International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XXXIX-B3, 2012
XXII ISPRS Congress, 25 August – 01 September 2012, Melbourne, Australia

# INTENSITY AND RANGE IMAGE BASED FEATURES FOR OBJECT DETECTION IN MOBILE MAPPING DATA

**Richard Palmer[1], Michael Borck[1], Geoff West[1] and Tele Tan[2]**

[1]Department of Spatial Sciences, [2]Department of Computing
Curtin University, GPO Box U1987, Perth 6845, Western Australia
{r.palmer, michael.borck}@postgrad.curtin.edu.au, {g.west, t.tan}@curtin.edu.au
Cooperative Research Centre for Spatial Information

**Commission III/4**

**KEY WORDS:** low-level features, image processing, point clouds, mobile and terrestrial mapping, 3-D features, 2-D features

**ABSTRACT:**

Mobile mapping is used for asset management, change detection, surveying and dimensional analysis. There is a great desire to automate these processes given the very large amounts of data, especially when 3-D point cloud data is combined with co-registered imagery - termed "3-D images". One approach requires low-level feature extraction from the images and point cloud data followed by pattern recognition and machine learning techniques to recognise the various high level features (or objects) in the images. This paper covers low-level feature analysis and investigates a number of different feature extraction methods for their usefulness. The features of interest include those based on the "bag of words" concept in which many low-level features are used *e.g.* histograms of gradients, as well as those describing the saliency (how unusual a region of the image is). These mainly image based features have been adapted to deal with 3-D images. The performance of the various features are discussed for typical mobile mapping scenarios and recommendations made as to the best features to use.

## 1 INTRODUCTION

Laser scanning is currently the averred method for the collection of surveying/mapping data but increasingly this is being augmented by 2-D imaging cameras. Co-registration of 2-D colour intensity maps collected from standard cameras with range measurements collected by laser scanners results in the creation of *3-D images*; 2-D images with every pixel having an associated range value. Recently, mapping systems based on stereoscopic imaging techniques have been used to produce similar 3-D images at the expense of reduced accuracy in range. The increasing use of mobile mapping systems based around such technology is resulting in the creation of very large amounts of data; mobile mapping systems operating along roads in urban centres typically collect full 360 degree panoramas every five or ten metres along the vehicle track. These datasets are very useful for a range of content analysis applications, but the speed of analysis is severely limited by the amount of costly and impractical manual processing needed to identify interesting features. There is a great need to improve upon the automated detection of content that is of interest to the user, so that a large proportion of time is not wasted looking through irrelevant data.

Processing data for the automatic identification of features or objects of interest is a core focus of computer vision research. Research has focussed on the analysis of very large cohorts of images because many people and organisations produce and share images and these must often be indexed and organised according to content. Websites such as Flickr (http://www.flicker.com) and Picasa (http://picasa.google.com), and the need to search the Web for images having specific content means many millions of images must be processed. Mobile mapping imagery requires similar processing to discover content for use in application areas such as asset management, change detection, surveying and dimensionality analysis. Mobile mapping data is distinct from regular 2-D imagery because of the availability of co-registered range information. This extra modality presents an interesting avenue for research because it offers the possibility of significantly increasing the speed and accuracy of existing 2-D image based feature detection methods.

Research into object detection has produced a large number of novel approaches to feature detection. The performance of features extracted from imagery is evaluated for a particular object detection task. This requires a task driven approach to the evaluation of features by first identifying the type of object in the imagery to be detected, before determining how accurate the object detection system that uses these features is in detecting the objects. Typically this requires much imagery with ground-truthed bounds defined around the objects to be detected. While there exists much intensity imagery (*e.g.* the PASCAL Visual Object Classes Challenge (Everingham et al., 2010)), there are no similar commonly available 3-D or range image datasets.

For the purposes of this research, a dataset from Earthmine was used consisting of a sequence of panoramas taken approximately every ten metres along the road within the Perth CBD, Western Australia. Each panorama consists of eight images projected onto the inside of a cube centred on the imaging camera array mounted on the mapping vehicle. Within the high resolution images, each colour pixel has co-registered against it the real world latitude, longitude and elevation at that point. 3-D images can be generated specifying the colour and range of each pixel in the image.

