CS365A Project

Report

Face Parts Labelling

Submitted in partial fulfillment of
the requirements for the course CS365A

Bachelor of Technology
in
Computer Science and Engineering

Submitted by

11372	Krishna Chaitanya K
13788	Vikas Jain

Under the guidance of
Prof Amitabha Mukherjee

Department of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY KANPUR
Kanpur, U.P. , India – 208 016

Winter Semester 2014-15

Abstract

Face Segmentation and Labelling are mid-level computer vision tasks. Conditional Random Fields(CRFs) are used for the segmentation and labelling task. CRFs are very useful and efficient tools for building models which can segment and label images. They are particularly useful to model the local interactions among the adjacent regions (superpixels). However, the CRFs may have difficulty deciding the boundary between the regions when there is less or no distiction between the features of the region[3]. Therefore in this project, the Restricted Boltzmann Machines(RBMs) are used complemantary to the CRFs to provide realistic labelling to the image which it does by providing a shape prior. The model is termed as GLOC model (global + local)[3]. We compare labelling performance of GLOC model and CRFs on the part label database. The GLOC model produces better results than CRF model.

Acknowledgments

Every year CS365A - Artificial Intelligence Programming Course is offered in IITK. The course involves a course project to be completed by a team of two students under the guidance of the course instructor. The course project gives the opportunity to the students to have experience in latest AI techniques and its applications. It would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them.

The authors would like to express their gratitude towards **Prof Amitabha Mukherjee** for his kind co-operation and encouragement which helped us in the completion of this project.

The authors would like to express their special gratitude and thanks to course TAs **Shashwat Chandra**, **Triya Bhattacharya**, **Nimisha Agarwal** and **Rajesh Shubhankar** for giving us attention and time.

Krishna Chaitanya K
Vikas Jain

April 2015
Indian Institute of Technology Kanpur

Problem Definition

The task is to segment the provided image into smaller regions and label the regions into three parts namely hair, skin and background. In this problem, facial hair is also considered as part of hair and the neck is also considered a part of the skin.

Figure 1: Face Image to label[3]

Figure 2: Grund Truth Labelling of the image[3]

Contents

List of Figures

1 Introduction

Grouping and organising image regions into logical and consistent parts, which share same attributes, are critical mid-level computer vision tasks. The fundamental techniques involved are segmentation and labelling the regions.

Image Segmentation is the technique to divide an image into smaller regions (groups of pixels). And image Labelling is labellling those smaller regions into some logical and known labels[11].

In our project, image segmentation is based on superpixels(group of pixels) which are obtained according to the boundaries of image. Once the superpixels are obtained, these superpixels are labelled into three categories namely *hair*, *skin* and *background*. Facial hair is also given the *hair* label. The labelling is done using *GLOC* model[3] which is a hybrid model of *Conditional Random Fields (CRFs)*[5] and *Restricted Boltzmann Machines (RBMs)*[10] in which *CRFs* ensures local interaction of the superpixel regions and *RBMs* ensures global shape prior of the image as well.

We demostrate the above GLOC model on *part label database* and also on the random images taken from the internet.

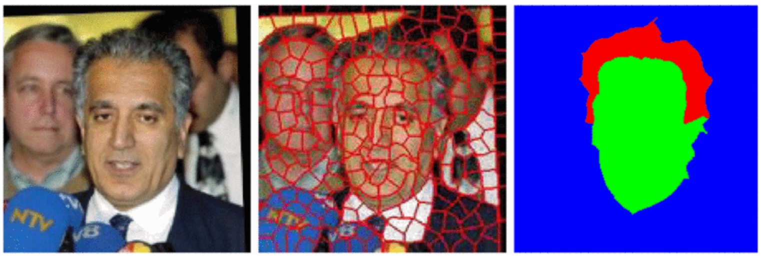

Figure 3: The left image shows a image from the dataset. The middle image shows the image segmented into superpixels. The right image shows the image labelled into three parts (hair, skin and background).[3]

2 Motivation

Face segmentation and labelling is an extremely important step in the recognition of faces because majority of face recognition methods work only with the labelled face images[4]. The overall performance and accuracy of a face recognition system depends on the correctness of the face area labelling, thus making face segmentation and labelling an extremely important task in a face recognition system[3]. The purpose of the face segmentation step is to extract the area containing the face from a given image which also contains other things.

In this work, we will segment the face into three regions and label them with hair, skin, and background labels suitably. This particular segmentation and labelling is chosen because in a paper by Huang[2], he remarked that a variety of high-level features, such as hair characteristics, gender and pose can often be deduced guided by the labelling of a face image into hair, skin and background segments.

The commonly used technique use CRFs alone. But, we used CRFs along with RBMs for better results. The motivation behind the strategy is primarily that CRFs fail when distinction between different region is not much. Hence RBMs provide global shape prior to overcome the problem[3].

