Survey on Conditional Random Fields in Object Recognition

Marie Mellor Mohammad Bayat Asya Vitko

October 2023

Abstract

This survey gives an overview of such a method used in image segmentation tasks as Conditional Random Fields (CRFs). It also provides descriptions of two alternative methods used for the task: convolutional neural networks (CNNs) and maximum entropy Markov Models (MEMMs). The survey provides a comparison of the three methods in terms of performance, computational efficiency, and practicality. Finally, the survey outlines applications of CRFs in the field of medical image segmentation.

1 Introduction

Object recognition is technology related to computer visions in which images or videos are processed to identify objects. Unlike humans, computers do not process images and videos as we would. People can easily identify different visual details like knowing the difference between a cup and a plate whereas computers need to be taught how to recognize what an image might contain.

In robotics, object recognition is used in a multitude of ways. Our focus was on object recognition in autonomous robotics. Many people are familiar with robot vacuum cleaners like the Roomba from iRobot. Autonomous robots like these provide some type of assistance or service to people, in the case of a robot vacuum it cleans a space all on its own without the need for user interference. But for this to work, the vacuum needs to be able to navigate throughout your home without disturbing you or the objects in your house. Object recognition in these sorts of systems allows these vacuums to traverse throughout a space and know what obstacles might be in their path so that they may go around.

Some of the more popular algorithms for object recognition are You Only Look Once (YOLO), Single-Shot Detector (SSD), and Region-based Convolutional Neural Networks (R-CNN)[4]. You Only Look Once, or YOLO is a real-time object recognition algorithm that is used for classification, image segmentation, and object detection. It aims to accurately detect and recognize objects with quicker results than previous models. YOLO utilizes a one-stage model for detection and processes an image in a single iteration to achieve relevant results. Single-shot Detector (SSD) is another one-stage detection model that focuses on identifying and predicting different classes and objects. It adapts the output spaces of bounding boxes in images and video frames using a deep neural network to generate a score for an object category. Region-based Convolutional Neural Networks (R-CNN) apply deep learning models for object recognition by selecting regions of an image and giving the model pre-defined labels. These labels are then used to categorize the objects in the image.

When comparing the algorithms, it is important to note that both YOLO and SSD have issues detecting smaller objects which can lower performance. R-CNN is much more accurate than both but is also considerably slower due to it being computationally expensive [3]. Each one of the object recognition algorithms has it's own bugs, the best-suited algorithm to use is dependant on the type of image being analyzed or the desired outcome. For black and white images, YOLO would be best. For complex images that require the best accuracy, R-CNN would be the best.

2 What are Conditional Random Fields

In the context of Conditional Random Fields (CRFs), CRFs arrange the object categories needed by utilizing parameters associated with features used to describe them, ie: color, shape, size etc. The number of object categories is directly related to the number of parameters which results in more complex models

which in turn, require more training as well as a more challenging design. More categories also mean that there must be a better distinction between them so that items with similar features but of different categories are not improperly classified.

The way that Conditional Random Fields model object recognition is through probability distribution P(y|x) [13]. This returns the probability of every different possible value to the random variables in y that an object can be classified as from x with the goal of finding the classification, or assignment to y, with the highest probability. Because objects cannot only be classified by one specific feature, the variables of an object must be broken down. For example, you cannot determine if an animal is a cat or a dog based only on the fact that it has eyes, ears, and nose. You would require more specific details to properly come to the conclusion of what animal it is. Similarly, CRFs rely on the assumptions from the random variables to break down the classification into smaller pieces.

The following figure by Ruiz-Sarmineto et al.[13] shows an example of object recognition using conditional random fields.

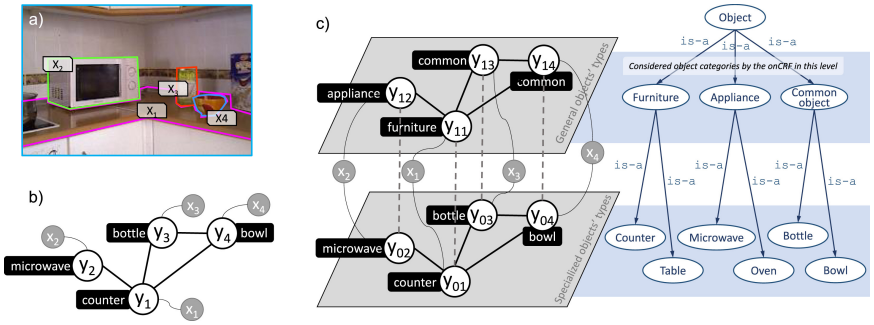

Figure 1: Fig.1 (a) scene capturing part of a kitchen with the observed objects labeled (x1, . . . , x4). (b) Standard CRF graph built from that scene, including as many nodes(y1, . . . , y4) as objects. (c) obCRF graph of the scene in (a) with two levels, indicating on the right the object categories considered at each level from an ontology.[13]

3 Comparison of CRFs to Some Alternative Object Recognition Methods

In this section we will briefly outline alternative methods to CRFs and sketch differences between them and CRFs in terms of performance, computational efficiency, practicality. Performance is a measure of how well a method achieves the goal of recognizing an object. Computational efficiency is a measure of the amount of resources (e.g. time) required to achieve the goal. Practicality is a measure of how easy or difficult it is to implement the method in practice.