Range image data has been used for object representation and to establish correspondences between an object's geometric model (*e.g.* derived from a generic CAD model of the object) and the object's representation in the range imagery (Arman et al., 1993), (Lavva et al., 2008), (Steder et al., 2009). However, due to the complexity and slowness of matching spatial models in range imagery, and the wide availability of intensity imagery, research has favoured extracting the appearance of an object to encode its discriminative qualities. In intensity images, *keypoints* or *interest points* have been proposed such as Harris keypoints (Harris and Stephens, 1988), SIFT (Lowe, 2004), SURF (Bay et al., 2006), and FAST (Rosten, 2006); blob detectors such as Maximally Stable Extremal Regions (MSER) (Matas et al., 2002),

and image operators such as the Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005) that encapsulates changes in the magnitude and orientation of contrast over a grid of small image patches. HOG features have shown excellent performance in their ability to recognise a range of different object types including natural objects as well as more artificial objects (Dalal and Triggs, 2005), (Felzenszwalb et al., 2010), (Schroff et al., 2008).

The success of such appearance based feature detection methods for intensity images led to the development of similar appearance based features for range images. Spin images (Johnson and Hebert, 1999) use 2-D histograms rotated around a reference point in space. Splash features (Stein and Medioni, 1992) are similar to HOG features in that they collect a distribution of surface normal orientations around a reference point. NARF (Normal Aligned Radial Feature) features (Steder et al., 2010) detect stable surface regions combined with large depth changes in object borders. The feature is designed to be stable across different viewpoints. Tripod operators (Pipitone and Adams, 1993) compactly encode surface shape information of objects by taking surface range measurements at the three corners of an equilateral triangle. Other range based descriptors include surface patch representations (Chen and Bhanu, 2007), surface normal based signatures (Li and Guskov, 2007), and tensor-based descriptors (Mian et al., 2006). However, for the most part, there is still little evidence that any of these range image based features are significantly better than any others for specific object detection tasks. Recent work has combined intensity based features with range to first segment images into planar range regions before using this information to guide the object detection process with intensity based features (Rapus et al., 2008), (Wei et al., 2011).

This paper reports upon a number of low-level feature extraction methods for their usefulness in describing salient image regions containing higher-level features/objects. The features of interest include those based on the "bag of words" concept in which many low-level features are used together to model the characteristics of an image region in order to measure the saliency relative to the whole image (section 2). These generate response maps indicating regions of interest in the image. Feature extraction methods that encode a greater amount of spatial and geometric information from range and intensity image regions are discussed in the context of their use in parts-based models for higher-level object detection (section 3). Extraction of rudimentary line segment information from the 3-D images for use in detecting and modelling/matching object geometry is also discussed. The paper concludes with a summary of future work and known issues to be addressed (section 4).

## 2 OBJECT SALIENCY

Given the large amount of data to be processed, it is necessary to first extract candidate regions with greater likelihood of containing higher-level features of interest. For a given object detection task, such as finding all bus shelters, a saliency detection method is required to return all approximate locations of bus shelters in the data. A consequence of this is a high false alarm rate. Subsequent processing is then more efficient because the total remaining amount of data is substantially reduced.

Some low-level features are more suited than others for discrimination between high-level features of interest. It is recognised that it is not possible to find a combination of one or more features that will detect all high-level features of interest. The selection of low-level features must be task driven; the objects to be detected must first be specified so that the combination of features that is most appropriate for the matching of such objects can be used. Machine learning approaches using a training set of manually chosen instances of the high-level features (positive examples) as well as instances of other high-level features not of the required class (negative examples) will determine the best low-level features to use and how they can be combined to satisfy the task.

### 2.1 Statistical based features

The statistical based features capture the scale invariant covariance of object structure. The histogram of an image is a plot of the number of pixels for each grey level value (or intensity values of a colour channel for colour images). The shape of the histogram provides information about the nature of the image (or a sub-region of the image). For example, a very narrow histogram implies a low contrast image, while a histogram skewed toward the high end implies a bright image, and a bi-modal histogram (or a histogram with multiple strong peaks) can imply the presence of one or more objects.

