A Survey of Semantic Segmentation

Martin Thoma
info@martin-thoma.de

***Abstract*—This survey gives an overview over different techniques used for pixel-level semantic segmentation. Metrics and datasets for the evaluation of segmentation algorithms and traditional approaches for segmentation such as unsupervised methods, Decision Forests and SVMs are described and pointers to the relevant papers are given. Recently published approaches with convolutional neural networks are mentioned and typical problematic situations for segmentation algorithms are examined. A taxonomy of segmentation algorithms is given.**

I. Introduction

Semantic segmentation is the task of clustering parts of images together which belong to the same object class. This type of algorithm has several use-cases such as detecting road signs [MBLAGJ+07], detecting tumors [MBVLG02], detecting medical instruments in operations [WAH97], colon crypts segmentation [CRSS14], land use and land cover classification [HDT02]. In contrast, non-semantic segmentation only clusters pixels together based on general characteristics of single objects. Hence the task of non-semantic segmentation is not well-defined, as many different segmentations might be acceptable.

Several applications of segmentation in medicine are listed in [PXP00].

Object detection, in comparison to semantic segmentation, has to distinguish different instances of the same object. While having a semantic segmentation is certainly a big advantage when trying to get object instances, there are a couple of problems: neighboring pixels of the same class might belong to different object instances and regions which are not connected my belong to the same object instance. For example, a tree in front of a car which visually divides the car into two parts.

This paper is organized as follows: It begins by giving a taxonomy of segmentation algorithms in Section II. A summary of quality measures and datasets which are used for semantic segmentation follows in Section III. A summary of traditional segmentation algorithms and their characteristics follows in Section V, as well as a brief, non-exhaustive summary of recently published semantic segmentation algorithms which are based on neural networks in Section VI. Finally, Section VII informs the reader about typical problematic cases for segmentation algorithms.

II. Taxonomy of Segmentation Algorithms

The computer vision community has published a wide range of segmentation algorithms so far. Those algorithms can be grouped by the kind of data they operate on and the kind of segmentation they are able to produce.

The following subsections will give four different criteria by which segmentation algorithms can be classified.

This survey describes fixed-class (see Section II-A), single-class affiliation (see Section II-B) algorithms which work on grayscale or colored single pixel images (see Section II-C) in a completely automated, passive fashion (see Section II-D).

A. Allowed classes

Semantic segmentation is a classification task. As such, the classes on which the algorithm is trained is a central design decision.

Most algorithms work with a fixed set of classes; some even only work on binary classes like *foreground vs background* [RM07], [CS10] or *street vs no street* [BKTT15].

However, there are also unsupervised segmentation algorithms which do not distinguish classes at all (see Section V-B) as well as segmentation algorithms which are able to recognize when they don't know a class. For example, in [GRC+08] a **void class** was added for classes which were not in the training set. Such a void class was also used in the MSRCv2 dataset (see Section III-B2) to make it possible to make more coarse segmentations and thus having to spend less time annotating the image.

B. Class affiliation of pixels

Humans do an incredible job when looking at the world. For example, when we see a glass of water standing on a table we can automatically say that there is the glass and behind it the table, even if we only had a single image and were not allowed to move. This means we simultaneously two labels to the coordinates of the glass: Glass and table. Although there is much more work being done on **single class affiliation** segmentation algorithms, there is a publication about **multiple class affiliation** segmentation [LRAL08]. Similarly, recent publications in pixel-level object segmentation used layered models [YHRF12].

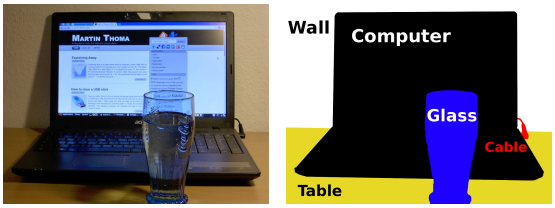
C. Input Data

The available data which can be used for the inference of a segmentation varies by application.

- **Grayscale vs colored**: Grayscale images are commonly used in medical imaging such as magnetic resonance (MR) imaging or ultrasonography whereas colored photographs are obviously widespread.
- **Excluding or including depth data**: RGB-D, sometimes also called range [HJBJ+96] is available in robotics, autonomous cars and recently also in consumer electronics such as Microsoft Kinect [Zha12].
- **Single image vs stereo images vs co-segmentation**: Single image segmentation is the most wide-spread kind of segmentation, but using stereo images was already tried in [BVZ01]. It can be seen as a more natural way of segmentation as most mammals have two eyes. It can also be seen as being related to having depth data.
  Co-segmentation as in [RMBK06], [CXGS12] is the problem of finding a consistent segmentation for multiple images. This problem can be seen in two ways: One the one hand, it can be seen as the problem of finding common objects in at least two images. On the other hand, every image after the first can be used as an additional source of information to find a meaningful segmentation. This idea can be extended to time series such as videos.
- **2D vs 3D**: Segmenting images is a 2D segmentation task where the smallest unit is called a *pixel*. In 3D data, such as volumetric X-ray CT images as they were used in [HHR01], the smallest unit is called a voxel.

D. Operation state

The operation state of the classifying machine can either be **active** as in [SUM+11], [SSA12] where robots can move objects to find a segmentation or **passive**, where the received image cannot be influenced. Among the passive algorithms, some segment in a completely **automatic** fashion, others work in an **interactive** mode. One example would be a system where the user clicks on the background or marks a coarse segmentation and the algorithm finds a fine-grained segmentation. [BJ00], [RKB04], [PS07] describe systems which work in an interactive mode.



(a) Example Scene (b) Visualization of a found segmentation

Figure 1: An example of a scene and a possible visualization of a found segmentation.

III. Evaluation and Datasets

A. Quality measures for evaluation

A performance measure is a crucial part of any machine learning system. As users of a semantic segmentation system expect correct results, the accuracy is the most commonly used performance measure, but there are other measures of quality which matter when segmentation algorithms are compared. This section gives an overview of those quality measures.

1) Accuracy: Showing the correctness of the segmentation hypotheses is done in most publications about semantic segmentation. However, there are a couple of different ways how this accuracy can be displayed. One way to give readers a first qualitative impression of the obtained segmentations is by showing examples such as Figure 1.

However, this can only support the explanation of particular problems or showcase special situation. For meaningful information about the overall accuracy, there are a couple of metrics how accuracy can be defined.

For this section, let $k \in \mathbb{N}$ be the number of classes, $n_{ij} \in \mathbb{N}_0$ with $i, j \in 1, \ldots, k$ be the number of pixels which belong to class i and were labeled as class j. Let $t_i = \sum_{j=1}^{k} n_{ij}$ be the total number of pixels of class i.

One way to compare segmentation algorithms is by the pixel-wise accuracy of the predicted segmentation as done in many publications [SWRC06], [CP08], [LSD14]. This is also called per-pixel rate and defined as $\frac{\sum_{i=1}^{k} n_{ii}}{\sum_{i=1}^{k} t_i}$. Taking the pixel-wise classification accuracy has two major drawbacks:

P1 Tasks like segmenting images for autonomous cars have large regions which have one class. This makes achieving classification accuracies of more than 30 % with a priori knowledge only possible. For example, a system might learn that a certain position of the image is most of the time “sky” while another position is most of the time “road”.

P2 The manually labeled images could have a more coarse labeling. For example, a human classifier could have labeled a region as "car" and the algorithm could have split that region into the general "car" and the more specific "wheel of a car"

Three accuracy metrics which do not suffer from problem P1 are used in [LSD14]:

- *mean accuracy*: $\frac{1}{k} \cdot \sum_{i=1}^{n} \frac{n_{ii}}{t_i}$
- *mean intersection over union*: $\frac{1}{k} \cdot \sum_{i=1}^{k} \frac{n_{ii}}{t_i + \sum_{j=1}^{k} n_{ji} - n_{ii}}$
- *frequency weighted intersection over union*: $\left(\sum_{p=1}^{k} t_p\right)^{-1} \sum_i \frac{t_i n_{ii}}{t_i + \sum_{j=1}^{k} n_{ji} - n_{ii}}$

Another problem might be pixels which cannot be assigned to one of the known classes. For this reason, [SWRC06] makes use of a void class. This class gets completely ignored for all quality measures. Hence the total number of pixels is assumed to be width $\cdot$ height $-$ number of void pixels.

