. Literature Review

Convolutional neural networks with many layers have recently been shown to achieve excellent results on many high-level tasks such as image classification, object detection and more recently also semantic segmentation. Particularly for semantic segmentation, a two-stage procedure is often employed. Hereby, convolutional networks are trained to provide good local pixel-wise features for the second step being traditionally a more global graphical model.

Alexander G. Schwing and Raquel Urtasun unifies this two-stage process into a single joint training algorithm. They demonstrate their method on the semantic image segmentation task and show encouraging results on the challenging PASCAL VOC 2012 dataset.

Anastasios Doulamis, Nikolaos Doulamis, Klimis Ntalianis, and Stefanos Kollias proposed an unsupervised video object (VO) segmentation and tracking algorithm based on an adaptable neural-network architecture. The proposed scheme comprises: 1) a VO tracking module and 2) an initial VO estimation module. Object tracking is handled as a classification problem and implemented through an adaptive network classifier, which provides better results compared to conventional motion-based tracking algorithms. Network adaptation is accomplished through an efficient and cost effective weight updating algorithm, providing a minimum degradation of the previous network knowledge and taking into account the current content conditions. A retraining set is constructed and used for this purpose based on initial VO estimation results. Two different scenarios are investigated. The first concerns extraction of human entities in video conferencing applications, while the second exploits depth information to identify generic VOs in stereoscopic video sequences. Human face body detection based on Gaussian distributions is accomplished in the first scenario, while segmentation fusion is obtained using color and depth information in the second scenario. A decision mechanism is also incorporated to detect time instances for weight updating.

Experimental results and comparisons indicate the good performance of the proposed scheme even in sequences with complicated content (object bending, occlusion).

Bharath Hariharan, Pablo Arbel´aez, Ross Girshick, and Jitendra Malik detects all instances of a category in an image and, for each instance, mark the pixels that belong to it. They call this task Simultaneous Detection and Segmentation (SDS). Unlike classical bounding box detection, SDS requires segmentation and not just a box. Unlike classical semantic segmentation, we require individual object instances.

They build on recent work that uses convolutional neural networks to classify category-independent region proposals (R-CNN), introducing a novel architecture tailored for SDS. They then use category-specific, topdown figure-ground predictions to refine our bottom-up proposals. They show a 7 point boost (16% relative) over our baselines on SDS, a 5 point boost (10% relative) over state-of-the-art on semantic segmentation, and state-of-the-art performance in object detection.

Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari presented a generic objectness measure, quantifying how likely it is for an image window to contain an object of any class. We explicitly train it to distinguish objects with a well-defined boundary in space, such as cows and telephones, from amorphous background elements, such as grass and road. The measure combines in a Bayesian framework several image cues measuring characteristics of objects, such as appearing different from their surroundings and having a closed boundary. These include an innovative cue to measure the closed boundary characteristic. In experiments on the challenging PASCAL VOC 07 dataset, we show this new cue to outperform a state-of-the-art saliency measure, and the combined objectness measure to perform better than any cue alone. We also compare to interest point operators, a HOG detector, and three recent works aiming at automatic object segmentation. Finally, they present two applications of objectness. In the first, we sample a small number windows according to their objectness probability and give an algorithm to employ them as location priors for modern class-specific object detectors. As they show experimentally, this greatly reduces the number of windows evaluated by the expensive class-specific model. In the second application, they use objectness as a complementary score in addition to the class-specific model, which leads to fewer false positives. As shown in several recent papers, objectness can act as a valuable focus of attention mechanism in many other applications operating on image windows, including weakly supervised learning of object categories, unsupervised pixelwise segmentation, and object tracking in video. Computing objectness is very efficient and takes only about 4 sec. per image.

Camille Couprie, Clement Farabet, Laurent Najman and Yann LeCun addresses multi-class segmentation of indoor scenes with RGB-D inputs. While this area of research has gained much attention recently, most works still rely on hand-crafted features. In contrast, they apply a multiscale convolutional network to learn features directly from the images and the depth information. They obtain state-of-the-art on the NYU-v2 depth dataset with an accuracy of 64.5%. They illustrate the labeling of indoor scenes in videos sequences that could be processed in real-time using appropriate hardware such as an FPGA.