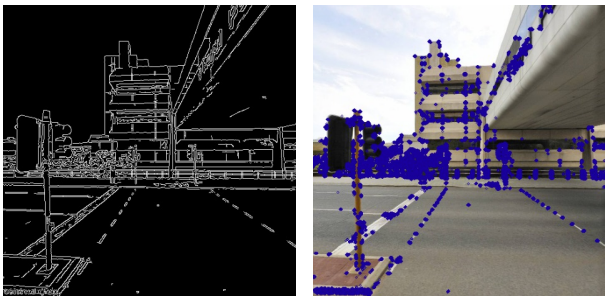
The histogram features considered in this paper are statistically based in that the histogram models the probability distribution of intensity levels in the image. These statistical features encode characteristics of the intensity level distribution for the image. A bright image will have a high mean and a dark image will have a low mean. High contrast regions have high variance, and low contrast images have low variance. The skew is positive when the tail of the histogram spreads out to the right (positive side), and is negative when the tail of the histogram spreads out to the left (negative side). High energy means that the number of different intensity levels in the region is low *i.e.,* the distribution is concentrated over only a small number of different intensity levels. Entropy is a measure of the number of bits required to encode the region data. Entropy increases as the pixel values in the image are distributed among a larger number of intensity levels. Complex regions have higher entropy and entropy tends to vary inversely with energy.

### 2.2 Localised keypoint, edge and corner features

Rosin (2009) argues for the density of edges as a measure of salience because interesting objects have more edges. Edge features have been very popular and range from simple differential measures of adjacent pixel contrasts such as the Sobel (Duda et al., 1973), Prewitt (Prewitt, 1970), and Robert's Cross (Roberts, 1963) operators to complex operators such as the Canny (Canny, 1986) and the Marr-Hildreth (Marr and Hildreth, 1980) edge detectors. Canny produces single pixel wide edges allowing edge linking, and exhibits good robustness to noise (see figure 1(a)). The simpler operators such as Sobel require a threshold and thinning to obtain single pixel wide edges. Corner detectors such as Harris (Harris and Stephens, 1988) have been popular because they produce features expressing a high degree of viewpoint invariance (see figure 1(b)). However, many of the features detected by the Canny operator are false corners and so cannot be semantically interpreted. More recently, keypoint detectors with stronger robustness to viewpoint invariance that detect fewer false features have been proposed such as SIFT (Lowe, 2004), SURF (Bay et al., 2006) and FAST (Rosten, 2006).

### 2.3 Saliency based features

Saliency based features take inspiration from aspects of the human visual system. This is a task driven process that analyses global image features to identify image regions containing more



(a) Canny (b) Harris

Figure 1: Edge and corner keypoint detectors.

"interesting" pixels. The images in figure 2 show high responses for regions with many edges representing busyness in the images or changes in intensity or frequency components of the image.

(a) Frequency-tuned (Achanta et al., 2009) (b) Maximal Symmetric Surround (Achanta and Süsstrunk, 2010)

Figure 2: Examples of saliency detectors.

Segmentation is the process of partitioning the image into multiple segments. Edge based saliency maps are used to segment the images into interesting and non-interesting regions by simple thresholding. Figure 3 demonstrates how this procedure drastically reduces the area of the image expected to contain meaningful information about the objects of interest.

(a) Edge based saliency map (Rosin, 2009) (b) Mask applied to image showing interesting region.

Figure 3: Saliency based segmentation using simple thresholding on saliency map.

The methods described do not require any kind of offline preprocessing to use, however they are also weak at detecting salient image regions while maintaining a low false alarm rate for higher-level features of interest. Learning a model of saliency offline is a more promising method for detecting salient image regions.