One way to deal with problem P1 and problem P2 is giving a *confusion matrix* as done in [SWRC06]. However, this approach is not feasible if many classes are given.

The F-measure is useful for binary classification task such as the KITTI road segmentation benchmark [FKG13] or crypt segmentation as done by [CRSS14]. It is calculated as "the harmonic mean of the precision and recall" [PH05]:

$$F_\beta = (1+\beta)^2 \frac{\mathrm{tp}}{(1+\beta^2) \cdot \mathrm{tp} + \beta^2 \cdot \mathrm{fn} + \mathrm{fp}}$$

where $\beta = 1$ is chosen in most cases and `tp` means *true positive*, `fn` means *false negative* and `fp` means *false positive*.

Finally, it should be noted that a lot of other measures for the accuracy of segmentations were proposed for non-semantic segmentation. One of those accuracy measures is *Normalized Probabilistic Rand* (NPR) index which was introduced in [UPH05] and evaluated in [CSI+09] on dermoscopy images. Other non-semantic segmentation measures were introduced in [MFTM01], but the reason for creating them seems to be to deal with the under-defined task description of non-semantic segmentation. These accuracy measures try to deal with different levels of coarsity of the segmentation. This is much less of a problem in semantic segmentation and thus those measures are not explained here.

2) Speed: A maximum upper bound on the execution time for the inference on a single image is a hard requirement for some applications. For example, in the case of autonomous cars an algorithm which classifies pixel as street or no-street and thus makes a semantic segmentation, every image needs to be processed within 20 ms [BKTT15]. This time is called **latency**.

Most papers do not give exact values for the time their application needs. One reason might be that this is very hardware, implementation and in some cases even data specific. For example, [HJBJ+96] notes that their algorithm needs 10 s on a Sun SparcStation 20. The fastest CPU ever produced for this system had 200 MHz. Comparing this directly with results which were obtained using an Intel i7-4820K with 3.9 GHz would not be meaningful.

However, it does still make sense to mention the execution time as well as the hardware in individual papers. This gives the interested reader the possibility to estimate how difficult it might be to adjust the algorithm to work in the required time-constraints.

Besides the latency, the **throughput** is another relevant characteristic of algorithms and implementations for semantic segmentation. For example, for the automatic description of images in order to enable text search the throughput is of much higher importance than latency.

3) Stability: A reasonable requirement on semantic segmentation algorithms is the stability of a segmentation over slight changes in the input image. When the image data is sightly blurred by smoke such as in Figure 4(c), the segmentation should not change. Also, two images which show a slight change in perspective should also only result in slight changes in the segmentation [PH05].

4) Memory usage: Peak memory usage matters when segmentation algorithms are used in devices like smartphones or cameras, or when the algorithms have to finish in a given time frame, run on the graphics processing unit (GPU) and consume so much memory for single image segmentation that only the latest graphic cards can be used. However, no publication were available mentioning the peak memory usage.

B. Datasets

The computer vision community produced a couple of different datasets which are publicly available. In the following, only the most widely used ones as well as three medical databases are described. An overview over the quantity and the kind of data is given by Table I.

1) PASCAL VOC: The PASCAL[1] VOC[2] challenge was organized eight times with different datasets: Once every year from 2005 to 2012 [EVGW+b].

[1] **p**attern **a**nalysis, **s**tatistical modelling and **c**omput**a**tional **l**earning, an EU network of excellence

[2] **V**isual **O**bject **C**lasses

Beginning with 2007, a segmentation challenge was added [EVGW+a].

The dataset consists of annotated photographs from www.flicker.com, a photo sharing website. There are multiple challenges for PASCAL VOC. The 2012 competition had five challenges of which one is a segmentation challenge where a single class label was given for each pixel. The classes are: aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, sofa, train, tv/monitor.

Although no new competitions will be held, new algorithms can be evaluated on the 2010, 2011 and 2012 data via http://host.robots.ox.ac.uk:8080/

The PASCAL VOC segmentation challenges use the *segmentation over union* criterion (see Section III-A).

2) MSRCv2: Microsoft Research has published a database of 591 photographs with pixel-level annotation of 21 classes: aeroplane, bike, bird, boat, body, book, building, car, cat, chair, cow, dog, face, flower, grass, road, sheep, sign, sky, tree, water. Additionally, there is a `void` label for pixels which do not belong to any of the 21 classes or which are close to the segmentation boundary. This allows a "rough and quick hand-segmentation which does not align exactly with the object boundaries" [SWRC06].

3) Medical Databases: The Warwick-QU Dataset consists of 165 images with pixel-level annotation of 5 classes: "healthy, adenomatous, moderately differentiated, moderately-to-poorly differentiated, and poorly differentiated" [CSM09]. This dataset is part of the Gland Segmentation (GlaS) challenge.

The DIARETDB1 [KKV+14] is a dataset of 89 images fundus images. Those images show the interior surface of the eye. Fundus images can be used to detect diabetic retinopathy. The images have four classes of coarse annotations: hard and soft exudates, hemorrhages and red small dots.

20 test and additionally 20 training retinal fundus images are available through the DRIVE data set [SAN+04]. The vessels were annotated. Additionally, [AP11] added vascular features.

The Open-CAS Endoscopic Datasets [MHMK+14] are 60 images taken from laparoscopic adrenalectomies and 60 images taken from laparoscopic pancreatic resections. Those are from 3 surgical procedures each. Half of the data was annotated by a medical expert for "medial instrument" and "no medical instrument". All images were labeled by anonymous untrained workers to which they refer to as *knowledge workers* (KWs). One crowd annotation was obtained for each image by a majority vote on a pixel basis of 10 segmentations given by 10 different KWs.

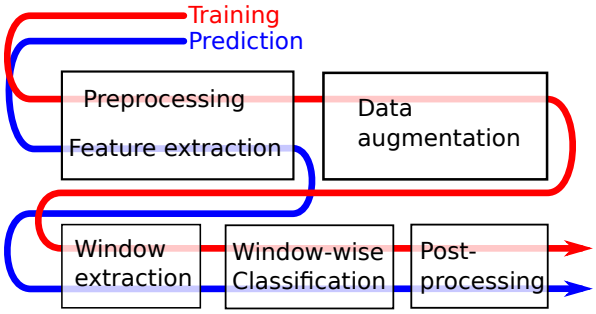


Figure 2: A typical segmentation pipeline gets raw pixel data, applies preprocessing techniques like scaling and feature extraction like HOG features. For training, data augmentation techniques such as image rotation can be applied. For every single image, patches of the image called *windows* are extracted and those windows are classified. The resulting semantic segmentation can be refined by simple morphologic operations or by more complex approaches such as Markov Random Fields (MRFs).

IV. Segmentation Pipeline

Typically, semantic segmentation is done with a classifier which operates on fixed-size feature inputs and a *sliding-window* approach [DT05], [YBCK10], [SCZ08]. This means a classifier is trained on images of a fixed size. The trained classifier is then fed with rectangular regions of the image which are called *windows*. Although the classifier gets an image patch of e.g. $51\,\text{px} \times 51\,\text{px}$ of the environment, it might only classify the center pixel or a subset of the complete window. This segmentation pipeline is visualized in Figure 2.

This approach was taken by [BKTT15] and a majority of the VOC2007 participants [EVGW+a]. As this approach has to apply the patch classifier $512 \cdot 512 = 262\,144$ times for images of size $512\,\text{px} \times 512\,\text{px}$, there are techniques for speeding it up such as applying a stride and interpolating the results.

Neural networks are able to apply the sliding window approach in a very efficient way by handling a trained network as a convolution and applying the convolution on the complete image.