There are two alternative methods that we will focus on: convolutional neural networks (CNNs), which are a type of neural networks, and maximum entropy Markov models (MEMM).

In what follows, we will first briefly outline what neural networks and MEMMs are, and then move on to comparing these methods to CRFs.

3.1 Convolutional Neural Networks

There is a large body of work on neural networks, and different types of neural networks exist for different tasks. In the field of image recognition such types of neural networks as convolutional neural networks (CNNs) [11], recurrent neural networks (RNNs) [10], recurrent convolutional neural networks (RCNNs) [7] achieve state-of-the-art results. We focus on comparing CRFs to such type of neural networks as CNNs, that is why provide an overview of what those are below.

The crucial difference of CNNs from other types of neural networks is that they use additional convolution layers. A convolution layer performs some operations on the input data, like multiplication, addition, etc., that transform the input by extracting patterns from the data. The result of the transformation by convolution layers is a so-called feature map. Feature maps are then used by fully connected layers, typical for NNs, to do final classification and recognition tasks. These layers are called fully connected, because in those layers every neuron in one layer is connected to every neuron in another layer. This is not necessarily true for convolution layers.

One of the immediate benefits of CNNs is that the convolution layer they use can learn hierarchies that may exist within features. For example, consider an image of a face. Initial convolution layers of a model may learn simple features present in an image of a face, like edges and color gradients. Another level of convolution layers may learn more abstract features, like facial contours: eyes, nose, lips, etc. Yet another level of convolution layers may be responsible for constructing feature maps that represent facial expressions. After all feature maps are extracted, they are flattened into one-dimensional layers, that is then passed on to fully connected layers for final classification or recognition.

One of the advantages that CNNs have over other types of neural networks is that they are robust to geometrical transformations [6], which means that they can recognize patterns even if those are scaled, or shifted, or rotated. This robustness is achieved by the fact that each neuron in convolution layer depends on a small region of an image, and thus is responsible for local features. This makes learning a feature less sensitive to its location in an image. Furthermore, when convolution is applied to an image, feature maps extract features from everywhere in an image and apply same weights to them no matter the origin of a feature.

3.2 Maximum Entropy Markov Models

Maximum entropy Markov Models are a class of techniques used for image segmentation tasks. They model conditional probability of a particular label to an object at hand given previous labels. Intuitively, the algorithm is trying to maximize the likelihood of a particular choice of labeling given the observed data.

Maximum entropy Markov Models use training data to find a distribution on labels that maximizes entropy and satisfies constraints of the data. Entropy of a distribution is computed as a negative sum over probabilities of each element in the state space weighted by the logarithm of that probability. Constraints from the data may take a form of expected values of certain features. Then the goal becomes to find weights for each constraint in such a way that the entropy of the distribution is maximized.

Maximum entropy Markov Models have an advantage over a similar technique called Hidden Markov Models (HMMs) in that more flexible modeling of the relationships between labels and the surrounding context is allowed [9]. Both MEMMs and CRFs are so-called discriminative models in contrast to HMMs which are generative models [9]. The crucial difference between these two types of models is in that discriminative models focus on assigning lables while generative models focus on how data was generated.

We provide an example of how MEMM can be used in an image segmentation task. An input to MEMM is an image, and it is represented as a grid, where each cell is a pixel in the image. Each cell is treated as an observation to be classified. States of MEMM correspond to different labels to be assigned to a pixel, like "background", "object", etc. Then transition probabilities are defined between states based on the content of each pixel. These are the likelihoods of a neighboring state receiving a particular label given the current label. Furthermore, observation probabilities are defined for each state based on the feature in the pixel. These are the likelihoods of a particular label being assigned based on the feature in a cell. The training process for MEMM consists in finding probabilities of labels.

