laws masks and color histograms. Since the originally proposed logistic regression algorithm performed poorly, we also present improved results using our boosting framework. Furthermore, additional to our edge-based method, we evaluated a modification adopting gradients computed with Sobel filters. As can be seen from the precision-recall curves [7, 2] depicted in Fig. 5, our method achieves significantly better results. Fig. 6 shows some example detections taken from the test set. As one can see, the detector successfully predicts the locations of grasp regions, showing only a small number of false positives. Finally, we tested our algorithm on images showing object classes not contained in the training set, see Fig. 7. The handles on the jar were detected as they resemble the mug handles - this can also be observed in the case of scissors. Additionally, the approach is able to detect not immediately apparent similarities between the flower stems and martini glass shafts.
Figure 6. Detection examples: Successful detections (red) and false positives (blue).

Figure 7. Meaningful detections (red) for classes not contained in the training set.

5. Conclusions

We presented a learning-based method for detecting grasp points in images of newly seen objects. Extensive tests have shown that our approach based on boosted histograms outperforms the state-of-the-art. It was demonstrated that the approach is capable of capturing grasping relevant information, achieving promising results on familiarly shaped object from classes not contained in the training set.

Currently we investigate extensions to automatically determine optimal blurring scale and descriptor aperture and the use of other cues. Integration in a stereo setup similar to the one presented in [13] is planned.

References