However, there are alternatives. Namely MRFs and Conditional Random Fields (CRFs) which take the information of the complete image and segment it in an holistic approach.

V. Traditional Approaches

Image segmentation algorithms which use traditional approaches, hence don't apply neural networks and make heavy use of domain knowledge, are wide-spread in the computer vision community. Features which can be used for segmentation are described in Section V-A, a very brief overview of unsupervised, non-semantic segmentation is given in Section V-B, Random Decision Forests are described in Section V-C, Markov Random Fields in Section V-E and Support Vector Machines (SVMs) in Section V-D. Postprocessing is covered in Section V-G.

It should be noted that algorithms can use combination of methods. For example, [TNL14] makes use of a combination of a SVM and a MRF. Also, auto-encoders can be used to learn features which in turn can be used by any classifier.

A. Features and Preprocessing methods

The choice of features is very important in traditional approaches. The most commonly used local and global features are explained in the following as well as feature dimensionality reduction algorithms.

1) Pixel Color: Pixel color in different image spaces (e.g. 3 features for RGB, 3 features for HSV, 1 feature for the gray-value) are the most widely used features. A typical image is in the RGB color space, but depending on the classifier and the problem another color space might result in better segmentations. RGB, YcBcr, HSL, Lab and YIQ are some examples used by [CRSS14]. No single color space has been proven to be superior to all others in all contexts [CJSW01]. However, the most common choices seem to be RGB and HSI. Reasons for choosing RGB is simplicity and the support by programming languages, whereas the choice of the HSI color space might make it simpler for the classifier to become invariant to illumination. One reason for choosing CIE-L*a*b* color space is that it approximates human perception of brightness [KP92]. It follows that choosing the L*a*b color space helps algorithms to detect structures which are seen by humans. Another way of improving the structure within an image is histogram equalization, which can be applied to improve contrast [PAA$^+$87], [RM07].

2) Histogram of oriented Gradients: Histogram of oriented gradients (HOG) features interpret the image as a discrete function $I : \mathbb{N}^2 \to \{\, 0, \ldots, 255 \,\}$ which maps the position (x, y) to a color. For each pixel, there are two gradients: The partial derivative of x and y. Now the original image is transformed to two feature maps of equal size which represents the gradient. These feature maps are splitted into patches and a histogram of the directions is calculated for each patch. HOG features were proposed in [DT05] and are used in [BMBM10], [FGMR10] for segmentation tasks.

3) SIFT: Scale-invariant feature transform (SIFT) feature descriptors describe keypoints in an image. The image patch of the size 16×16 around the keypoint is taken. This patch is divided in 16 distinct parts of the size 4×4. For each of those parts a histogram of 8 orientations is calculated similar as for HOG features. This results in a 128-dimensional feature vector for each keypoint.

It should be emphasized that SIFT is a global feature for a complete image.

SIFT is described in detail in [Low04] and are used in [PTN09].

4) BOV: Bag-of-visual-words (BOV), also called *bag of keypoints*, is based on vector quantization. Similar to HOG features, BOV features are histograms which count the number of occurrences of certain patterns within a patch of the image. BOV are described in [CDF$^+$04] and used in combination with SIFT feature descriptors in [CP08].

5) Poselets: Poselets rely on manually added extra keypoints such as "right shoulder", "left shoulder", "right knee" and "left knee". They were originally used for human pose estimation. Finding those extra keypoints is easily possible for well-known image classes like humans. However, it is difficult for classes like airplanes, ships, organs or cells where the human annotators do not know the keypoints. Additionally, the keypoints have to be chosen for every single class. There are strategies to deal with those problems like viewpoint-dependent keypoints. Poselets were used in [BMBM10] to detect people and in [BBMM11] for general object detection of the PASCAL VOC dataset.

6) Textons: A *texton* is the minimal building block of vision. The computer vision literature does not give a strict definition for textons, but edge detectors could be one example. One might argue that deep learning techniques with Convolution Neuronal Networks (CNNs) learn textons in the first filters.

An excellent explanation of textons can be found in [ZGWX05].

7) Dimensionality Reduction: High-resolution images have a lot of pixels. Having one or more feature per pixel results in well over a million features. This makes training difficult while the higher resolution might not contain much more information. A simple approach to deal with this is downsampling the high-resolution image to a low-resolution variant. Another way of doing dimensionality reduction is principal component analysis (PCA), which is applied by [COWR11]. The idea behind PCA is to find a hyperplane on which all

feature vectors can be projected with a minimal loss of information. A detailed description of PCA is given by [Smi02].

One problem of PCA is the fact that it does not distinguish different classes. This means it can happen that a perfectly linearly separable set of feature vectors becomes not separable at all after applying PCA.

There are many other techniques for dimensionality reduction. An overview and a comparison over some of them is given by [vdMPvdH09].

B. Unsupervised Segmentation

Unsupervised segmentation algorithms can be used in supervised segmentation as another source of information or to refine a segmentation. While unsupervised segmentation algorithms can never be semantic, they are well-studied and deserve at least a very brief overview.

Semantic segmentation algorithms store information about the classes they were trained to segment while non-semantic segmentation algorithms try to detect consistent regions or region boundaries.

1) Clustering Algorithms: Clustering algorithms can directly be applied on the pixels, when one gives a feature vector per pixel. Two clustering algorithms are k-means and the mean-shift algorithm.

The k-means algorithm is a general-purpose clustering algorithm which requires the number of clusters to be given beforehand. Initially, it places the k centroids randomly in the feature space. Then it assigns each data point to the nearest centroid, moves the centroid to the center of the cluster and continues the process until a stopping criterion is reached. A faster variant is described in [Har75].

k-means was applied by [CLP98] for medical image segmentation.

Another clustering algorithm is the mean-shift algorithm which was introduced by [CM02] for segmentation tasks. The algorithm finds the cluster centers by initializing centroids at random seed points and iteratively shifting them to the mean coordinate within a certain range. Instead of taking a hard range constraint, the mean can also be calculated by using any kernel. This effectively applies a weight to the coordinates of the points. The mean shift algorithm finds cluster centers at positions with a highest local density of points.

2) Graph Based Image Segmentation: Graph-based image segmentation algorithms typically interpret pixels as vertices and an edge weight is a measure of dissimilarity such as the difference in color [FH04], [Fel]. There are several different candidates for edges. The 4-neighborhood (north, east, south west) or an 8-neighborhood (north, north-east, east, south-east, south, south-west, west, north-west) are plausible choices. One way to cut the edges is by building a minimum spanning tree and removing edges above a threshold. This threshold can either be constant, adapted to the graph or adjusted by the user. After the edge-cutting step, the connected components are the segments.

A graph-based method which ranked 2$^{\text{nd}}$ in the Pascal VOC 2010 challenge [EVGW^{+}10] is described in [CS10]. The system makes heavy use of the multi-cue contour detector globalPb [MAFM08] and needs about 10 GB of main memory [CS11].

3) Random Walks: Random walks belong to the graph-based image segmentation algorithms. Random walk image segmentation usually works as follows: Seed points are placed on the image for the different objects in the image. From every single pixel, the probability to reach the different seed points by a random walk is calculated. This is done by taking image gradients as described in Section V-A for HOG features. The class of the pixel is the class of which a seed point will be reached with highest probability. At first, this is an interactive segmentation method, but it can be extended to be non-interactive by using another segmentation methods output as seed points.

4) Active Contour Models: Active contour models (ACMs) are algorithms which segment images roughly along edges, but also try to find a border which is smooth. This is done by defining a so called *energy function* which will be minimized. They were initially described in [KWT88]. ACMs can be used to segment an image or to refine segmentation as it was done in [AM98] for brain MR images.

5) Watershed Segmentation: The watershed algorithm takes a grayscale image and interprets it as a height map. Low values are catchment basins and the higher values between two neighboring catchment basins is the watershed. The catchment basins should contain what the developer wants to capture. This implies that those areas must be dark on grayscale images. The algorithm starts to fill the basins from the lowest point. When two basins are connected, a watershed is found. The algorithm stops when the highest point is reached.