C.P. Town and D. Sinclair demonstrates an approach to content based image retrieval founded on the semantically meaningful labelling of images by high level visual categories. The image labelling is achieved by means of a set of trained neural network classifiers which map segmented image region descriptors onto semantic class membership terms. It is argued that the semantic terms give a good estimate of the salient features which are important for discrimination in image retrieval. Furthermore, it is shown that the choice of visual categories such as grass or sky which mirror high level human perception allows the implementation of intuitive and versatile query composition interfaces and a variety of image similarity metrics for content based retrieval.

Christian Szegedy, Alexander Toshev and Dumitru Erhan shown outstanding performance on image classification tasks. In this paper we go one step further and address the problem of object detection using DNNs, that is not only classifying but also precisely localizing objects of various classes. They present a simple and yet powerful formulation of object detection as a regression problem to object bounding box masks. We define a multi-scale inference procedure which is able to produce high-resolution object detections at a low cost by a few network applications. State-of-the-art performance of the approach is shown on Pascal VOC.

Cl´ement Farabet, Camille Couprie, Laurent Najman, Yann Lecun propose a method that uses a multiscale convolutional network trained from raw pixels to extract dense feature vectors that encode regions of multiple sizes centered on each pixel. The method alleviates the need for engineered features, and produces a powerful representation that captures texture, shape and contextual information. They report results using multiple post-processing methods to produce the final labeling. Among those, they propose a technique to automatically retrieve, from a pool of segmentation components, an optimal set of components that best explain the scene; these components are arbitrary, e.g. they can be taken from a segmentation tree, or from any family of over-segmentations. The system yields record accuracies on the Sift Flow Dataset (33 classes) and the Barcelona Dataset (170 classes) and near-record accuracy on Stanford Background Dataset (8 classes), while being an order of magnitude faster than competing approaches, producing a 320 × 240 image labeling in less than a second, including feature extraction.

Clément Farabet, Camille Couprie, Laurent Najman and Yann LeCun proposed scene parsing method here starts by computing a tree of segments from a graph of pixel dissimilarities. Simultaneously, a set of dense feature vectors is computed which encodes regions of multiple sizes centered on each pixel. The feature extractor is a multiscale convolutional network trained from raw pixels. The feature vectors associated with the segments covered by each node in the tree are aggregated and fed to a classifier which produces an estimate of the distribution of object categories contained in the segment. A subset of tree nodes that cover the image are then selected so as to maximize the average "purity" of the class distributions, hence maximizing the overall likelihood that each segment will contain a single object. The convolutional network feature extractor is trained end-to-end from raw pixels, alleviating the need for engineered features. After training, the system is parameter free. The system yields record accuracies on the Stanford Background Dataset (8 classes), the Sift Flow Dataset (33 classes) and the Barcelona Dataset (170 classes) while being an order of magnitude faster than competing approaches, producing a 320 × 240 image labeling in less than 1 second.

Hongsheng Li, Rui Zhao, and Xiaogang Wang present highly efficient algorithms for performing forward and backward propagation of Convolutional Neural Network (CNN) for pixelwise classification on images. For pixelwise classification tasks, such as image segmentation and object detection, surrounding image patches are fed into CNN for predicting the classes of centered pixels via forward propagation and for updating CNN parameters via backward propagation. However, forward and backward propagation was originally designed for whole-image classification. Directly applying it to pixelwise classification in a patch-by-patch scanning manner is extremely inefficient, because surrounding patches of pixels have large overlaps, which lead to a lot of redundant computation. The proposed algorithms eliminate all the redundant computation in convolution and pooling on images by introducing novel d-regularly sparse kernels. It generates exactly the same results as those by patch-by-patch scanning. Convolution and pooling operations with such kernels are able to continuously access memory and can run efficiently on GPUs. A fraction of patches of interest can be chosen from each training image for backward propagation by applying a mask to the error map at the last CNN layer. Its computation complexity is constant with respect to the number of patches sampled from the image. Experiments have shown that our proposed algorithms speed up commonly used patch-by-patch scanning over 1500 times in both forward and backward propagation. The speedup increases with the sizes of images and patches. Source code of GPU implementation is ready to be released to the public.

The topic of semantic segmentation has witnessed considerable progress due to the powerful features learned by convolutional neural networks (CNNs). The current leading approaches for semantic segmentation exploit shape information by extracting CNN features from masked image regions. This strategy introduces artificial boundaries on the images and may impact the quality of the extracted features. Besides, the operations on the raw image domain require to compute thousands of networks on a single image, which is time-consuming.