### 2.4 Learning based features

In order to detect salient regions of an image, a model of saliency can be learned for comparison against new images from training data. The Support Vector Machine (SVM) is a method of supervised machine learning based on the theory of statistical learning (Cortes and Vapnik, 1995). The theory behind the SVM guarantees that any $N$ dimensional feature space is linearly separable in $N + M$ dimensions (where $M$ is not excluded from being possibly infinite). The SVM finds a separating hyperplane (in $N + M$ dimensional space) between two classes of training data (the positive and the negative examples). The placement of this hyperplane is such that the distances between the hyperplane and the closest training instances (the support vectors) on either side of the plane are maximised. Since noise in the training data cannot be avoided, the SVM is extended to incorporate a "soft-margin" around the hyperplane to allow training instance outliers to sit on the wrong side of the hyperplane. The complete learning algorithm seeks to maximise the distance of the support vectors to the hyperplane, while minimising the distances from the separating hyperplane of training instances found to be on the wrong side of the separating hyperplane. Finally, since training data isn't always linearly separable in the provided $N$ dimensional feature space, a kernel function can be used to place the data into a higher dimensionality feature space to increase the likelihood that a separating hyperplane with a good fit to the data can be found. The kernel function may be a high degree polynomial (or worse) on the training data, but this does not incur any extra processing overhead since the training data only ever appears as a dot product of vectors inside the kernel function. SVMs have demonstrated excellent performance in a number of similar studies (Felzenszwalb et al., 2010), (Dalal and Triggs, 2005), (Lin et al., 2011) concerning object detection.

Bounding boxes are positioned around examples of the objects to be identified in a set of training images (see figure 4). Features for these bounded regions are calculated and then concatenated as $N$-dimensional feature vectors (where $N$ is the number of features used) to generate a set of positive training examples. The same number of negative feature vectors are randomly generated (from image regions that do not contain the objects of interest). The positive and negative examples are passed to an SVM for training and five-fold cross-validation is performed, varying the parameters to the kernel functions of the SVM to identify an optimal model without overfitting to the training data (linear and radial basis functions are evaluated for their performance during cross-validation). The generated model represents a weight vector which is multiplied (as the scalar product) with a feature vector calculated from a new (previously unseen) image region to determine whether the image region is salient or not. The feature measurements explored in our approach consist of: Histogram of Orientations (over whole image sub-regions), edge density, Harris keypoint density, FAST keypoint density, mean depth of the range image (in the image sub-region), standard deviation of the intensity histogram, skew of the intensity histogram, energy of the intensity histogram, and entropy of the intensity histogram.

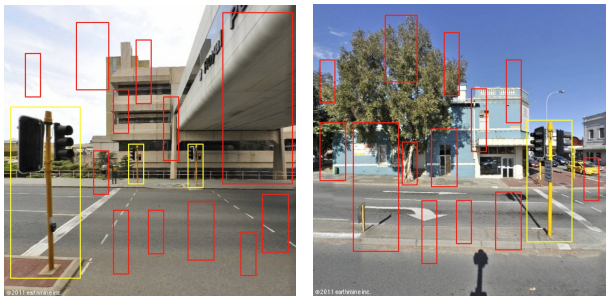


Figure 4: Training images displaying positive training instances (yellow) and negative instances (red).

Initial results are promising with a 85% correct identification rate

and a 20% false alarm rate using HOG features. Future work will explore different combinations of feature, SVM parameters and kernel functions to maximise the correct detection rate while minimising false alarms.

### 2.5 Range image based features

Each pixel of the high resolution intensity images has co-registered with it a range value. This allows a range map of the intensity image to be calculated which is used as an additional feature in the learning process described in section 2.4. The range map is also used to segment the intensity image as it is expected that objects of interest (*e.g.* street furniture) are located within a certain distance from the camera (the position of the camera is known *a priori*). The range map is thresholded to create a mask which is combined with the saliency response map created by any of the other methods to further reduce the area of the image to be passed to the next stage of processing. Figure 5 displays how the range map is combined with an edge-based saliency map to produce a final segmentation of the image.

(a) Range map (b) Mask applied to original image

Figure 5: Range based segmentation using simple thresholding on range map.