A detailed description of the watershed segmentation algorithm is given in [RM00].

The watershed segmentation was used in [JLD03] to segment white blood cells. As the authors describe, the segmentation by watershed transform has two flaws: Over-segmentation due to local minima and thick watersheds due to plateaus.

C. Random Decision Forests

Random Decision Forests were first proposed in [Ho95]. This type of classifier applies techniques called *ensemble learning*, where multiple classifiers are trained and a combination of their hypotheses is used. One ensemble learning technique is the *random subspaces* method where each classifier is trained on a random subspace of the feature space. Another ensemble learning technique is *bagging*, which is training the trees on random subsets of the training set. In the case of Random Decision Forests, the classifiers are decision trees. A decision tree is a tree where each inner node uses one or more features to decide in which branch to descend. Each leaf is a class.

One strength of Random Decision Forests compared to many other classifiers like SVMs and neural networks is that the scale of measure of the features (nominal, ordinal, interval, ratio) can be arbitrary. Another advantage of Random Decision Forests compared to SVMs, for example, is the speed of training and classification.

Decision trees were extensively studied in the past 20 years and a multitude of training algorithms have been proposed (e.g. ID3 in [Qui86], C4.5 in [Qui93]). Possible training hyperparameters are the measure to evaluate the "goodness of split" [Min89], the number of decision trees being used, and if the depth of the trees is restricted. Typically in the context of classification, decision trees are trained by adding new nodes until each leaf contains only nodes of a single class or until it is not possible to split further. This is called a *stopping criterion*.

There are two typical training modes: *Central axis projection* and *perceptron training*. In training, for each node a hyperplane is searched which is optimal according to an error function.

Random Decision Forests with texton features (see Section V-A6) are applied in [SJC08] for segmentation. In the [MSC] dataset, they report a per-pixel accuracy rate of $66.9\,\%$ for their best system. This system requires $415\,\text{ms}$ for the segmentation of $320\,\text{px} \times 213\,\text{px}$ images on a single $2.7\,\text{GHz}$ core. On the Pascal VOC 2007 dataset, they report an average per-pixel accuracy for their best segmentation system of $42\,\%$.

An excellent introduction to Random Decision Forests for semantic segmentation is given by [SCZ08].

D. SVMs

SVMs are well-studied binary classifiers which can be described by five central ideas. For those ideas, the training data is represented as $(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i$ is the feature vector and $y_i \in \{ -1, 1 \}$ the binary label for training example $i \in \{ 1, \dots, m \}$.

1) If data is linearly separable, it can be separated by a hyperplane. There is one hyperplane which maximizes the distance to the next datapoints (*support vectors*). This hyperplane should be taken:

$$\begin{aligned} &\underset{\mathbf{w},b}{\text{minimize}}\, \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{s.t. } \forall_{i=1}^{m} y_i \cdot \underbrace{(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)}_{\text{sgn applied to this gives the classification}} \geq 1 \end{aligned}$$

2) Even if the underlying process which generates the features for the two classes is linearly separable, noise can make the data not separable. The introduction of *slack variables* to relax the requirement of linear separability solves this problem. The trade-off between accepting some errors and a more complex model is weighted by a parameter $C \in \mathbb{R}_0^+$. The bigger C, the more errors are accepted. The new optimization problem is:

$$\begin{aligned} &\underset{\mathbf{w}}{\text{minimize}}\, \frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \sum_{i=1}^{m} \xi_i \\ &\text{s.t. } \forall_{i=1}^{m} y_i \cdot (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \end{aligned}$$

Note that $0 \leq \xi_i \leq 1$ means that the data point is within the margin, whereas $\xi_i \geq 1$ means it is misclassified. An SVM with $C > 0$ is also called a *soft-margin SVM*.

3) The primal problem is to find the normal vector $\mathbf{w}$ and the bias b. The dual problem is to express $\mathbf{w}$ as a linear combination of the training data $\mathbf{x}_i$:

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i$$

where $y_i \in \{ -1, 1 \}$ represents the class of the training example and α_i are Lagrange multipliers. The usage of Lagrange multipliers is explained with some examples in [Smi04]. The usage of the Lagrange multipliers α_i changes the optimization problem depend on the α_i which are weights for the feature vectors. It turns out that most α_i will be zero. The non-zero weighted vectors are called *support vectors*.
The optimization problem is now, according to [Bur98]:

$$\begin{aligned} &\underset{\alpha_i}{\text{maximize}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ &\text{s.t. } \forall_{i=1}^{m} 0 \leq \alpha_i \leq C \\ &\text{s.t. } \sum_{i=1}^{m} \alpha_i y_i = 0 \end{aligned}$$

4) Not every dataset is linearly separable. This problem is approached by transforming the feature vectors $\mathbf{x}$ with a non-linear mapping Φ into a higher dimensional (probably ∞-dimensional) space. As the feature vectors $\mathbf{x}$ are only used within scalar product $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$, it is not necessary to do the transformation. It is enough to do the calculation

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

This function K is called a *kernel*. The idea of never explicitly transforming the vectors $\mathbf{x}_i$ to the higher dimensional space is called the *kernel trick*. Common kernels include the polynomial kernel

$$K_P(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_i, \mathbf{x}_j \rangle + r)^p$$

of degree p and coefficient r, the Gaussian radial basis function (RBF) kernel

$$K_{\text{Gauss}}(\mathbf{x}_i, \mathbf{x}_j) = e^{\frac{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$$

and the sigmoid kernel

$$K_{\tanh}(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \langle \mathbf{x}_i, \mathbf{x}_j \rangle - r)$$

where the parameter γ determines how much influence single training examples have.

5) The described SVMs can only distinguish between two classes. Common strategies to expand those binary classifiers to multi-class classification is the *one-vs-all* and the *one-vs-one* strategy. In the one-vs-all strategy n classifiers have to be trained which can distinguish one of the n classes against all other classes. In the one-vs-one strategy $\frac{n^2-n}{2}$ classifiers are trained; one classifier for each pair of classes.

A detailed description of SVMs can be found in [Bur98].

SVMs are used by [YHRF12] on the 2009 and 2010 PASCAL segmentation challenge [EVGW+10]. They did not hand their classifier in to the challenge itself, but calculated an average rank of 7 among the different categories.

[FGMR10] also used an SVM based method with HOG features and achieved the 7th rank in the 2010 PASCAL segmentation challenge by mean accuracy. It needs about 2 s on a 2.8 GHz 8-core Intel processor.

E. Markov Random Fields

MRFs are undirected probabilistic graphical models which are wide-spread model in computer vision. The overall idea of MRFs is to assign a random variable for each feature and a random variable for each pixel which gets labeled as shown in Figure 3. For example, a MRF which is trained on images of the size $224\,\text{px} \times 224\,\text{pixel}$ and gets the raw RGB values as features has

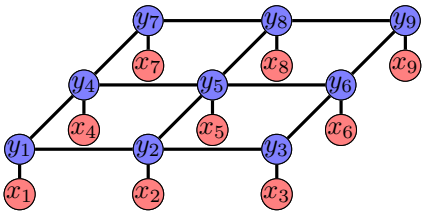


Figure 3: CRF with 4-neighborhood. Each node x_i represents a pixel and each node y_i represents a label.

$$\underbrace{224 \cdot 224 \cdot 3}_{\text{input}} + \underbrace{224 \cdot 224}_{\text{output}} = 200\,704$$

random variables. Those random variables are conditionally independent, given their local neighborhood. These (in)dependencies can be expressed with a graph.

Let $G = (\mathcal{V}, \mathcal{E})$ be the associated undirected graph of an MRF and $\mathcal{C}$ be the set of all maximal cliques in that graph. Nodes represent random variables $\mathbf{x}, \mathbf{y}$ and edges represent conditional dependencies. Just like in he 4-neighborhood [SWRC06] and the 8-neighborhood are reasonable choices for constructing the graph.