In this paper, **Jifeng Dai, Kaiming He and Jian Sun** propose a method to exploit shape information via masking convolutional features. The proposal segments (e.g., super-pixels) are treated as masks on the convolutional feature maps. The CNN features of segments are directly masked out from these maps and used to train classifiers for recognition. They further propose a joint method to handle objects and "stuff" (e.g., grass, sky, water) in the same framework. State-of-the-art results are demonstrated on benchmarks of PASCAL VOC and new PASCALCONTEXT, with a compelling computational speed.

Jonathan Long, Evan Shelhamer and Trevor Darrell show that convolutional networks by themselves, trained end-to-end, pixelsto-pixels, exceed the state-of-the-art in semantic segmentation. Their key insight is to build "fully convolutional" networks that take input of arbitrary size and produce correspondingly-sized output with efficient inference and learning. They define and detail the space of fully convolutional networks, explain their application to spatially dense prediction tasks, and draw connections to prior models. They adapt contemporary classification networks (AlexNet, the VGG net, and GoogLeNet) into fully convolutional networks and transfer their learned representations by fine-tuning to the segmentation task. They then define a novel architecture that combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations. Their fully convolutional network achieves state-of-the-art segmentation of PASCAL VOC (20% relative improvement to 62.2% mean IU on 2012), NYUDv2, and SIFT Flow, while inference takes less than one fifth of a second for a typical image.

Jose M. Alvarez, Yann LeCun, TheoGevers and Antonio M. Lopez referred to the process of assigning an object label (e.g., building, road, sidewalk, car, pedestrian) to every pixel in an image. Common approaches formulate the task as a random field labeling problem modeling the interactions between labels by combining local and contextual features such as color, depth, edges, SIFT or HoG. These models are trained to maximize the likelihood of the correct classification given a training set. However, these approaches rely on hand–designed features (e.g., texture, SIFT or HoG) and a higher computational time required in the inference process. Therefore, in this paper, we focus on estimating the unary potentials of a conditional random field via ensembles of learned features.We propose an algorithm based on convolutional neural networks to learn local features from training data at different scales and resolutions. Then, diversification between these features is exploited using a weighted linear combination. Experiments on a publicly available database show the effectiveness of the proposed method to perform semantic road scene segmentation in still images. The algorithm outperforms appearance based methods and its performance is similar compared to state–of–the–art methods using other sources of information such as depth, motion or stereo.

Joseph J. Lim C. Lawrence Zitnick Piotr Doll´ar proposed a novel approach to both learning and detecting local contour-based representations for mid-level features. Their features, called sketch tokens, are learned using supervised mid-level information in the form of hand drawn contours in images. Patches of human generated contours are clustered to form sketch token classes and a random forest classifier is used for efficient detection in novel images. They demonstrate our approach on both top down and bottom-up tasks. They show state-of-the-art results on the top-down task of contour detection while being over 200_ faster than competing methods. They also achieve large improvements in detection accuracy for the bottom-up tasks of pedestrian and object detection as measured on INRIA and PASCAL, respectively. These gains are due to the complementary information provided by sketch tokensto low-level features such as gradient histograms.

Joao Carreira and **Cristian Sminchisescu** presented a novel framework for generating and ranking plausible objects hypotheses in an image using bottom up processes and mid-level cues. The object hypotheses are represented as figure-ground segmentations, and are extracted automatically, without prior knowledge about properties of individual object classes, by solving a sequence of constrained parametric min-cut problems (CPMC) on a regular image grid. They then learn to rank the object hypotheses by training a continuous model to predict how plausible the segments are, given their mid-level region properties. They show that this algorithm significantly outperforms the state of the art for low-level segmentation in the VOC09 segmentation dataset. It achieves the same average best segmentation covering as the best performing technique to date, 0.61 when using just the top 7 ranked segments, instead of the full hierarchy in. Their method achieves 0.78 average best covering using 154 segments. In a companion paper, they also show that the algorithm achieves state-of-the art results when used in a segmentation-based recognition pipeline.

Deep Convolutional Neural Networks (DCNNs) have recently shown state of the art performance in high level vision tasks, such as image classification and object detection. This work brings together methods from DCNNs and probabilistic graphical models for addressing the task of pixel-level classification (also called ”semantic image segmentation”).