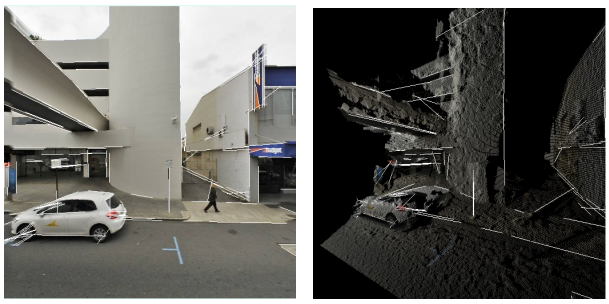
## 3 OBJECT DETECTION

The aim of an object detection system is to identify the category/class and location in an image of one or more objects of interest. The solution requires that the system internally represents models of the categories of object to be identified so that it can compare these models to locations in previously unseen images in order to identify when and if an instance of that model (an object) is present. Ideally one model should enable recognition of all such objects in a category, and be robust to the great variation of objects possible within a given category, as well as the great variation in how these objects may appear in an image (different viewpoints, different scales, varying lighting conditions). This means that the system must minimise the false negative detection rate. In addition, each model must be distinctive enough to preclude the possibility of confusing an instance of one model/class for another (or the null class representing no object). This is equivalent to minimising the false positive detection rate.

The presence of range data with mobile mapping data should theoretically allow for more accurate object identification because of the extra information available. In this paper, range information has been used to help segment the image into regions more likely to contain high-level features of interest. This range information can be further used to calculate geometric properties of the images and their content.

### 3.1 Object geometry

The identification of object edges, lines and corners can be used to infer the presence of straight lines or other geometric shapes in the image using feature extractors such as the Hough transform (Duda and Hart, 1972). If found together in non-random configurations, line features may be combined to form perceptual groups (Lowe, 1985). Once an object has been detected, such perceptual groups become doubly useful for the problem of object pose estimation. Figure 6 shows extraction of line information from a 3-D Earthmine image. These lines are first detected in the 2-D intensity image using a probabilistic version of the Hough transform to find line segments. Each point along a detected line is then queried against the co-registered range data. A line found in the 2-D image is rejected if the range along its length does not scale linearly. To allow for noise in the range information, a parameter specifies the degree of allowed range variation along the length of the line. The range points are fit to the 2-D lines using standard linear regression and end-points for the lines determined. It is possible to discriminate between edge type lines and intensity based lines by querying the linearity (in range) of short lines orthogonal to and crossing the detected line. Though providing quite a coarse estimation of scene geometry, these lines can later be used when comparing the geometric model of a learned class with detected objects to better approximate their locations in space.



(a) Hough lines in original intensity image (b) Line segments projected into space via linear regression in range

Figure 6: Line segments detected via Hough transform projected into 3-D space using linear regression in range.

### 3.2 Modelling Schema

Many objects (such as people or animals) are highly articulated and any model of their appearance or geometry must be able to cater for the wide range of pose variation intrinsic to these types of object. Non-natural objects often have fewer individually movable components and there is far less variation in how adjacent parts of the same object appear in relation to one another.

Methods based on pictorial structures (Felzenszwalb and Huttenlocher, 2005) and deformable parts-based models (Felzenszwalb et al., 2010) have demonstrated success in their ability to detect objects even when viewed from an unusual viewpoint, or when their parts are obscured due to occlusion with other scene elements or their location at the edge of an image. A model is a hierarchy of parts where a single part is the child of a root part having features computed at half the resolution of the parts. The placement of each of the parts in the model is conditionally independent of its sibling parts given its root. Figure 7 shows an example of a deformable parts model for the side and front view of a car using HOG features.

### 3.3 Detection Method

Modelling of independent object parts using HOG features has been used in this paper to detect cars in intensity images using a variant of the approach described by Felzenszwalb et al, (2010).

Figure 7: Example of deformable parts model using HOG features (Felzenszwalb et al., 2010)

The intensity image for testing is first scaled into an image pyramid of several different resolutions of HOG feature maps. At each scale of the feature pyramid, the root filter for a model of the object of interest is cross correlated with the feature map. This results in the generation of a response map for the root filter. This is repeated for each of the child parts using the feature map in the pyramid calculated at twice the resolution of the root filter. The detection process is performed independently for each part and the response maps for each part are transformed according to the best detection(s) of the root. Groupings of the detected parts that match learned anchor part positions in the car model are favoured over part configurations more distant from the learned anchors using a deformation cost function (the parameters of which are learned based on the observed variability of the parts in the training data). This produces an overall response map for complete root and part detections. The largest responses are thresholded and a bounding box calculated as the convex hull of a car's individual part detections. Finally, the scale of the bounding boxes for each detected object are rescaled and translated to match the original image dimensions. This places the part and root bounding boxes in the correct locations for the original image.