Typically, random variables $\mathbf{y}$ represent the class of a single pixel, random variables $\mathbf{x}$ represent a pixel values and edges represent pixel neighborhood in computer vision problems segmentation problems where MRFs are used. Accordingly, the random variables $\mathbf{y}$ live on $1, \ldots,$ nr of classes and the random variables $\mathbf{x}$ typically live on $0, \ldots, 255$ or $[0, 1]$.

The probability of $\mathbf{x}, \mathbf{y}$ can be expressed as

$$P(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} e^{-E(\mathbf{x}, \mathbf{y})}$$

where $Z = \sum_{\mathbf{x}, \mathbf{y}} e^{-E(\mathbf{x}, \mathbf{y})}$ is a normalization term called the *partition function* and E is called the *energy function*. A common choice for the energy function is

$$E(\mathbf{x}, \mathbf{y}) = \sum_{c \in \mathcal{C}} \psi_c(\mathbf{x}, \mathbf{y})$$

where ψ is called a *clique potential*. One choice for cliques of size two $\mathbf{x}, \mathbf{y} = (x_1, x_2)$ is [KP06]

$$\psi_c(x_1, x_2) = w\delta(x_1, x_2) = \begin{cases} +w & \text{if } x_1 \neq x_2 \\ -w & \text{if } x_1 = x_2 \end{cases}$$

According to [Mur12], the most common way of inference over the posterior MRF in computer vision problems is Maximum A Posteriori (MAP) estimation.

Detailed introductions to MRFs are given by [BKR11], [Mur12]. MRFs are used by [ZBS01] and [MSB12] for image segmentation.

F. Conditional Random Fields

CRFs are MRFs where all clique potentials are conditioned on input features [Mur12]. This means, instead of learning the distribution $P(\mathbf{y}, \mathbf{x})$, the task is reformulated to learn the distribution $P(\mathbf{y}|\mathbf{x})$. One consequence of this reformulation is that CRFs need much less parameters as the distribution of $\mathbf{x}$ does not have to be estimated. Another advantage of CRFs compared to MRFs is that no distribution assumption about $\mathbf{x}$ has to be made.

A CRF has the partition function Z:

$$Z(\mathbf{x}) = \sum_{\mathbf{y}} P(\mathbf{x}, \mathbf{y})$$

and joint probability distribution

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{y}_c|\mathbf{x})$$

The simplest way to define the clique potentials ψ is the count of the class $\mathbf{y}_c$ given $\mathbf{x}$ added with a positive smoothing constant to prevent the complete term from getting zero.

CRFs as described in [LRKT09] have reached top performance in PASCAL VOC 2010 [VOC10] and are also used in [HZCP04], [SWRC06] for semantic segmentation.

A method similar to CRFs was proposed in [GBVdW$^+$10]. The system of Gonfaus et.al. ranked 1$^{\text{st}}$ by mean accuracy in the segmentation task of the PASCAL VOC 2010 challenge [EVGW$^+$10].

An introduction to CRFs is given by [SM11].

G. Post-processing methods

Post-processing refine a found segmentation and remove obvious errors. For example, the morphological operations *opening* and *closing* can remove noise. The opening operation is a dilation followed by a erosion. This removes tiny segments. The closing operation is a erosion followed by a dilation. This removes tiny gaps in otherwise filled regions. They were used in [CLP98] for biomedical image segmentation.

Another way of refinement of the found segmentation is by adjusting the segmentation to match close edges. This was used in [BBMM11] with an ultra-metric contour map [AMFM09].

Active contour models are another example of a post-processing method [KWT88].

VI. Neural Networks for Semantic Segmentation

Artificial neural networks are classifiers which are inspired by biologic neurons. Every single artificial neuron has some inputs which are weighted and sumed up. Then, the neuron applies a so called *activation function* to the weighted sum and gives an output. Those neurons can take either a feature vector as input or the output of other neurons. In this way, they build up feature hierarchies.

The parameters they learn are the *weights* $w \in \mathbb{R}$. They are learned by gradient descent. To do so, an error function — usually cross-entropy or mean squared error — is necessary. For the gradient descent algorithm, one sees the labeled training data as given, the weights as variables and the error function as a surface in this weight-space. Minimizing the error function in the weight space adapts the neural network to the problem.

There are lots of ideas around neural networks like regularization, better optimization algorithms, automatically building up architectures, design choices for activation functions. This is not explained in detail here, but some of the mayor breakthroughs are outlined.

CNNs are neural networks which learn image filters. They drastically reduce the number of parameters which have to be learned while being still general enough for the problem domain of images. This was shown by Alex Krizhevsky et al. in [KSH12]. One major idea was a clever regularization called *dropout training*, which set the output of neurons while training randomly to zero. Another contribution was the usage of an activation function called *rectified linear unit*:

$$\varphi_{\text{ReLU}}(x) = \max(0, x)$$

Those are much faster to train than the commonly used sigmoid activation functions

$$\varphi_{\text{Sigmoid}}(x) = \frac{1}{e^{-x} + 1}$$

Krizhevsky et al. implemented those ideas and participated in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). The best other system, which used SIFT features and Fisher Vectors, had a performance of about $25.7\,\%$ while the network by Alex Krizhevsky et al. got $17.0\,\%$ error rate on the ILSVRC-2010 dataset. As a preprocessing step, they downsampled all images to a fixed size of $256\,\text{px} \times 256\,\text{px}$ before they fed the features into their network. This network is commonly known as *AlexNet*.

Since AlexNet was developed, a lot of different neural networks have been proposed. One interesting example is [PC13], where a recurrent CNN for semantic segmentation is presented.

Another notable paper is [LSD14]. The algorithm presented there makes use of a classifying network such as AlexNet, but applies the complete network as an image filter. This way, each pixel gets a probability distribution for each of the trained classes. By taking the most likely class, a semantic segmentation can be done with arbitrary image sizes.

A very recent publication by Dai et al. [DHS15] showed that segmentation with much deeper networks is possible and achieves better results.

More detailed explanations to neural networks for visual recognition is given by [LKJ15].

VII. Possible Problems in the Data for Segmentation Algorithms

Different segmentation workflows have different problems. However, there are a couple of special cases which should be tested. Those cases might not occur often in the training data, but it could still happen in the productive system.

I am not aware of any systematic work which examined the influence of problems such as the following.

A. Lens Flare

Lens flare is the effect of light getting scattered in the lens system of the camera. The testing data set of the KITTI road evaluation benchmark [FKG13] has a couple of photos with this problem. Figure 4(a) shows an extreme example of lens flare.

B. Vignetting

Vignetting is the effect of a photograph getting darker in the corners. This can have many reasons, for example filters on the camera blocking light at the corners.

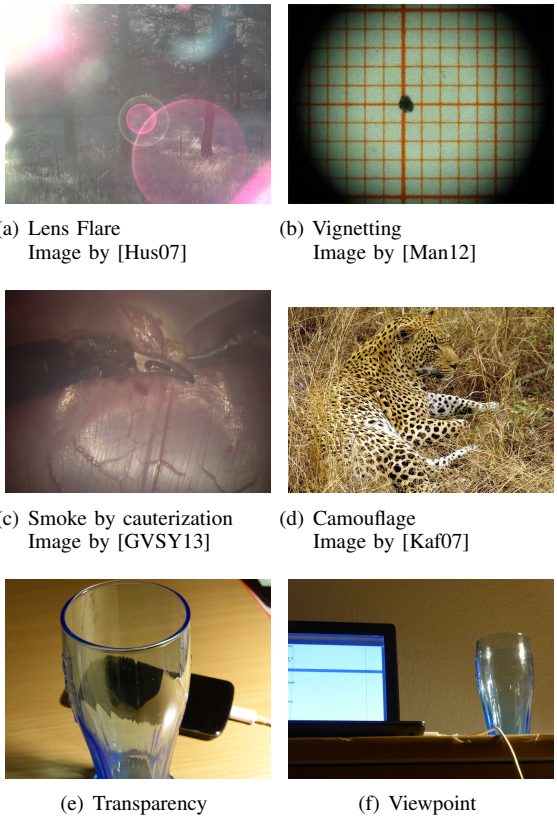
C. Blurred images

Images can be blurred for a couple of reasons. A problem with the lenses mechanics, focusing on the wrong point, too quick movement, smoke or foam. One example of a blurred image is Figure 4(c), which was taken during an in vivo porcine procedure of diaphragm dissection. The smoke was caused by cauterization.

