Structural image and video understanding
Lou, Z.

Citation for published version (APA):
Lou, Z. (2016). Structural image and video understanding

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
The composition of the world around us is not random. To a large extent, most scenes are well structured. For example, Figure 1.1 shows a typical image of my daily working place. Many structured entities are present in this image. At a larger scale, the image is composed of sky, background and (fore)ground. At a smaller scale, the composition of windows and doors are useful to interpret the type and architecture of the building. At an even smaller scale, trees and pedestrians contain their own characteristic compositions and structures. A structures appear everywhere in the world. The structure may be fixed such as faces or more ‘abstract’ such as cups. For humans to recognize a cup, it is not necessary that one has seen all cups before. Once the abstract structure of cups has been learned, a new cup can easily be recognized. Hence, structural information is important for automatic image and video understanding.

Figure 1.1: A typical image of my daily working place and its structures.

Structures refer to patterns and their relationships. For human faces, there are two eyes at the upper part, a nose in the middle and a mouth at the bottom part of a face. This ‘fixed’ composite structure helps us to recognize faces. Exploiting this type of structure, people can also estimate attributes of faces, such as age and gender. By comparing corresponding parts of different faces, one can differentiate between elderly people and kids.

Similar to faces, objects also has structures which are helpful in the understanding of images. Although objects may not have a ‘fixed’ structure like faces, the structure helps in differentiating one object from another. For example, a cup is normally a hollow
1. Introduction

cylinder with a bottom surface. Cups may have differences but a cup can not have the same structure as a chair, for example. Indeed, when we think about specific objects, such as cups, we have an ‘abstract’ structure in our mind. Exploiting this abstract structure is important for image understanding.

There is a long tradition in computer vision to exploit structural information for image understanding. In the early days of computer vision, shape matching has extensively been investigated. As shown in figure 1.2, objects such as elephants, fishes and airplanes are represented by simple, rigid templates. These approaches aim to detect objects by matching the objects with reference shapes or models. However, in a real world scenario, most of the structures are complex and non-rigid.

![Figure 1.2: Levels of structures investigated in this thesis.](image)

One step further is to consider objects composed of loosely connected components, such as faces [19, 67, 122]. Human faces contain parts such as two eyes (top), a nose (middle) and a mouth (bottom). These entities may vary in shape and size. As a consequence, many face recognition methods start with aligning faces to ensure that the corresponding parts of different faces are matching. After face alignment, machine learning pipelines are used to learn a mapping from facial features to attributes like age and gender. Directly learning a mapping may not be able to learn the structural information. So structural modeling is needed to better understand facial images.

Since the 1990’s, research has focused on general scene understanding containing looser compositions than faces. The work of [41, 104] translates scene understanding into a classification or clustering problem. The global geometry of scenes [68, 130] generates a 3D-layout from a single image. Providing training data with pixel-wise labeling, [72, 113] estimate the geometrical information of generic images. However, these algorithms are limited to indoor scenes and they need training data with pixel-wise labeling.

Image alignment [6, 11, 87–89, 92, 115, 119, 124] is another way exploiting image structures. Images from similar scenes usually have (a part of) common structures. In general, there are two types of correspondence algorithms for image alignment: parametric and non-parametric algorithms. Parametric algorithms assume that a global transformation model holds between two images. Key-points are used to estimate the parameters of the model. Non-parametric algorithms directly match pixels in two images. The two types of algorithms each have their advantages and disadvantages. Parametric models are good for matching images with a large variance as long as there are many key points present. Non-parametric algorithms match images as long as the variance is
small. To address the challenge of image alignment exploiting image structures, we perform piece-wise image alignment considering global structures by taking the advantage of both parametric and non-parametric algorithm.

Recently, research has been extended to generic video content understanding [2, 40, 85, 93, 109, 129, 143]. One of the main challenges in video understanding is the recognition of (moving) objects. Assuming that one object may occur in several videos, the inter-video structures and time-based structures are important for determining the primary objects across videos.

Although structural analysis is important for image and video understanding and many successes have already been achieved, there are still a number of aspects which need more attention. In human face analysis, when a classifier is learned to map image features [62, 63, 142] to attributes such as age, expression and gender, relationships are ignored between the states of local regions and the attributes of the image. Ignoring these relationships may lead to the inability to learn the hidden co-occurrence of local states and attributes. In this thesis, the aim is to model and learn implicit and explicit structures of images and videos.

