Chapter 2

Literature and Related work

2.1 Survey on Scene labelling

In this chapter, we have surveyed the related state-of-art work that has included the objectives of our thesis. The scene parsing based image segmentation is widely used to detect the predefined objects in an image. Scene parsing or semantic classification is a process of labelling each pixel of an image to a semantic category. Semantic image labelling applications are widespread in 3-D images, video processing, still 2-D images, and volumetric data-sets. By looking at the broader prospect, the scene parsing is a core area in computer vision that leads to a holistic understanding of an image. An increasing number of applications have paved the way to extract complete information about objects in a scene. Scene parsing and semantic segmentation based applications have conducted in the past by using traditional machine learning applications. But deep learning-based applications that have achieved tremendous success in terms of accuracy and time has now turned the tables. Therefore, we have divided the scene classification based parsing methods in two parts: (1) Traditional methods (2) Deep Learning based methods. Traditionally the old established streams of computer vision and machine learning are more mature towards the aspect of clarity than deep learning. The dynamically changing nature of this domain has resulted in a hard starting point. To navigate with the pace of evolution, we have found that semantic labelling is confusing because the sheer amount of literature has been produced in recent times. In this section, we have reviewed some of the renowned state-of-arts.

2.2 Scene Labelling/Parsing based techniques

Scene parsing for image data is one of the interesting problems in computer vision. Still, the problem becomes more interesting when complete image data is not available, such as point clouds Lidar data or image in-painting, and some image portions need to be estimated using graph and reconstruction techniques. These data-sets need to be interpolated to generate an image, and then scene parsing took place. In [30], a graph-based framework has been proposed for the generation and semantic parsing of images from point cloud to detect the streets. Similarly, [26] has suggested the non-parametric label transfer by using an adaptive fixed-size window. Scene parsing in some other perspectives [31, 12, 6], such as pyramidal , panoptic, and 3-D scene parsing with reconstruction, have been conducted efficiently by using global information in different regions, segregation of semantic and instance segmentation, and a holistic grammar for scene representation.

2.2.1 Machine Learning based techniques for Scene Labelling

In this section, we have discussed the classical machine learning-based approaches for scene parsing. In [18], an aerial image segmentation has been introduced that derived an intelligent grammar for rules derivation to segment the aerial objects in satellite scenes. This acknowledges and recognizes the vast number of objects in an image scene with proportionate and few rules. In [24], a geometric parsing of image scenes was performed by an optimization model that iteratively computes the edge groups in line segments, group them in state line, collect the lines as parallel families, and finally locate the horizon and zenith in that image. This process considers uncertainty in scene parsing. In some references like [25], a speech and natural language processing technique that uses the parsing graphs, has been used to generate the similar parsing graph for the image that holds the image objects with their spatial characteristics in the form of the scene object graph. Sometimes the machine learning-based classifiers are confounded for the complicated scenes due to their variable nature and occlusions. Then, an automatic relation learning method [3] has proposed that learns the interactions and apply them into a holistic interpretation for a scene labelling. In [3], a parsing algorithm for middle-level representation of natural images has been proposed in which image curves have grouped with image segmentation using Bayesian inferences. The work [10], has proposed a stochastic grammar-based approach that generates the synthetic 3-D scene and large two-dimensional images for the training samples of a machine learning-based process.

Machine learning based methods on Hyper-spectral images (HSI) : Hyperspectral images are the high dimensional and spectrally mixed images that contain a massive amount of information and objects within it. In [15], a pixel un-mixing based regression and matrix factorization methods has been proposed to un-mix the multi-spectral and hyper-spectral images both. The pixels are noisy due to sensor errors and class-mixing. Therefore, image de-noising has performed before the machine predictions. Recently, an adaptive and dictionary learning-based method [4] has proposed that considers the three main characteristics in predictions, namely: (1) sparsity in the spectral-spatial domain, (2) correlated spectral domain, and (3) auto-correlation in the space. Kernel-based filter methods such as [13], have used to transform the linear spaces to non-linear decision boundaries because hyper-spectral data is so much mixed in its space that can not be easily separable by linear kernels. Therefore, a mercer product was introduced to generate the non-linear decision boundary. Anomaly detection in HSI images can be performed by using a constrain adversarial auto-encoder [29] that represents the features in a better way.

2.2.2 Deep Learning based techniques for Scene Labelling

Deep learning-based methods have become one of the key application domains for semantic segmentation and scene parsing. Several data-sets have released for researchers to verify their algorithms. After the evolution of deep convolution neural networks, the semantic labelling and scene parsing have achieved the tremendous success [14]. In [1], an iris segmentation for bio-metric images has performed for low quality images. A fully convoluted deep-CNN has introduced to perform on several large public databases and the low-quality images. In medical imaging, some semi-supervised methods have also been used for brain tissue labelling and parsing [8]. Annotation of medical images is a difficult process; therefore, scene parsing took place with relatively small samples. In [2], a systematic review on fully connected deep CNN has introduced for efficient scene parsing and background subtraction. For background subtraction, the authors have used CDnet-2012 data-set and performed the task in two steps: (1) need of background subtraction and (2) the adequacy of deep neural network architectures. In [27], a data augmentation technique, used in the chapter-4, 6, and 7, have applied for the creation of large data-sets. Subsequently, we have performed on eye patches of the headset data-set. Such methods have sufficiently low complexity; therefore, they can be used in embedded frameworks. The work [7], has proposed an enhanced U-net architecture for biomedical image segmentation for multi-model architecture. The authors have developed a MultiResUNet architecture as a next-generation U-net for scene parsing and semantic segmentation.