D. Other Problems

If the following effects can occur at all and if they are problems depends heavily on the problem domain and the used model.

1) Partial Occlusions: Segmentation systems which employ a model of the objects which should be segmented might suffer from partial occlusions.



(a) Lens Flare
Image by [Hus07]

(b) Vignetting
Image by [Man12]

(c) Smoke by cauterization
Image by [GVSY13]

(d) Camouflage
Image by [Kaf07]

(e) Transparency

(f) Viewpoint

Figure 4: Examples of images which might cause semantic segmentation systems to fail.

2) Camouflage: Some objects, like animals in the wild, actively try to hide (see Figure 4(d) as an example). In other cases it might just be bad luck that objects are hard for humans to detect. This problem has two interesting aspects: On the one hand, the segmenting system might suffer from the same problems as humans do. On the other hand, the segmenting system might be better than humans are, but it is forced to learn from images labeled by humans. If the labels are wrong, the system is forced to learn something wrong.

3) Semi-transparent Occlusion: Some objects like drinking glasses can be visible and still leave the object behind them visible as shown in Figure 4(e). This is mainly a definition problem: Is the seen pixel the glass label or the smartphone label?

4) Viewpoints: Changes in viewpoints can be a problem, if they don't occur in the training data. For example, an image captioning system which was trained on photographs of professional photographers might not have photos from the point of view of a child. This is visualized in Figure 4(f).

VIII. Discussion

Ohta et al. wrote [OKS78] 38 years ago. It is one of the first papers mentioning semantic segmentation. In this time, a lot of work was done and many different directions have been explored. Different kinds of semantic segmentation have emerged.

This paper presents a taxonomy of those kinds of semantic segmentation and a brief overview of completely automatic, passive, semantic segmentation algorithms.

Future work includes a comparative study of those algorithms on publicly available dataset such as the ones presented in Table I. Another open question is the influence of the problems described in Section VII. This could be done using a subset of the thousands of images of Wikipedia Commons, such as https://commons.wikimedia.org/wiki/Category:Blurring for blurred images.

A combination of different classifiers in an ensemble would be an interesting option to explore in order to improve accuracy. Another direction which is currently studied is combining classifiers such as neural networks with CRFs [ZJRP$^+$15].

References

[AM98] M. S. Atkins and B. T. Mackiewich, "Fully automatic segmentation of the brain in mri," *Medical Imaging, IEEE Transactions on*, vol. 17, no. 1, pp. 98–107, Feb. 1998. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=668699

[AMFM09] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "From contours to regions: An empirical evaluation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, Jun. 2009, pp. 2294–2301. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5206707

[AP11] G. Azzopardi and N. Petkov, "Detection of retinal vascular bifurcations by trainable v4-like filters," in *Computer Analysis of Images and Patterns*. Springer, 2011, pp. 451–459. [Online]. Available: http://www.cs.rug.nl/~imaging/databases/retina_database/retinalfeatures_database.html

[BBMM11] T. Brox, L. Bourdev, S. Maji, and J. Malik, "Object segmentation by alignment of poselet activations to image contours," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, Jun. 2011, pp. 2225–2232. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5995659

[BJ00] Y. Boykov and M.-P. Jolly, "Interactive organ segmentation using graph cuts," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2000*. Springer, 2000, pp. 276–286. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-540-40899-4_28

[BKR11] A. Blake, P. Kohli, and C. Rother, *Markov random fields for vision and image processing*. Mit Press, 2011.

[BKTT15] S. Bittel, V. Kaiser, M. Teichmann, and M. Thoma, "Pixel-wise segmentation of street with neural networks," *arXiv preprint arXiv:1511.00513*, 2015. [Online]. Available: http://arxiv.org/abs/1511.00513

[BMBM10] L. Bourdev, S. Maji, T. Brox, and J. Malik, "Detecting people using mutually consistent poselet activations," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 168–181. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-15567-3_13#page-1

[Bur98] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[BVZ01] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 11, pp. 1222–1239, 2001. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=969114

[CDF$^+$04] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1-22. Prague, 2004, pp. 1–2.

[CJSW01] H.-D. Cheng, X. Jiang, Y. Sun, and J. Wang, "Color image segmentation: advances and prospects," *Pattern recognition*, vol. 34, no. 12, pp. 2259–2281, 2001.

[CLP98] C. W. Chen, J. Luo, and K. J. Parker, "Image segmentation via adaptive k-mean clustering and knowledge-based morphological operations with biomedical applications," *Image Processing, IEEE Transactions on*, vol. 7, no. 12, pp. 1673–1683, Dec.

1998. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=730379

[CM02] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603–619, 2002. [Online]. Available: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1000236

[COWR11] C. Chen, J. Ozolek, W. Wang, and G. K. Rohde, "A pixel classification system for segmenting biomedical images using intensity neighborhoods and dimension reduction," in *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*. IEEE, 2011, pp. 1649–1652. [Online]. Available: https://www.andrew.cmu.edu/user/gustavor/chen_isbi_11.pdf

[CP08] G. Csurka and F. Perronnin, "A simple high performance approach to semantic segmentation." in *BMVC*, 2008, pp. 1–10. [Online]. Available: http://www.xrce.xerox.com/layout/set/print/content/download/16654/118653/file/2008-023.pdf

[CRSS] A. Cohen, E. Rivlin, I. Shimshoni, and E. Sabo, "Colon crypt segmentation website." [Online]. Available: http://mis.haifa.ac.il/~ishimshoni/SegmentCrypt/Download.htm

[CRSS14] ——, "Memory based active contour algorithm using pixel-level classified images for colon crypt segmentation," *Computerized Medical Imaging and Graphics*, Nov. 2014. [Online]. Available: http://mis.haifa.ac.il/~ishimshoni/SegmentCrypt/Active%20contour%20based%20on%20pixel-level%20classified%20image%20for%20colon%20crypts%20segmentation.pdf

[CS10] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3241–3248.

[CS11] ——, "Cpmc: Constrained parametric min-cuts for automatic object segmentation," Feb. 2011. [Online]. Available: http://www.maths.lth.se/matematiklth/personal/sminchis/code/cpmc/

[CSI+09] M. E. Celebi, G. Schaefer, H. Iyatomi, W. V. Stoecker, J. M. Malters, and J. M. Grichnik, "An improved objective evaluation measure for border detection in dermoscopy images," *Skin Research and Technology*, vol. 15, no. 4, pp. 444–450, 2009. [Online]. Available: http://arxiv.org/abs/1009.1020

[CSM09] L. P. Coelho, A. Shariff, and R. F. Murphy, "Nuclear segmentation in microscope cell images: a hand-segmented dataset and comparison of algorithms," in *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*. IEEE, 2009, pp. 518–521. [Online]. Available: http://murphylab.web.cmu.edu/data

[CXGS12] M. D. Collins, J. Xu, L. Grady, and V. Singh, "Random walks based multi-image segmentation: Quasiconvexity results and gpu-based solutions," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1656–1663. [Online]. Available: http://pages.cs.wisc.edu/~jiaxu/pub/rwcoseg.pdf

[DHS15] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," *arXiv preprint arXiv:1512.04412*, 2015.

