UNIVERSITY OF AMSTERDAM

Structural Image and Video Understanding in Computer Vision

Z. Lou

In this thesis, we have discussed how to exploit the structures in several computer vision topics. The five chapters addressed five computer vision topics using the image structures. In chapter 2, we proposed a structural model to jointly predict the age, expression and gender of a face. By modeling the facial regions with latent variables, we learn the relationship of different variables and improve the prediction accuracy of each tasks. In chapter 3, we proposed a framework to generate the 3D reconstruction from a single image. We firstly predict the geometrical structure of a scene, then generate the 3D layout of the image using the predicted geometrical structure as prior information. In chapter 4, we extract the primary object across videos. By building a graphical model on top of several videos, the primary object is extracted from each frame. In chapter 5, we using deep learning to learn the features and structures of images to predict the illuminant of scenes. Since deep learning benefits from large training dataset, we proposed a data augmentation method to generate large size of training images. In chapter 6, an image alignment algorithm is presented. In this work, we proposed a piecewise based algorithm to address the problem of aligning images with large view-point difference and non-planar assumption. In the following sections, we give more detailed summary of each chapters and conclude this thesis.

Chapter 2:Expression-Invariant Age Estimation Using Structured Learning

Accurately estimating the human age is vital for human facial analysis. However, in some cases, human faces are recorded under expressions. Those expressions highly harm the estimating accuracy because of the wrinkles exposed with expressions. To this end, we proposed a new algorithm to learn an expression-invariant age estimation predictor. In chapter 2, we jointly learn age and expression in a latent structural SVM framework. This is conducted by adding a latent layer between the image features and the output labels (i.e. age, expression).

The proposed algorithm is evaluated on three datasets (i.e. FACES, Lifespan and NEMO). Compared to independent learning algorithms which are trained for one single task, the proposed method improves for 14.43\%, 37.75\% and 9.3\% for the FACES, Lifespan and NEMO datasets respectively. We have also shown the model is very flexible for adding more tasks (e.g. gender estimation) for extension.

Chapter 3:Extracting 3D Layout from a Single Image Using Global Image Structures

Scene understanding is one of the key challenges in computer vision. We noticed that most of the scene images can be classified into a set of limited categories based on their geometrical structures. Therefore, to generate the 3D layout of a single image, we firstly predict the geometrical structure of the image. Then using the predicted geometrical structure as a prior, the 3D layout of an image is generated.

Instead of learning a direct mapping from the input image to the geometrical structures category, a set of templates have been used to model the subtopics of different geometrical subregions. With those templates, the relationship of the subtopics and the output type is learned. The experiments have shown that the classification accuracy improves 4.5\% comparing to other state-of-the-art algorithms. Using the predicted geometrical structural type, a random walk based segmentation algorithm is used to generate the 3D layout. To evaluate the proposed algorithm, we collect a large dataset with ground-truth layout labeling. The experiments show that the proposed algorithm outperforms other state-of-the-art algorithms by 11.7\%.

Chapter 4:Extracting Primary Objects by Video Co-Segmentation

Extracting objects from a video has many applications, such as video summary and video editing. Existing approaches mostly use motion and objectness features to extract object in a video. These traditional methods may fail in the case of multiple objects or still objects. Observing that one object may occurs in different videos, a structural model is exploited to model the intra and inter video relationships of the primary object. By inferring on the model, one object proposal is selected from each frame. Further, we build an appearance model on the selected object proposals and refine the segmentation on each frame.

The proposed algorithm is evaluated on two datasets and compared with other state-of-the-art algorithms. In comparison with other algorithms, it improves 1.5\% on MOViCS and 7.8\% on the New Sport Dataset. Beside better accuracy, the object extracted by the proposed algorithm has semantic meaning, namely the primary. By repeating the algorithm multiple times, multiple objects are extracted. This is very flexible for deceding the number of objects to extract.

Chapter 5:Color Constancy by Deep Learning

Traditionally, structural learning algorithms use handcrafted features with predefined structures for training. In this chapter, we have proposed a framework using deep learning for color constancy. The color constancy problem is formulated into a regression problem. A three-steps training strategy is proposed to learn better features and structures for color constancy. Since deep learning benefits from large size of training data, we proposed a method to generate more training image.

The proposed algorithm is evaluated on two widely used dataset, namely the Grayball dataset and the Colorchecker dataset. Compare to other algorithms, the proposed algorithm has mainly has two

advantage. Firstly, it is more accurate. On the Grayball dataset, the proposed algorithm improve the state-of-the-art by 9\%. On the Colorchecker dataset, the proposed algorithm has comparable result with other algorithm. Second, it is very efficient. Using GPU, the proposed algorithm can process image with more than 100 fps which means real-time performance.

Chapter 6: Image Alignment by Piecewise Planar Region Matching

Image is structurally organized. Using those structure, we align two images pixel by pixel. Traditional image alignment algorithms assuming affine transformation are constrained to planar scenes and small view point difference. In this chapter, we have proposed a new algorithm which can handle non-affine transformation and large view point difference. Specifically, one image is segmented into small pieces, affine transformation is applied on each piece. To make all the transformations on each piece consist to each other, a global constraint is exploited to ensure that the boundary point of each piece is tightly aligned with each other. The combined optimization problem is solved in a EM like framework.

The experiments show that the proposed algorithm is able to align two images with large view point difference. When the viewpoint difference is large, the proposed algorithm outperforms other algorithms. And it can also handle images with small view point difference. Some applications of this algorithm are also shown in the experiments. By aligning two images, we can edit one image using the content of another image which is recorded in another view point.