3 Previous Work

A class of very popular and effective tools used for the segementation of images including the face segmentation are *CRFs (Conditional Random Fields)*[1]. The *CRFs*[5] are very powerful tools used for modelling the region boundaries by looking at the local interaction between adjacent regions which, in the scope of this project, are *superpixels*.

However, CRFs have a big limitation i.e. they do not deal with the issue of long distance interactions between superpixels which are not adjacent to each other[3]. This leads to problems in face parts labelling. For example, CRFs can help us clearly model the boundaries, in the case where there is sufficient distinction between the two adjacent regions which we want to seperate. But suppose there is insufficient distinction between the two regions, like in the case of the background having the same color as the skin tone of the person as shown in figure 4, the CRF model fails to label the image regions correctly as it is unable to distinguish between the background and the skin.

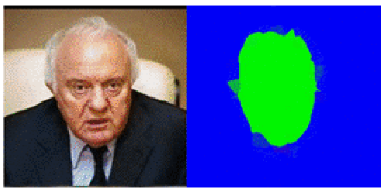

Figure 4: Example of CRF failure due to indistinct boundary[3]

There are several other methods which have been proposed for face segmentation. One of these methods[12] builds a model based on hair colour and then employs a region growing algorithm which modifies the hair region.

Another model [6] was built upon by making and training mixture models for color distributions of hair,skin and background.

4 Methodology

To overcome the drawbacks of the CRFs as described before, we make use of another graphical model called RBM(Restricted Boltzmann Machine)[3]. This graphical model is used to model the global shape of the skin, hair and background regions. These modeled object shapes act as priors which complement the working of the CRFs and help rule out erroneous labellings which do not meet the prior.

4.1 Preliminary Models

The model proposed is a hybrid model[3] of the following two graphical models :

4.1.1 Conditional Random Fields(CRFs)

The conditional random field[5] is a graphical model which as described earlier, models the local interaction between adjacent regions. This property makes it a very useful model for structured output prediction. It also finds a lot of application in computer vision. In keeping with the aim of this project, the conditional distribution and energy function is defined as the following[3]:

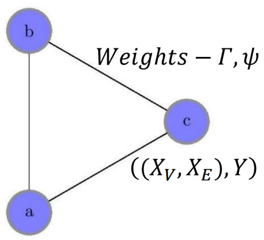

Figure 5: Graphical CRF Model

$$P_{\text{crf}}(\mathcal{Y}|\mathcal{X}) \propto \exp(-E_{\text{crf}}(\mathcal{Y}, \mathcal{X})) \tag{1}$$

$$E_{\text{crf}}(\mathcal{Y}, \mathcal{X}) = E_{\text{node}}(\mathcal{Y}, \mathcal{X}_{\mathcal{V}}) + E_{\text{edge}}(\mathcal{Y}, \mathcal{X}_{\mathcal{E}}) \tag{2}$$

$$E_{\text{node}}(\mathcal{Y}, \mathcal{X}_{\mathcal{V}}) = -\sum_{s \in \mathcal{V}} \sum_{l=1}^{L} \sum_{d=1}^{D_n} y_{sl} \Gamma_{ld} x_{sd} \tag{3}$$

$$E_{\text{edge}}(\mathcal{Y}, \mathcal{X}_{\mathcal{E}}) = -\sum_{(i,j)\in\mathcal{E}} \sum_{l,l'=1}^{L} \sum_{e=1}^{D_e} y_{il} y_{jl'} \Psi_{ll'e} x_{ije} \tag{4}$$

The meanings of the various notations is as follows:

- $\mathcal{Y}$: Region label.
- $\mathcal{X}_{\mathcal{V}}$: Region node feature vector.
- $\mathcal{X}_{\mathcal{E}}$: Region edge feature vector.
- $\mathcal{X}$: $(\mathcal{X}_{\mathcal{V}}, \mathcal{X}_{\mathcal{E}})$
- Ψ : Edge weights.
- Γ : Node weights.

The various parameters are trained to maximize the conditional log-likelihood of the training data[3].

$$\max_{\Gamma,\Psi} \sum_{m=1}^{M} \log P_{\text{crf}}(\mathcal{Y}^{(m)} | \mathcal{X}^{(m)}).$$

The general algorithms used for learning here are LBFGS and the mean-field approximation or loopy belief propagation(LBF) is used for inference[3].