[DT05] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, June 2005, pp. 886–893 vol. 1. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1467360

[EVGW+a] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. [Online]. Available: http://host.robots.ox.ac.uk:8080/pascal/VOC/voc2007/index.html

[EVGW+b] ——, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. [Online]. Available: http://host.robots.ox.ac.uk:8080/pascal/VOC/voc2012/index.html

[EVGW+10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[EVGW+12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "Visual object classes challenge 2012 (voc2012)," 2012. [Online]. Available: http://host.robots.ox.ac.uk:8080/pascal/VOC/voc2012/index.html

[Fel] P. F. Felzenszwalb, "Graph based image segmentation." [Online]. Available: http://cs.brown.edu/~pff/segment/

[FGMR10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[FH04] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004. [Online]. Available: http://link.springer.com/article/10.1023/B:VISI.0000022288.19776.77

[FKG13] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *International Conference on Intelligent Transportation Systems (ITSC)*, 2013. [Online]. Available: http://www.cvlibs.net/datasets/kitti/eval_road.php

[GBVdW+10] J. M. Gonfaus, X. Boix, J. Van de Weijer, A. D. Bagdanov, J. Serrat, and J. Gonzalez, "Harmony potentials for joint classification and segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3280–3287.

[GRC+08] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller, "Multi-class segmentation with relative location prior," *International Journal of Computer Vision*, vol. 80, no. 3, pp. 300–316, Apr. 2008.

[GVSY13] S. Giannarou, M. Visentini-Scarzanella, and G.-Z. Yang, "Probabilistic tracking of affine-invariant anisotropic regions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 130–143, 2013.

[Har75] J. A. Hartigan, *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

[HDT02] C. Huang, L. Davis, and J. Townshend, "An assessment of support vector machines for land cover classification," *International Journal of remote sensing*, vol. 23, no. 4, pp. 725–749, 2002.

[HHR01] S. Hu, E. Hoffman, and J. Reinhardt, "Automatic lung segmentation for accurate quantitation of volumetric x-ray ct images," *Medical Imaging, IEEE*

Transactions on, vol. 20, no. 6, pp. 490–498, Jun. 2001.

[HJBJ+96] A. Hoover, G. Jean-Baptiste, X. Jiang, P. J. Flynn, H. Bunke, D. B. Goldgof, K. Bowyer, D. W. Eggert, A. Fitzgibbon, and R. B. Fisher, "An experimental comparison of range image segmentation algorithms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 7, pp. 673–689, Jul. 1996. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=506791

[Ho95] T. K. Ho, "Random decision forests," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, vol. 1. IEEE, 1995, pp. 278–282. [Online]. Available: http://ect.bell-labs.com/who/tkh/publications/papers/odt.pdf

[Hus07] Hustvedt, "File:cctv lens flare.jpg," Wikipedia Commons, Nov. 2007. [Online]. Available: https://commons.wikimedia.org/wiki/File:CCTV_Lens_flare.jpg

[HZCP04] X. He, R. Zemel, and M. Carreira-Perpindn, "Multiscale conditional random fields for image labeling," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, Jun. 2004, pp. II–695–II–702 Vol.2. [Online]. Available: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1315232

[JLD03] K. Jiang, Q.-M. Liao, and S.-Y. Dai, "A novel white blood cell segmentation scheme using scale-space filtering and watershed clustering," in *Machine Learning and Cybernetics, 2003 International Conference on*, vol. 5, Nov 2003, pp. 2820–2825 Vol.5. [Online]. Available: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1260033

[Kaf07] L. Kaffer, "File:great male leopard in south afrika-jd.jpg," Wikipedia Commons, Jul. 2007. [Online]. Available: https://commons.wikimedia.org/wiki/File:Great_male_Leopard_in_South_Afrika-JD.JPG

[KKV+14] V. Kalesnykiene, J.-k. Kamarainen, R. Voutilainen, J. Pietilä, H. Kälviäinen, and H. Uusitalo, "Diaretdb1 diabetic retinopathy database and evaluation protocol," 2014. [Online]. Available: http://www2.it.lut.fi/project/imageret/diaretdb1/

[KP92] J. M. Kasson and W. Plouffe, "An analysis of selected computer interchange color spaces," *ACM Transactions on Graphics (TOG)*, vol. 11, no. 4, pp. 373–405, 1992.

[KP06] Z. Kato and T.-C. Pong, "A markov random field image segmentation model for color textured images," *Image and Vision Computing*, vol. 24, no. 10, pp. 1103–1114, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885606001223

[KSH12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[KWT88] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, Jan. 1988. [Online]. Available: http://link.springer.com/article/10.1007/BF00133570

[LKJ15] F.-F. Li, A. Karpathy, and J. Johnson, "CS231n: Convolutional neural networks for visual recognition," 2015. [Online]. Available: http://cs231n.stanford.edu/

[Low04] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. [Online]. Available: http://dx.doi.org/10.1023/B%3AVISI.0000029664.99615.94

[LRAL08] A. Levin, A. Rav-Acha, and D. Lischinski, "Spectral matting," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 10, pp. 1699–1712, 2008. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4547428

[LRKT09] L. Ladický, C. Russell, P. Kohli, and P. Torr, "Associative hierarchical crfs for object class image segmentation," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 739–746. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5459248

[LSD14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *arXiv preprint arXiv:1411.4038*, 2014. [Online]. Available: http://arxiv.org/abs/1411.4038

[MAFM08] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik, "Using contours to detect and localize junctions in natural images," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, June 2008, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4587420

[Man12] M. Manske, "File:randabschattung mikroskop kamera 6.jpg," Wikipedia Commons, Dec. 2012. [Online]. Available: https://commons.wikimedia.org/wiki/File:Randabschattung_Mikroskop_Kamera_6.JPG

[MBLAGJ+07] S. Maldonado-Bascon, S. Lafuente-Arroyo, P. Gil-Jimenez, H. Gomez-Moreno, and F. Lopez-Ferreras, "Road-sign detection and recognition based on support vector machines," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 8, no. 2, pp. 264–278, Jun. 2007. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4220659

[MBVLG02] N. Moon, E. Bullitt, K. Van Leemput, and G. Gerig, "Automatic brain and tumor segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2002*. Springer, 2002, pp. 372–379.

[MFTM01] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 416–423. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=937655

[MHMK+14] L. Maier-Hein, S. Mersmann, D. Kondermann, S. Bodenstedt, A. Sanchez, C. Stock, H. G. Kenngott, M. Eisenmann, and S. Speidel, "Can masses of non-experts train highly accurate image classifiers?" in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*. Springer, 2014, pp. 438–445. [Online]. Available: http://opencas.webarchiv.kit.edu/?q=node/26

[Min89] J. Mingers, "An empirical comparison of selection measures for decision-tree induction," *Machine Learning*, vol. 3, no. 4, pp. 319–342, 1989. [Online]. Available: http://dx.doi.org/10.1023/A%3A1022645801436

[MSB12] G. Moser, S. B. Serpico, and J. A. Benediktsson, "Markov random field models for supervised land

cover classification from very high resolution multispectral remote sensing images," in *Advances in Radar and Remote Sensing (TyWRRS), 2012 Tyrrhenian Workshop on*. IEEE, 2012, pp. 235–242. [Online]. Available: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6381135

[MSC] "Object class recognition image database." [Online]. Available: http://research.microsoft.com/vision/cambridge/recognition/

[MSR] "Image understanding - research data," Microsoft Research. [Online]. Available: http://research.microsoft.com/en-us/projects/objectclassrecognition/

[Mur12] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[OKS78] Y.-i. Ohta, T. Kanade, and T. Sakai, "An analysis system for scenes containing objects with substructures," in *Proceedings of the Fourth International Joint Conference on Pattern Recognitions*, 1978, pp. 752–754.

[PAA+87] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0734189X8780186X

[PC13] P. H. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene parsing," *arXiv preprint arXiv:1306.2795*, 2013. [Online]. Available: http://arxiv.org/abs/1306.2795v1

[PH05] C. Pantofaru and M. Hebert, "A comparison of image segmentation algorithms," *Robotics Institute*, p. 336, 2005. [Online]. Available: http://riweb-backend.ri.cmu.edu/pub_files/pub4/pantofaru_caroline_2005_1/pantofaru_caroline_2005_1.pdf

[PS07] A. Protiere and G. Sapiro, "Interactive image segmentation via adaptive weighted distances," *Image Processing, IEEE Transactions on*, vol. 16, no. 4, pp. 1046–1057, 2007. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4130436

[PTN09] N. Plath, M. Toussaint, and S. Nakajima, "Multi-class image segmentation using conditional random fields and global classification," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 817–824.