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy and Alan L. Yuille show that responses at the final layer of DCNNs are not sufficiently localized for accurate object segmentation. This is due to the very invariance properties that make DCNNs good for high level tasks. They overcome this poor localization property of deep networks by combining the responses at the final DCNN layer with a fully connected Conditional Random Field (CRF). Qualitatively, their "DeepLab" system is able to localize segment boundaries at a level of accuracy which is beyond previous methods. Quantitatively, our method sets the new state-of-art at the PASCAL VOC-2012 semantic image segmentation task, reaching 71.6% IOU accuracy in the test set. They show how these results can be obtained efficiently: Careful network re-purposing and a novel application of the ’hole’ algorithm from the wavelet community allow dense computation of neural net responses at 8 frames per second on a modern GPU.

Mohammadreza Mostajabi, Payman Yadollahpour and Gregory Shakhnarovich introduce a purely feed-forward architecture for semantic segmentation. They map small image elements (superpixels) to rich feature representations extracted from a sequence of nested regions of increasing extent. These regions are obtained by ”zooming out” from the superpixel all the way to scene-level resolution. This approach exploits statistical structure in the image and in the label space without setting up explicit structured prediction mechanisms, and thus avoids complex and expensive inference. Instead superpixels are classified by a feedforward multilayer network. Their architecture achieves new state of the art performance in semantic segmentation, obtaining 64.4% average accuracy on the PASCAL VOC 2012 test set.

Ning Zhang et al. in his paper "Part based R-CNN's for fine grained category detection" showed that semantic part localization can facilitate ne-grained categorization by explicitly isolating subtle appearance dierences associated with specifc object parts. Methods for pose-normalized representations have been proposed, but generally presume bounding box annotations at test time due to the difficulty of object detection. They proposed a model for ne-grained categorization that overcomes these limitations by leveraging deep convolutional features computed on bottom-up region proposals. Their method learns whole object and part detectors, enforces learned geometric constraints between them, and predicts a ne-grained category from a pose-normalized representation. Experiments on the Caltech- UCSD bird dataset conrm that this method outperforms state-of-the-art ne-grained categorization methods in an end-to-end evaluation without requiring a bounding box at test time.

Pablo Arbeaez, Bharath Hariharau, Chunhui Gu, Saurabh Gupta, Lubomir Bourdev and Jitendra Malik addressed the problem of segmenting and recognizing objects in real world images, focusing on challenging articulated categories such as humans and other animals. For this purpose, they propose a novel design for region-based object detectors that integrates efficiently top-down information from scanning-windows part models and global appearance cues. Their detectors produce class-specific scores for bottom-up regions, and then aggregate the votes of multiple overlapping candidates through pixel classification. They evaluate our approach on the PASCAL segmentation challenge, and report competitive performance with respect to current leading techniques. On VOC2010, their method obtains the best results in 6/20 categories and the highest performance on articulated objects.

Pedro H. O. Pinheiro Ronan propose an approach consisting of a recurrent convolutional neural network which allows them to consider a large input context, while limiting the capacity of the model. Collobert Scene parsing is a technique that consist on giving a label to all pixels in an image according to the class they belong to. To ensure a good visual coherence and a high class accuracy, it is essential for a scene parser to capture image long range dependencies. In a feed-forward architecture, this can be simply achieved by considering a sufficiently large input context patch, around each pixel to be labeled. Contrary to most standard approaches, our method does not rely on any segmentation methods, nor any task-specific features. The system is trained in an end-to-end manner over raw pixels, and models complex spatial dependencies with low inference cost. As the context size increases with the built-in recurrence, the system identifies and corrects its own errors. Their approach yields state-of-the-art performance on both the Stanford Background Dataset and the SIFT Flow Dataset, while remaining very fast at test time.

Philipp Kr¨ahenb¨uhl, Vladlen Koltun, Most state-of-the-art techniques for multi-class image segmentation and labeling use conditional random fields (CRF) defined over pixels or image regions. While region-level models often feature dense pairwise connectivity, pixel-level models are considerably larger and have only permitted sparse graph structures. The paper considers fully connected CRF models defined on the complete set of pixels in an image. The resulting graphs have billions of edges, making traditional inference algorithms impractical. The contribution is a highly efficient approximate inference algorithm for fully connected CRF models in which the pairwise edge potentials are defined by a linear combination of Gaussian kernels. Experiments demonstrates that dense connectivity at the pixel level substantially improves segmentation and labeling accuracy.