3.3 Comparison of NNs, MEMMs, and CRFs

3.3.1 Performance

The main drawback of MEMMs is that they may exhibit label bias [14]. This means that MEMMs might assign higher probability to labels that occur frequently in the data although it might not be appropriate for the context. The reason for label bias is on a high level that MEMMs (being Markov models) are local and cannot capture global relations that may exist in the data. Going back to the example of how MEMMs operate, each label depends only on its left neighbor in the sequence of cells and on the feature in its corresponding cell. Even intuitively this seems like a very small scope to base inference of the label value from. CRFs improve over MEMMs in that they relax the assumption of

conditional independence of the observed data [12]. Now it may be that features in the neighboring cells effect weights of labels to a current cell.

Still, the main drawback of CRFs and MEMMs is that they are local (CRFs, being essentially Markov models, do not move beyond small regions of an image when making inferences about them). Thus, these models may fail to capture global dependencies in the image or hierarchical features which arise when the task at hand involves recognizing objects placed in some context [5]. To the contrary, NNs, in particular CNNs, overcome these challenges, and are capable of capturing global relations that exist within raw data, as described in Section 3.1.

3.3.2 Computational Efficiency

Training CRFs and MEMMs on large-scale datasets can be computationally expensive. The reason is that the associated optimization problem, which is to compute the partition function (i.e., to find the normalization term that ensures that all probabilities in the model sum up to 1), may become very resource demanding, especially in object recognition tasks, where images may have large number of pixels [5]. In general, finding a partition function is a notoriously hard task: this is typically a #P-complete problem. In the context of counting problems, the partition function is the number of solutions to a problem. To give an example, finding the exact value of a partition function of the uniform distribution on perfect matchings in a bipartite graph is #P-complete [15].

Neural networks, for example CNNs, are also computationally demanding. They require a lot of memory and processing time. They work on machines that have CPU with GPU, and cannot be used on small memory devices like cellphones [1]. A significant amount of resources is spent by CNNs on the problem of finding local minima: a necessary task for updating the weights. Different optimization algorithms exist [6], but they do have large running times.

3.3.3 Practicality

On the one hand, NNs use hidden layers while CRFs can avoid that. This is desirable in practice because this allows one to avoid problems around finding local minima. Usually, the problems around local minima include the following. An algorithm used to search for local minima may exhibit slow convergence rate, which worsens computational efficiency and negatively effects practicality [6]. Furthermore, an algorithm may be sensitive to learning rate selection in a sense that success of an algorithm depends on a good choice of value of a parameter [6]. Another problem that often occurs in CNNs is a so-called vanishing gradient problem [6]. Specifically, when a network has many layers, as CNNs do, the weights are learnt through backpropagation. The values propagated from final layers to initial ones are gradients of a loss function that represent how much a particular weight should increase or decrease. Since there are so many layers, values of the gradient get multiplied more and more, and become very small, i.e., vanish. This results in that earlier layers of the network may

fail to learn anything meaningful about the input. Still, CRFs have their own optimization problem to overcome, which is the demanding computation of a partition function, as mentioned in the previous section.

On the other hand, CRFs and MEMMs depend on handcrafted features or feature engineering. It may be time consuming and even not that feasible to design effective features, so the performance of CRFs may be limited due to poor quality of the chosen features [5]. Contrary to this, NNs have a capability to learn features automatically. This is exactly the benefit of the layered structure and use of backpropagation.

4 Application of CRF in medical images

In many different computer vision applications, CRF is coupled with other machine learning algorithms for more accurate image segmentation [2]. In the medical field, the same approach has helped to make more sophisticated image processing. For example, for medical image analysis, initially, the image gets processed by an ML algorithm like Convolutional Neural Networks (CNN), which can extract more complex features and then pass these pre-processed features to the conditional random field algorithm for more accurate image segmentation.[2]

The quality of image segmentation is crucial in accurately dividing medical images into distinct regions, aiding in the diagnosis and treatment planning for various medical conditions [2]. In traditional image segmentation, the details that need to be separated in the images using the imaging machine are chosen manually. For example, the machine could be configured to focus on the intensity feature of the region [2]. However, this method isn't perfect and can miss important differences between two regions that have similar shades. In a research study on segmentation in medical images, an alternative CRF called, **Posterior CRF** was proposed which outperforms the existing methods.[2]

Posterior Conditional Random Fields is a new approach that combines the prediction of Convolutional Neural Networks for more accurate segmentation. The way this new approach works is that after the CNN extracts the various patterns in the image, it passes these pattern classifications to CRF which uses these learned features along with the spatial information it captures from the image to produce more accurate image segmentation than solely depending on the intensity and spatial features which could result in poor classification. [2]