### 3.4 Results

Figure 8 shows results of detections of cars in the Earthmine intensity images. Thresholds were chosen manually in these results to determine the few best detections. The bounding box algorithm as used simply computes the convex hull of the object's parts. Better methods that consider the amount of deformation of a part as a factor to scale the position of the bounding box should result in more accurate object localisations.

## 4 CONCLUSIONS AND FUTURE WORK

This study has addressed two different stages in the feature detection process. In the initial stages when the relative proportion of data is high, accuracy is traded for speed in a coarse grained task driven approach to saliency detection and image segmentation. In the second stage, when the relative proportion of the data is lower, speed is traded for the more detailed processing required for the detection of particular objects. The proposed saliency detection method uses simple feature detectors working over the whole image in a sliding window approach to identify image sub-regions that are more likely to contain high-level features of interest. The output from the first stage is a response map which can be thresholded to identify image sub-region candidates for the second stage of processing that uses more complex feature vectors incorporating HOG style features derived from both the intensity and range imagery. The second stage detection process cross correlates these feature vectors with the image sub-region candidates provided from the first stage to identify promising object locations. A final stage (not discussed in this paper) will detect the pose of the detected objects by comparing the parts of the object with detected geometric features in the 3-D image.

Figure 8: Sample car detections in the Earthmine intensity images. Blue boxes denote individual part detections, while yellow boxes denote detection of whole object instances. Note the erroneous double detection of the car on the left of the image in the bottom right example.

### 4.1 Future Work

One of the biggest factors determining speed of detection is the requirement to evaluate all possible scales of an object in the intensity image. The addition of range information removes this need and the object's size can be learned along with model parameters.

Prior knowledge of how the data were collected and frequency and occurrence of low-level features extracted from the images in an offline processing step can be incorporated within a probabilistic framework to help guide the search for higher level (more complex) objects of interest. This can be considered a context dependent extension of our existing approach to detecting object saliency.

Finally, in future work, the effectiveness of the extended saliency and high-level feature detection methods will be tested against a larger set of 3-D images in order to assess the broader viability of the methods for object detection.

## ACKNOWLEDGEMENTS

This work is supported by the Cooperative Research Centre for Spatial Information, whose activities are funded by the Australian Commonwealth's Cooperative Research Centres Programme. It provides PhD scholarships for Michael Borck and Richard Palmer and partially funds Prof. Geoff West's position. The authors would like to thank John Ristevski and Anthony Fassero from Earthmine for making available the dataset used in this work.

## REFERENCES

Achanta, R. and Süsstrunk, S., 2010. Saliency detection using maximum symmetric surround. In: Proceedings of the 17th IEEE International Conference on Image Processing, IEEE, pp. 2653–2656.

Achanta, R., Hemami, S., Estrada, F. and Susstrunk, S., 2009. Frequency-tuned salient region detection. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 1597–1604.

Arman, F., Aggarwal, J. K. and Aggarvval, J. K., 1993. Model-based object recognition in dense-range images—a review. ACM Computing Surveys 25(1), pp. 5–43.

Bay, H., Tuytelaars, T. and Van Gool, L., 2006. Surf: Speeded up robust features. In: A. Leonardis, H. Bischof and A. Pinz (eds), Computer Vision ECCV 2006, Lecture Notes in Computer Science, Vol. 3951, Springer Berlin / Heidelberg, pp. 404–417.

Canny, J., 1986. A computational approach to edge detection. IEEE transactions on pattern analysis and machine intelligence 8(6), pp. 679–98.

Chen, H. and Bhanu, B., 2007. 3D free-form object recognition in range images using local surface patches. Pattern Recognition Letters 28(10), pp. 1252–1262.