Pierre Sermanet David Eigen , Xiang Zhang Michael Mathieu Rob Fergus Yann LeCun present an integrated framework for using Convolutional Networks for classification, localization and detection. They show how a multiscale and sliding window approach can be efficiently implemented within a ConvNet. They also introduce a novel deep learning approach to localization by learning to predict object boundaries. Bounding boxes are then accumulated rather than suppressed in order to increase detection confidence. They show that different tasks can be learned simultaneously using a single shared network. This integrated framework is the winner of the localization task of the ImageNet Large Scale Visual

Recognition Challenge 2013 (ILSVRC2013) and obtained very competitive results for the detection and classifications tasks. In post-competition work, they establish a new state of the art for the detection task. Finally, they release a feature extractor from our best model called OverFeat.

Rainer Lienhart and Axel Wernicke propose a novel method for localizing and segmenting text in complex images and videos. Text lines are identified by using a complex-valued multilayer feed-forward network trained to detect text at a fixed scale and position. The network's output at all scales and positions is integrated into a single text-saliency map, serving as a starting point for candidate text lines. In the case of video, these candidate text lines are refined by exploiting the temporal redundancy of text in video. Localized text lines are then scaled to a fixed height of 100 pixels and segmented into a binary image with black characters on white background. For videos, temporal redundancy is exploited to improve segmentation performance. Input images and videos can be of any size due to a true multiresolution approach. Moreover, the system is not only able to locate and segment text occurrences into large binary images, but is also able to track each text line with sub-pixel accuracy over the entire occurrence in a video, so that one text bitmap is created for all instances of that text line. Therefore, their text segmentation results can also be used for object-based video encoding such as that enabled by MPEG-4.

Richard Socher richard, Cliff Chiung-Yu Lin, Andrew Y. Ng, and Christopher D. Manning introduce a max-margin structure prediction architecture based on recursive neural networks that can successfully recover such structure both in complex scene images as well as sentences. The same algorithm can be used both to provide a competitive syntactic parser for natural language sentences from the Penn Treebank and to outperform alternative approaches for semantic scene segmentation, annotation and classification. Recursive structure is commonly found in the inputs of different modalities such as natural scene images or natural language sentences. Discovering this recursive structure helps us to not only identify the units that an image or sentence contains but also how they interact to form a whole. For segmentation and annotation their algorithm obtains a new level of state-of-theart performance on the Stanford background dataset (78.1%). The features from the image parse tree outperform Gist descriptors for scene classification by 4%.

Ross Girshick in 2015 presented the technique "Fast R-CNN" . In his paper he proposed Fast R-CNN, a clean and fast framework for object detection. Compared to traditional R-CNN, and its accelerated version SPPnet, Fast R-CNN trains networks using a multi-task loss in a single training stage. The multi-task loss simplifies learning and improves detection accuracy. Unlike SPPnet, all network layers can be updated during fine- uning. They show that this difference has practical ramifications for very deep networks, such as VGG16, where mAP suffers when only the fully-connected layers are updated. Compared to "slow" R-CNN, Fast RCNN is 9 faster at training VGG16 for detection, 213 faster at test-time, and achieves a significantly higher mAP on PASCAL VOC 2012. Compared to SPPnet, Fast R-CNN trains VGG16 3 faster, tests 10 faster, and is more accurate.

Object detection performance, as measured on the canonical PASCAL VOC dataset, has plateaued in the last few years. The best-performing methods are complex ensemble systems that typically combine multiple low-level image features with high-level context. In this paper, **Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik** propose a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 30% relative to the previous best result on VOC 2012—achieving a mAP of 53.3%. Their approach combines two key insights: (1) one can apply high-capacity convolutional neural networks (CNNs) to bottom-up region proposals in order to localize and segment objects and (2) when labeled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significant performance boost. Since They combine region proposals with CNNs, we call our method R-CNN: Regions with CNN features. We also compare R-CNN to OverFeat, a recently proposed sliding-window detector based on a similar CNN architecture. They find that R-CNN outperforms OverFeat by a large margin on the 200-class ILSVRC2013 detection dataset.