2.2.3 Relaxation based techniques for Scene Labelling

Label Relaxation techniques are widely used to enhance the outcome of a predictive classifier. In most cases, the authors have used pott's model in relaxation, which is also modified as a conditional random field(CRF) or Markov random field(MRF). In this section, we have discussed some bench-mark work for label optimization and semantic scene relaxation-based techniques. In [16], a comprehensive survey on discrete and continuous labelling has performed by using pott's model. Both discrete and continuous approaches, such as MRF and partial differential equations (PDE), respectively, have been used in the study. Such optimizations are NP-Hard problems. Therefore, the only local solution can be expected. The technique for discrete and continuous problems has suggested as MRF and Primal-Dual solutions. Curve matching is an important aspect of many computer vision problems, therefore a relaxation guided curve matching method has also applied that matches the previously matched points to the next points to guide an image neighbourhood curve [19]. In [21], a kernel and spectral clustering methods have introduced with the MRF based label relaxation to improve the cluster interpretation of the image. In [11], discrete energy minimization techniques have been introduced to study the modern inference methods in the image. The variational approaches also perform the image fusion to preserve the salient information and improving the contrast [17]. The work [23], has proposed another CRF based variational approach to improve the clustering result by allocating a semantic label to each pixel and fusion of their spatial context. A superposing based approach with superpixels has introduced with a non-parametric method for scene parsing with given categories [22]. In [5], the authors have suggested a method to differentiate the real and occluded part of an image and remove the occluded background. An efficient feature matching with non-negative orthogonal relaxation (NOR) [9] has been performed so that an IQP(integer quadratic solving) has solved by one-to-one matching, which is also an NP-hard problem.

2.2.4 Evaluation Matrices and Validation

In this section, we have discussed the evaluation matrices for scene parsing and image labelling methods. If pixel-wise labelled ground-truth is available, then the semantically predicted image can be directly compared with ground-truth for accurate measurement. Image labelling tools can produce semantically labelled data-sets. Image labelling is a time-consuming process for RGB and Gray images. Therefore, we have used standard labelled data-sets such as pascal-voc and Sift-flow. However, algorithms can be tested on some other self-produced labelled data-sets also. In the case of hyperspectral images (HSI), the creation of a labelled ground-truth is costly for sensor-based high dimensional images. Therefore, we have tested our methods on the publicly available ground-truth HSI image sets. In this section, we have described the accuracy metrics for evaluating predictive frameworks in semantic image labelling.

2.2.4.1 Overall Accuracy (OA)

Overall accuracy details about the proportion of correctly predicted data and given reference data. It is denoted as percentages where 100% shows the perfect prediction by the classifier for reference testing set. OA is easy to understand for mapping the user and producers with accurate information. It is defined as:

$$OA = \frac{\sum_{i=1}^{c} p_{i,i}}{N} \tag{2.1}$$

Where $p_{i,i}$ N is the total number of pixels as:

$$N = \sum_{i=1}^{c} \sum_{j=1}^{c} p_{i,j} \tag{2.2}$$

2.2.4.2 Average Accuracy (AA)

Average accuracy is the class-wise average of the user's accuracy. AA provides the complete class-wise performance of classifier which includes each kind of prediction .i.e Tp(true positive),Tn(true negative), Fp(false positive), Fn(false negatives). AA is defined as:

$$AA = \frac{\sum_{i=1}^{c} \frac{c_{i,i}}{n_i}}{c} \tag{2.3}$$

2.2.4.3 Kappa value (K)

Kappa values are computed with the help of confusion or error matrices. Kappa is defined as:

$$K = \frac{ObservedAccuracy - AgreementChance}{1 - AgreementChance} \tag{2.4}$$

2.2.4.4 Intersection Over Union (IOU)

This matrix is efficient in deep learning-based applications where batch accuracy has been computed to estimate the epoch convergence. It is also known as the Jaccard index. In the Jaccard index, two same-sized data-sets, i.e., predictions and ground-truth, are mapped with each other to know the correct and wrong predictions. It is defined as:

$$IOU = \frac{1}{c}\sum_{i=1}^{c} \frac{c_{i,i}}{Tp1_i + Tp2_i - c_{i,i}} \tag{2.5}$$

where,$p_{i,j}$ is prediction in on and off diogonals of error matrix and the true positives is:

$$Tp_i = \sum_{j=1}^{c} p_{i,j} \tag{2.6}$$

2.2.4.5 Precision, Recall, and Fscore

Tp, Tn, Fp, and Fn are the true positive, true negative, false positive, and false-negative predictions then Precision, Recall, Kappa, and F-score can be defined as:

$$precision = \frac{Tp}{Tp + Fp} \tag{2.7}$$

$$recall = \frac{Tp}{Tp + Fn} \tag{2.8}$$

$$Fscore = 2 \times \frac{recall \times precision}{recall + precision} \tag{2.9}$$