Cortes, C. and Vapnik, V., 1995. Support-vector networks. Machine Learning 20, pp. 273–297. 10.1007/BF00994018.

Dalal, N. and Triggs, B., 2005. Histograms of Oriented Gradients for Human Detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Ieee, pp. 886–893.

Duda, R. O. and Hart, P. E., 1972. Use of the Hough transformation to detect lines and curves in pictures. Communications of the ACM 15(April 1971), pp. 11–15.

Duda, R. O., Hart, P. E. and Stork, D. G., 1973. Pattern Classification and Scene Analysis. 1 edn, Wiley-Interscience, New York, USA.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A., 2010. The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88(2), pp. 303–338.

Felzenszwalb, P. F. and Huttenlocher, D. P., 2005. Pictorial Structures for Object Recognition. International Journal of Computer Vision 61(1), pp. 55–79.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D. and Ramanan, D., 2010. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), pp. 1627–45.

Harris, C. and Stephens, M., 1988. A combined corner and edge detector. Alvey vision conference pp. 147–152.

Johnson, A. E. and Hebert, M., 1999. Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(5), pp. 433–449.

Lavva, I., Hameiri, E. and Shimshoni, I., 2008. Robust methods for geometric primitive recovery and estimation from range images. IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society 38(3), pp. 826–45.

Li, X. and Guskov, I., 2007. 3D object recognition from range images using pyramid matching. In: 2007 IEEE 11th International Conference on Computer Vision, Ieee, pp. 1–6.

Lin, Y., Lv, F., Zhu, S., Yang, M., Cour, T., Yu, K., America, N. E. C. L., Cao, L. and Huang, T., 2011. Large-scale image classification: fast feature extraction and SVM training. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, pp. 1689–1696.

Lowe, D. G., 1985. Perceptual organization and visual recognition. Vol. 5, Kluwer Academic Pub.

Lowe, D. G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), pp. 91–110.

Marr, D. and Hildreth, E., 1980. Theory of edge detection. Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain) 207(1167), pp. 187–217.

Matas, J., Chum, O., Urban, M. and Pajdla, T., 2002. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. bmvc pp. 384–393.

Mian, A. S., Bennamoun, M. and Owens, R., 2006. Three-dimensional model-based object recognition and segmentation in cluttered scenes. IEEE transactions on pattern analysis and machine intelligence 28(10), pp. 1584–601.

Pipitone, F. and Adams, W., 1993. Rapid recognition of freeform objects in noisy range images using tripod operators. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE Comput. Soc. Press, pp. 715–716.

Prewitt, J. M. S., 1970. Object enhancement and extraction. Academic Press, New York, USA.

Rapus, M., Munder, S., Baratoff, G. and Denzler, J., 2008. Pedestrian recognition using combined low-resolution depth and intensity images. In: Intelligent Vehicles Symposium, 2008 IEEE, pp. 632 –636.

Roberts, L. G., 1963. Machine Perception Of Three-Dimensional Solids. Phd, Lincoln Laboratory, Department of Electrical Engineering, Massachusetts Institute of Technology.

Rosin, P. L., 2009. A simple method for detecting salient regions. Pattern Recognition 42(11), pp. 2363 – 2371.

Rosten, E., 2006. Machine learning for high-speed corner detection. Machine Learning pp. 1–14.

Schroff, F., Criminisi, A. and Zisserman, A., 2008. Object class segmentation using random forests. In: Proceedings of the British Machine Vision Conference.

Steder, B., Grisetti, G., Van Loock, M. and Burgard, W., 2009. Robust on-line model-based object detection from range images. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Ieee, pp. 4739–4744.

Steder, B., Rusu, R. B., Konolige, K. and Burgard, W., 2010. NARF: 3D range image features for object recognition. In: Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS).

Stein, F. and Medioni, G., 1992. Structural indexing: Efficient 3D object recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 14(2), pp. 125–145.

Wei, X., Phung, S. L. and Bouzerdoum, A., 2011. Pedestrian sensing using time-of-flight range camera. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on, pp. 43 –48.