[PXP00] D. L. Pham, C. Xu, and J. L. Prince, "A survey of current methods in medical image segmentation," *Annual Review of Biomedical Engineering*, vol. 2, no. 1, pp. 315–337, 2000, pMID: 11701515. [Online]. Available: http://dx.doi.org/10.1146/annurev.bioeng.2.1.315

[Qui86] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, Aug. 1986. [Online]. Available: http://dx.doi.org/10.1023/A%3A1022643204877

[Qui93] ——, *C4.5: Programs for Machine Learning*, P. Langley, Ed. Morgan Kaufmann Publishers, Inc., 1993.

[RKB04] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004. [Online]. Available: http://delivery.acm.org/10.1145/1020000/1015720/p309-rother.pdf

[RM00] J. B. Roerdink and A. Meijster, "The watershed transform: Definitions, algorithms and parallelization strategies," *Fundam. Inform.*, vol. 41, no. 1-2, pp. 187–228, 2000.

[RM07] J. Reynolds and K. Murphy, "Figure-ground segmentation using a hierarchical conditional random field," in *Computer and Robot Vision, 2007. CRV '07. Fourth Canadian Conference on*, May 2007, pp. 175–182. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4228537

[RMBK06] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, June 2006, pp. 993–1000. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1640859

[SAN+04] J. Staal, M. D. Abràmoff, M. Niemeijer, M. Viergever, B. Van Ginneken *et al.*, "Ridge-based vessel segmentation in color images of the retina," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 4, pp. 501–509, 2004. [Online]. Available: http://www.isi.uu.nl/Research/Databases/DRIVE/

[SCZ08] F. Schroff, A. Criminisi, and A. Zisserman, "Object class segmentation using random forests." in *BMVC*, 2008, pp. 1–10. [Online]. Available: http://research.microsoft.com/pubs/72423/Criminisi_bmvc2008.pdf

[SJC08] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, Jun. 2008, pp. 1–8. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4587503

[SM11] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Machine Learning*, vol. 4, no. 4, pp. 267–373, 2011. [Online]. Available: http://homepages.inf.ed.ac.uk/csutton/publications/crftutv2.pdf

[Smi02] L. I. Smith, "A tutorial on principal components analysis," *Cornell University, USA*, vol. 51, p. 52, 2002.

[Smi04] B. T. Smith, "Lagrange multipliers tutorial in the context of support vector machines," *Memorial University of Newfoundland St. John's, Newfoundland, Canada*, Jun. 2004.

[SSA12] D. Schiebener, J. Schill, and T. Asfour, "Discovery, segmentation and reactive grasping of unknown objects." in *Humanoids*, 2012, pp. 71–77. [Online]. Available: http://h2t.anthropomatik.kit.edu/pdf/Schiebener2012.pdf

[SUM+11] D. Schiebener, A. Ude, J. Morimotot, T. Asfour, and R. Dillmann, "Segmentation and learning of unknown objects through physical interaction," in *Humanoid Robots (Humanoids), 2011 11th IEEE-RAS International Conference on*. IEEE, 2011, pp. 500–506. [Online]. Available: http://ieeexplore.ieee.org/ielx5/6086637/6100798/06100843.pdf

[SWRC06] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 1–15. [Online]. Available: http://link.springer.com/chapter/10.1007/11744023_1

[TNL14] J. Tighe, M. Niethammer, and S. Lazebnik, "Scene parsing with object instances and occlusion ordering," in *Computer Vision and*

Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014, pp. 3748–3755. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6909874

[UPH05] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "A measure for objective evaluation of image segmentation algorithms," in *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*. IEEE, 2005, pp. 34–34. [Online]. Available: http://repository.cmu.edu/cgi/viewcontent.cgi?article=1365&context=robotics

[vdMPvdH09] L. J. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research*, vol. 10, no. 1-41, pp. 66–71, 2009.

[VOC10] "Voc2010 preliminary results," 2010. [Online]. Available: http://host.robots.ox.ac.uk/pascal/VOC/voc2010/results/index.html

[WAH97] G.-Q. Wei, K. Arbter, and G. Hirzinger, "Automatic tracking of laparoscopic instruments by color coding," in *CVRMed-MRCAS'97*, ser. Lecture Notes in Computer Science, J. Troccaz, E. Grimson, and R. Mösges, Eds. Springer Berlin Heidelberg, 1997, vol. 1205, pp. 357–366. [Online]. Available: http://dx.doi.org/10.1007/BFb0029257

[YBCK10] Z. Yin, R. Bise, M. Chen, and T. Kanade, "Cell segmentation in microscopy imagery using a bag of local bayesian classifiers," in *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*, Apr. 2010, pp. 125–128. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5490399

[YHRF12] Y. Yang, S. Hallman, D. Ramanan, and C. C. Fowlkes, "Layered object models for image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 9, pp. 1731–1743, Sep. 2012. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6042883

[ZBS01] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *Medical Imaging, IEEE Transactions on*, vol. 20, no. 1, pp. 45–57, 2001. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=906424

[ZGWX05] S.-C. Zhu, C.-E. Guo, Y. Wang, and Z. Xu, "What are textons?" *International Journal of Computer Vision*, vol. 62, no. 1-2, pp. 121–143, 2005.

[Zha12] Z. Zhang, "Microsoft kinect sensor and its effect," *MultiMedia, IEEE*, vol. 19, no. 2, pp. 4–10, Feb. 2012.

[ZJRP+15] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537. [Online]. Available: http://www.robots.ox.ac.uk/~szheng/papers/CRFasRNN.pdf

GLOSSARY

ACM active contour model. 6

BOV bag-of-visual-words. 5

CNN Convolution Neuronal Network. 5, 9
CRF Conditional Random Field. 4, 8, 9, 11

GPU graphics processing unit. 3

HOG histogram of oriented gradients. 5, 6, 8

ILSVRC ImageNet Large-Scale Visual Recognition Challenge. 9

MAP Maximum A Posteriori. 8
MR magnetic resonance. 2, 6
MRF Markov Random Field. 4, 8

PCA principal component analysis. 5

RBF radial basis function. 8

SIFT scale-invariant feature transform. 5
SVM Support Vector Machine. 4, 6–8

APPENDIX A
TABLES

Database	Image Resolution (width $\times$ height)	Number of Images	Number of Classes	Channels	Data source
Colon Crypt DB	$(302\,\text{px} - 1116\,\text{px}) \times (349\,\text{px} - 875\,\text{px})$	389	2	3	[CRSS]
DIARETDB1	$1500\,\text{px} \times 1500\,\text{px}$	89	4	3	[KKV+14]
KITTI Road	$(1226\,\text{px} - 1242\,\text{px}) \times (370\,\text{px} - 376\,\text{px})$	289	2	3	[FKG13]
MSRCv1	$(213\,\text{px} - 320\,\text{px}) \times (213\,\text{px} - 320\,\text{px})$	240	9	3	[MSR]
MSRCv2	$(213\,\text{px} - 320\,\text{px}) \times (162\,\text{px} - 320\,\text{px})$	591	23	3	[MSR]
Open-CAS Endoscopic Datasets	$640\,\text{px} \times 480\,\text{px}$	120	2	3	[MHMK+14]
PASCAL VOC 2012	$(142\,\text{px} - 500\,\text{px}) \times (\ 71\,\text{px} - 500\,\text{px})$	2913	20	3	[EVGW+12]
Warwick-QU	$(567\,\text{px} - 775\,\text{px}) \times (430\,\text{px} - 522\,\text{px})$	165	5	3	[CSM09]

Table I: An overview over publicly available image databases with a semantic segmentation ground trouth.