Shaoqing Ren et. al published paper on Faster R-CNN, using it for Real-Time Object Detection with Region Proposal Networks. State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations. Advances like SPPnet and Fast R-CNN have reduced the running time of these detection networks, exposing region proposal computation as a bottleneck. In this work, they introduced a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position. RPNs are trained end-to-end to generate high-quality region proposals, which are used by Fast R-CNN for detection. With a simple alternating optimization, RPN and Fast R-CNN can be trained to share convolutional features. For the very deep VGG-16 model, their detection system has a frame rate of 5fps (including all steps) on a GPU, while achieving state-of-the-art object detection accuracy on PASCAL VOC 2007 (73.2% mAP) and 2012 (70.4% mAP) using 300 proposals per image.

Shuai Zheng Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su,Dalong Du, Chang Huang, and Philip H. S. Torr Pixel-level labeling tasks, such as semantic segmentation, play a central role in image understanding. Recent approaches have attempted to harness the capabilities of deep learning techniques for image recognition to tackle pixellevellabelling tasks. One central issue in this methodology is the limited capacity of deep learning techniques to delineate visual objects. To solve this problem, we introduce a new form of convolutional neural network that combines the strengths of Convolutional Neural Networks (CNNs) and Conditional Random Fields

(CRFs)-based probabilistic graphical modelling. To this end, Conditional Random Fields as Recurrent Neural Networks is formulated. This network, called CRF-RNN, is then plugged in as a part of a CNN to obtain a deep network that has desirable properties of both CNNs and CRFs. Importantly, the system fully integrates CRF modelling with CNNs, making it possible to train the whole deep network end-to-end with the usual back-propagation algorithm, avoiding offline post processing methods for object delineation. The proposed method to the problem of semantic image segmentation, obtaining top results on the challenging Pascal VOC 2012 segmentation benchmark.

S. Ji and H.W. Park proposed Two-step image segmentation algorithm, which is based on region coherency for the segmentation of color image. The first step is the

watershed segmentation, and the next one is the region merging using artificial neural networks. Spatially homogeneous regions are obtained by the first step, but the regions are over segmented. The second step merges the over segmented regions. The proposed method exploits the luminance and chrominance difference components of color image to verifv region coherency. The YUV color coordinate system is used in this work.

Graph cut optimization is one of the standard workhorses of image segmentation since for binary random field representations of the image, it gives globally optimal results and there are efficient polynomial time implementations. Often, the random field is applied over a flat partitioning of the image into non-intersecting elements, such as pixels or super-pixels. In the paper **Victor Lempitsky, Andrea Vedaldi and Andrew Zisserman** show that if, instead of a flat partitioning, the image is represented by a hierarchical segmentation tree, then the resulting energy combining unary and boundary terms can still be optimized using graph cut (with all the corresponding benefits of global optimality and efficiency). As a result of such inference, the image gets partitioned into a set of segments that may come from different layers of the tree. They apply this formulation, which they call the pylon model, to the task of semantic segmentation where the goal is to separate an image into areas belonging to different semantic classes. The experiments highlight the advantage of inference on a segmentation tree (over a flat partitioning) and demonstrate that the optimization in the pylon model is able to flexibly choose the level of segmentation across the image. Overall, the proposed system has superior segmentation accuracy on several datasets (Graz-02, Stanford background) compared to previously suggested approaches.

3. Problem Statement

Semantic image segmentation refers to the problem of assigning a semantic label (such as “person”, “car” or “dog”) to every pixel in the image. **Semantic segmentation** (or pixel classification) associates one of the pre-defined class labels to each pixel. The input image is divided into the regions, which correspond to the objects of the scene or "stuff" (in terms of Heitz and Koller (2008)). In the simplest case pixels are classified w.r.t. their local features, such as colour and/or texture features (Shotton et al., 2006). Markov Random Fields could be used to incorporate inter-pixel relations. To evaluate the performance of our proposed approach we are going to use performance parameter named "**mean intersection-over-union (IOU) score** ". For each class, Intersection over Union (IU) score is=true positive / (true positive + false positive + false negative).

True positives: the number of correctly classified pixels.

False positives: the number of pixels wrongly classified.

False negatives: the number of pixels wrongly not classifed.

We will try to achieve a mean intersection-over-union (IOU) score of 60-70% on PASCAL VOC segmentation benchmark.