4.1.2 Restricted Boltzman machine(RBM)

Restricted Boltzman machine[10] is a fully connected, bipartite and undirected graphical model. The nodes are divided into two parts: hidden and visible layers. The joint distribution in the scope of this project is defined as the following when there are R^2 visible nodes and K hidden nodes[3]:

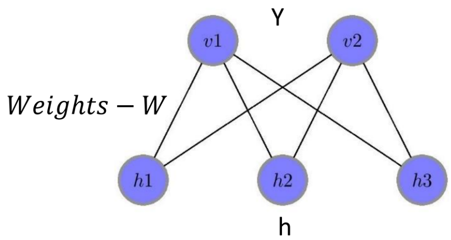

Figure 6: Graphical RBM Model

$$P_{\text{rbm}}(\mathcal{Y}, \mathbf{h}) \propto \exp(-E_{\text{rbm}}(\mathcal{Y}, \mathbf{h})) \tag{5}$$

$$E_{\text{rbm}}(\mathcal{Y}, \mathbf{h}) = -\sum_{r=1}^{R^2}\sum_{l=1}^{L}\sum_{k=1}^{K} y_{rl} W_{rlk} h_k - \sum_{k=1}^{K} b_k h_k - \sum_{r=1}^{R^2}\sum_{l=1}^{L} c_{rl} y_{rl} \tag{6}$$

The meanings of the various notations is as follows:

- $\mathcal{Y}$: Region label.
- $\mathbf{h}$: Hidden node label.
- W : Connection weights between hidden and visible nodes.
- b_k : Hidden bias
- c_{rl} : Visible bias.

Here again, the various parameters are trained to maximize the conditional log-likelihood of the training data[3].

$$\max_{W,b,c} \sum_{m=1}^{M} \log\left(\sum_{\mathbf{h}} P_{\text{rbm}}(\mathcal{Y}^{(m)}, \mathbf{h})\right).$$

The general algorithm used for learning here is stochastic gradient descent where the gradient can be approximated by using contrastive divergence[3].

4.2 GLOC Model

The final model combines the best of the CRFs (local consitency) and the RBMs(global shape prior) thus giving rise to its name GLOC(GLObal and LOCal)[3].

Mathematically, this model can be defined by its probability distribution given as follows[3]:

$$P_{\text{gloc}}(\mathcal{Y}|\mathcal{X}) \propto \sum_{\mathbf{h}} \exp(-E_{\text{gloc}}(\mathcal{Y}, \mathcal{X}, \mathbf{h})) \tag{7}$$

$$E_{\text{gloc}}(\mathcal{Y}, \mathcal{X}, \mathbf{h}) = E_{\text{crf}}(\mathcal{Y}, \mathcal{X}) + E_{\text{rbm}}(\mathcal{Y}, \mathbf{h}) \tag{8}$$

Where all the notations are as defined before.

We can see that the two models are combined by defining the energy function as the sum of the CRF and RBM energies. By combining the two energies, it physically translates into a case where the RBM model acts as a shape prior because if the RBM energy is high(meaning that the shape deviates from the prior), the total energy increases thereby decreasing the likelihood of that labeling scheme.

The model parameters (W, b, c, Ψ, Γ) are trained to maximize the conditional log likelihood of the training data[3].

$$\max_{W,b,c,\Gamma,\Psi} \sum_{m=1}^{M} \log P_{\mathrm{gloc}}(\mathcal{Y}^{(m)}|\mathcal{X}^{(m)}).$$

4.2.1 Virtual Pooling Layer

But, the RBM graphical model has a fixed number of visible nodes and the images in the dataset have slighly varying number of superpixels(200 to 250 per image).Therefore the existing model needs some modification[3].

This modification is done in the form of a virtual pooling layer[3]. The image is divided into an $(R \times R)$ grid. Now as shown in figure 7, the top 2 layers act as the RBM where the virtual pooling layer acts as the visible layer of the RBM. The labels and feature vectors of a node in the vitual pooling layer is determined by the superpixels of the image which overlap with the grid associated with that node. The extent of contribution is determined by its proportion of overlap area[3].

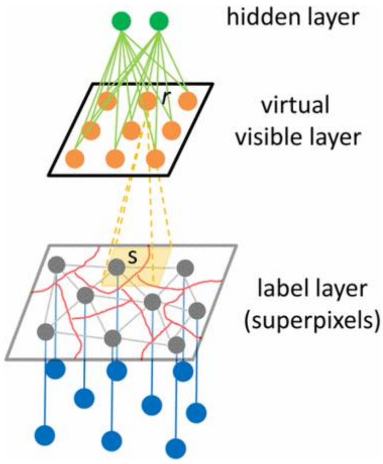

Figure 7: GLOC model with virtual pooling layer [3].

4.2.2 Features Used

The following features are used as stated in the paper [3]:

Node Features:

- Color: K-Means is run over the pixel space giving a normalized histogram over 64 bins.

- Texture: Normalized histogram over 64 textons which are generated according to [7].
- Position: Proportion of the superpixel within each grid of the image. The grids are formed by dividing the image into (8×8) equal parts.

Edge features:

- Sum of probabilities of boundary along the border.[8]
- Euclidean Distance between mean color histogram.
- Chi-squared distance between texture histograms.