Medical Image Analysis using

CT scans are crucial for measuring the diameter of the pulmonary artery and the ratio of the diameter of the pulmonary artery to the diameter of the aorta, which correlates with invasive measures of pulmonary artery pressure [8]. To illustrate how Posterior Conditional Random Fields (PCRF) can be applied to CT-Scan image segmentation, this section will discuss research done on aorta and pulmonary artery segmentation in non-contrast CT scans. In non-contrast

scans, the conventional CRF processes the segmentation based on the intensity values. Since the aorta and pulmonary artery share similar intensity values, it makes it difficult to make inferences between them. For this reason, the author argues that the features learned from CNN are more informative than just intensity value. PCRF uses these learned features to process the scans.[2]

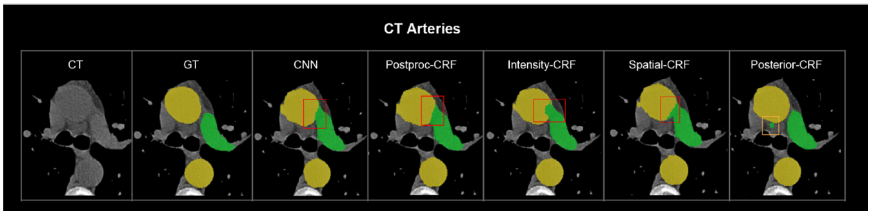

Figure 2: From left to right : (1) Original image (2) Manual annotation (3) CNN baseline (4) Postproc-CRF (5) Intensity-CRF (6) Spatial-CRF (7) Posterior-CRF. Yellow: Aorta Green: Pulmonary artery [2]

Figure 2 illustrates the comparison of different methods that are used for segmentation. By inspecting the results, it is clear that the boundary line or edges between the Aorta and pulmonary artery is more clear. In evaluating the PCRF method, the author mentions that the qualitative comparison between different methods demonstrates much better performance by having better segmentation of blurred boundaries and very small objects which is the power of PCRF compared to other methods[2]. The high qualitative performance of PCRF would provide medical professionals with more tools to investigate the underlying diseases more accurately and prescribe a medical treatment plan with higher confidence.

5 Conclusions

In this survey we described CRFs as an approach to image segmentation. We also described several alternative methods used for this task, such as CNNs and MEMMs. We compared CRFs, CNNs, and MEMMs based on their performance, computational efficiency, and practicality. The conclusions are that all the approaches are computationally demanding and suffer from certain practical complications, like computation of the partition function or problem of finding local minima. It is worth mentioning that CNNs is a state-of-the-art approach to image segmentation, i.e., it shows the best performance of all the methods considered here. MEMMs suffer from label bias. We also briefly outlined how CRFs are being used in medical image segmentation.

References

[1] Bansal, M., Kumar, M., and Kumar, M. 2d object recognition techniques: state-of-the-art work. *Archives of Computational Methods in Engineering 28* (2021), 1147–1161.

[2] Chen, S., Zhang, Z., Zhang, W., Liao, S., and Wang, L. An end-to-end approach to segmentation in medical images with cnn and posterior-crf. *Computerized Medical Imaging and Graphics 93* (2021), 101966.

[3] Devanathan, H. The basics of object detection: Yolo, ssd, r-cnn, Oct 2022.

[4] Hansen, U. S. Object detection: Models, use cases, examples, Apr 2023.

[5] Koller, D., and Friedman, N. *Probabilistic graphical models: principles and techniques*. 2009.

[6] Li, Z., Liu, F., Yang, W., Peng, S., and Zhou, J. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems* (2021).

[7] Liang, M., and Hu, X. Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 3367–3375.

[8] Matsuoka, S., Washko, G. R., Dransfield, M. T., Yamashiro, T., San Jose Estepar, R., Diaz, A., Silverman, E. K., Patz, S., and Hatabu, H. Pulmonary arterial enlargement and acute exacerbations of copd. *New England Journal of Medicine 367*, 10 (2012), 913–921.

[9] McCallum, A., Freitag, D., Pereira, F. C., et al. Maximum entropy markov models for information extraction and segmentation. In *Icml* (2000), vol. 17, pp. 591–598.

[10] Medsker, L. R., and Jain, L. Recurrent neural networks. *Design and Applications 5*, 64-67 (2001), 2.

[11] O'Shea, K., and Nash, R. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458* (2015).

[12] Quattoni, A., Collins, M., and Darrell, T. Conditional random fields for object recognition. *Advances in neural information processing systems 17* (2004).

[13] Ruiz-Sarmiento, J.-R., Galindo, C., Monroy, J., Moreno, F.-A., and Gonzalez-Jimenez, J. Ontology-based conditional random fields for object recognition. *Knowledge-Based Systems 168* (2019), 100–108.

[14] Sutton, C., McCallum, A., et al. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning 4*, 4 (2012), 267–373.

[15] Valiant, L. G. The complexity of computing the permanent. *Theoretical computer science 8*, 2 (1979), 189–201.