Humans have the ability to estimate the age of a person based on facial appearance, even when showing a smile or surprise. However, for the current state-of-the-art computer vision, facial expressions strongly influence the results. Wrinkles caused by smiling may probably result in overestimating of the age of a person. To address this, our first research question is:

**RQ 1** Can we perform expression-invariant age estimation by exploiting the structural information of faces?

The research question is addressed in chapter 2. Facial appearance changes by expression-related muscles may overlap with aging-induced facial changes. To better understand the two types of changes, a graphical model is used to jointly learn age and expression. In this model, instead of predicting the age and expression directly from facial features, a set of latent variables is exploited to model the states of the sub-regions of a face. Then both age and expression are predicted based on the states of the latent variables and facial features.

In chapter 2, we can use a fixed structure to learn the relations between attributes like age and expression. However, general scenes may have looser structures. From a single image, it is difficult to extract the depth information directly from the pixel values. We revisited the research question of [100]:

**RQ 2** Is it possible to learn the scene structure and automatically generate the 3D layout from one single image?

We address this question in chapter 3. Extracting the pixel-level 3D-layout from a single image [68, 69, 72, 84, 86, 113, 130] is important for different applications such as object localization, image and video categorization. Traditionally, the 3D-layout is derived by solving a pixel-level classification problem. However, the image-level 3D-structure can be beneficial for extracting pixel-level 3D-layout since it implies the way how pixels in the image are organized. We propose an approach that firstly predicts the global image structure and then we use the global structure for fine-grained pixel-level
1. Introduction

3D-layout extraction. Specifically, image features are extracted based on multiple layout
templates. We then learn a discriminative model for classifying the global layout at the
image-level. By using latent variables, we implicitly model the sub-level semantics of the
image, which enrich the expressiveness of our model. After the image-level structure is
obtained, it is used as prior knowledge in a random walk based segmentation framework
to infer a pixel-wise 3D-layout.

So far, we have been discussed the use of structural information for static images.
Dynamic structures in videos are equally important for automatical video understanding,
and they are driven by objects. So, the next research question is:

RQ 3 Can we extract the primary object across videos using the dynamic structural
information?

The question is discussed in chapter 4. Video object segmentation is a challenging
problem. Without human annotation or other prior information, it is hard to select a
meaningful primary object from a single video. So extracting the primary object across
videos is a more promising approach. However, existing algorithms consider the problem
as foreground/background segmentation. Therefore, we propose an algorithm that learns
the model of the primary object by representing the frames/videos in a graphical model.
The probabilistic graphical model is built across a set of videos based on an object pro-
posal algorithm. Our approach considers appearance, spatial and temporal consistency
of primary objects.

In the previous chapters, we have shown that exploiting structural information is help-
ful for image and video understanding. For color constancy, various types of structural
information such as high-light edges and shadow are being used to predict the illuminant.
However, previous research shows that by using deep learning, features and structures are
learned automatically. Recent developments lead to the next research question.

RQ 4 Can we use deep neural nets to automatically learn features and structures to
predict the color of the illuminant?

The question is addressed in chapter 5. Computational color constancy aims to es-
timate the color of the light source. The performance of many vision tasks, such as
object detection and scene understanding, may benefit from color constancy by using
the corrected object colors. Since traditional color constancy methods are based on spe-
cific assumptions, none of those methods can be used as a universal predictor. Further,
shallow learning schemes are used for training-based color constancy, possibly suffering
from limited learning capabilities. In this chapter, we propose a new framework using
Deep Neural Networks to obtain accurate light source estimation. We reformulate color
constancy as a deep nets based regression approach to estimate the color of the light
source. The model is trained using datasets of more than a million images. Leveraging
on large training data and deep models, the proposed algorithm improve the performance
of color constancy.

For the previous chapters, we have focused on learning the structure from images and
subsequently applying the model to new images. However, even for two images of the
same scene, taken under different viewpoints, patterns can be recognized and matched.
Therefore, our last research question is:
RQ 5 Can we align two images under large viewpoint differences using local and global structures?

This question is addressed in chapter 6. Robust image registration is a challenging problem, especially when dealing with severe changes in illumination and viewpoint. Previous methods assume a global homography [115, 119]. They are able to align images under predefined constraints (i.e. planar scenes and parallax-free camera motion). These constraints may not hold for natural scenes and uncontrolled imaging conditions. Therefore, in this chapter, we propose a novel method which approximates image regions with planes by incorporating piecewise local geometric models. The approximated planar regions are obtained by exploiting a hierarchical figure-ground segmentation method. Each such planar region assumes an affine transformation. To achieve the alignment of the planar regions, an energy function is defined which employs intensity, key-point descriptor and geometric information under a global constraint. By re-segmenting and re-merging planar regions iteratively in an energy minimization framework, the method is able to align images even under significant changes in illumination and viewpoint.