4.3 Algorithm Used

4.3.1 Learning

- We maximize the conditional log likelihood using contrastive divergence[3].
- It relies on the approximation of the gradient of the log likelihood based on a short Markov chain.

4.3.2 Inference

- Since the joint inference of superpixel labels and hidden nodes is intractable, we use mean-field approximation[3].
- The approximated distribution is such that it minimizes the Kullback-Leibler distance between the approximate and original distribution.

5 Dataset Used

Part Label Database is used for learning and testing the GLOC model as well for the previous CRF model. This database contains labelings of 2927 face images into Hair/Skin/Background labels. We used 1500 images for training, 500 used for validation, and 927 used for testing[3]. `http://vis-www.cs.umass.edu/lfw/part_labels/`

We have also taken random 20 images from the internet and have tested the GLOC model on those images after resizing them to (250×250) to keep them consistent with the rest of the database.

6 Code Used

We have modified and made suitable changes in the following codes available for our dataset. The various files were suitably modified to add our own database of 20 images.

- Feature Generation: `http://vis-www.cs.umass.edu/code/gloc/gloc_features.zip`
- GLOC Model: `http://vis-www.cs.umass.edu/code/gloc/gloc.zip`

7 Results

The parameters used for the GLOC as well CRFs model are[3]:

- Number of Visible Nodes: 576(24×24)
- Number of Hidden Nodes: 400
- Image Size: 250×250
- Number of Superpixels: 200-250 per image

Accuracy is measured using the precentage of superpixels correctly labelled corresponding to ground truth labels.

7.1 On Part Labels Database

The total number of images are 2927 in the database which are divided in following part:

- **Training** : 1500
- **Test** : 500
- **Validation** : 927

The accuracy obtained on the database on different models are:

- CRFs Alone: 93.3356%
- GLOC Model: 94.946%

Error Reduction using GLOC: 25.39%

Successful Example:

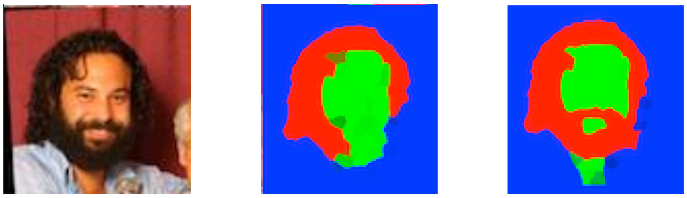

Figure 8: The left image shows a image from the dataset. The middle image shows the CRF result. The right image shows the GLOC result.[3]

7.2 On the Dataset generated

We have taken random 20 images from the internet and run CRF and GLOC Model on these images for testing and the accuracy obtained is: The total number of superpixels in the images: 4573

- CRFs Alone: 94.42% (4318 superpixels correctly labelled)
- GLOC Model: 95.17% (4352 superpixels correctly labelled)

The accuracy is better than the one on the part labels database as the images in our own database are more clear and the boundaries in the images are more distinct.

8 Future Work

Deep Boltzmann Machine(DBMs)[9] can be used in place of Restricted Boltzmann Machines. DBMs have deep architecture and have more layers involved. The results are expected to be better with DBMs but computation and training will be slow.

References

[1] Xuming He, Richard S Zemel, and MA 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*, volume 2, pages II–695. IEEE, 2004.

[2] Gary B Huang, Manjunath Narayana, and Erik Learned-Miller. Towards unconstrained face recognition. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.

[3] Andrew Kae, Kihyuk Sohn, Honglak Lee, and Erik Learned-Miller. Augmenting crfs with boltzmann machine shape priors for image labeling. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2019–2026. IEEE, 2013.

[4] Patrik Kamencay, Martina Zachariasova, Robert Hudec, Roman Jarina, Miroslav Benco, and Jan Hlubik. A novel approach to face recognition using image segmentation based on spca-knn method. *Radioengineering*, 22(1), 2013.

[5] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[6] Kuang-chih Lee, Dragomir Anguelov, Baris Sumengen, and Salih Burak Gokturk. Markov random field models for hair and face segmentation. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.

[7] Jitendra Malik, Serge Belongie, Jianbo Shi, and Thomas Leung. Textons, contours and regions: Cue integration in image segmentation. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 918–925. IEEE, 1999.

[8] David R Martin, Charless C Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using brightness and texture. In *Advances in Neural Information Processing Systems*, pages 1255–1262, 2002.

[9] Ruslan Salakhutdinov and Geoffrey E Hinton. Deep boltzmann machines. pages 448–455, 2009.

[10] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798. ACM, 2007.

[11] Wikipedia. Image segmentation — wikipedia, the free encyclopedia, 2015. [Online; accessed 16-March-2015].

[12] Yaser Yacoob and Larry S Davis. Detection and analysis of hair. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(7):1164–1169, 2006.