Computer Vision and Image Understanding 131 (2015) 128–144

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Hierarchical Bayesian models for unsupervised scene understanding

Daniel M. Steinberg *, Oscar Pizarro, Stefan B. Williams

Australian Centre for Field Robotics, The University of Sydney, Sydney, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 21 November 2013
Accepted 10 June 2014
Available online 19 June 2014

Keywords:
Scene understanding
Unsupervised learning
Clustering
Hierarchical Bayesian models
Topic models
Variational Bayes

ABSTRACT

For very large datasets with more than a few classes, producing ground-truth data can represent a substantial, and potentially expensive, human effort. This is particularly evident when the datasets have been collected for a particular purpose, e.g. scientific inquiry, or by autonomous agents in novel and inaccessible environments. In these situations there is scope for the use of unsupervised approaches that can model collections of images and automatically summarise their content. To this end, we present novel hierarchical Bayesian models for image clustering, image segment clustering, and unsupervised scene understanding. The purpose of this investigation is to highlight and compare hierarchical structures for modelling context within images based on visual data alone. We also compare the unsupervised models with state-of-the-art supervised and weakly supervised models for image understanding. We show that some of the unsupervised models are competitive with the supervised and weakly supervised models on standard datasets. Finally, we demonstrate these unsupervised models working on a large dataset containing more than one hundred thousand images of the sea floor collected by a robot.

1. Introduction

In many real-world applications involving the collection of visual data, obtaining ground truth from a human expert can be very costly or even infeasible. For example, remote autonomous agents operating in novel environments like extra-planetary rovers and autonomous underwater vehicles (AUVs) are very effective at collecting huge quantities of visual data. Sending all of this data back to human operators quickly is hard since communication is usually bandwidth limited. In these situations it may be desirable to have algorithms operating on these vehicles that can summarise the data in unsupervised but semantically meaningful ways.

Similarly, many scientific datasets may contain terabytes of visual data that require expert knowledge to label it in a manner which is suitable for scientific inference. Obtaining such knowledge for large datasets can be a large drain on research resources. Again, it would be desirable to have algorithms that can separate this data automatically and in semantically meaningful ways, so the attention of the domain experts can be focused on subsets of the visual data for further labelling. In Section 6 we present a large visual dataset collected by an AUV that exhibits exactly this problem.

Recently there has been much focus on the computer vision problem of *scene understanding*, whereby multiple sources of information and various contextual relationships are used to create holistic scene models. Typically the aim in scene understanding is to improve scene recognition tasks while taking advantage of scene labels or annotations [1–4], accompanying caption or body text [5], or even contextual relationships between image labels and low-level visual features [6,7]. Most of these approaches are weakly supervised, semi-supervised or supervised in nature, and not much attention has been given to fully unsupervised, visual-data only holistic scene understanding.

In this article we wish to explore how unsupervised, or visual-data only, techniques can be applied to the problem of scene understanding. To this end we experiment with well established unsupervised models for clustering, such as Bayesian mixture models [8] and latent Dirichlet allocation [9]. These models cluster coarse whole-image descriptors, or cluster individual parts of images (but not simultaneously). We also explore models that can cluster data on multiple levels simultaneously (e.g. image segments or parts, and images), which are similar to the models presented in [4,10]. These models discover the relationships between objects in images, and then define scene types as distributions of these objects. Also, by knowing the scene type, contextual information is used to aid in finding objects within scenes. Finally we present a new model that can cluster multiple sources of visual information, such as segment and image descriptors. This model takes advantage of holistic image descriptors, which may encode

* Corresponding author.
E-mail addresses: d.steinberg@acfr.usyd.edu.au (D.M. Steinberg), o.pizarro@acfr.usyd.edu.au (O. Pizarro), stefanw@acfr.usyd.edu.au (S.B. Williams).

http://dx.doi.org/10.1016/j.cviu.2014.06.004
1077-3142/

spatial layout, as well as modelling scene types as distributions of objects.

All of these models are compared on standard computer vision datasets as well as a large AUV dataset for scene and object discovery. Emphasis is placed on scene category discovery, since we compare these unsupervised methods to state-of-the-art weakly-, semi- and supervised techniques for scene understanding. We also compare these models for object discovery in two of the experiments.

In the next section we review the most relevant literature to place this work in context. We then present the hierarchical Bayesian models we use for unsupervised scene understanding in Section 3 and in Section 4 we present variational Bayes algorithms for learning these models. In Section 5 we describe the image and image-segment descriptors we use, since these play a large part in the performance of these purely visual-data driven models. Then in Section 6 we empirically compare all of the aforementioned models, and summarise our results in Section 7.

2. Relevant literature

Visual context, such as the spatial structure of images, and position and co-occurrence of objects within scenes provide semantic information that aids object and scene recognition in our visual cortex [11,12]. Similarly, semantic information about images can be derived from the volumes of textual data that accompanies these images in the form of tags, captions and paragraph text on the Internet. Consequently, there has recently been a lot of research focusing on holistic image "understanding", where these sources of information are fused in order to improve scene and object recognition tasks.

An early attempt at combining annotation information with scene modelling proposed in [1] extends latent Dirichlet allocation (LDA) [9] to use both visual and textual data for inferring image tags in untagged images. This is essentially a "weak" form of supervision, where the exact image classes are unknown, but some semantically relevant information is still used in training the model. Subsequent research, [2–4,13–15] (amongst others) present hierarchical Bayesian models that can simultaneously classify scenes and recognise objects. These models can be supervised at the scene level, object level, or both. They can also use "weak" labels, or annotations, at the image or object levels [2–4]. Typically the features used to represent each image in these models are the proportions of super-pixel clusters (objects) contained within the images. The super-pixels are usually described by a combination of bag-of-words (BoW) features, such as quantised SIFT descriptors and quantised attributes like colour, texture and shape.

Li et al. [3] present a hierarchical Bayesian model that has a principled way of dealing with "noisy" or irrelevant object tags. Essentially a trade off is made between the model's certainty of the distribution of tags that correspond to a visual object class, and the distribution of tags that are irrelevant to the current object class. If an object class has a strong associated posterior distribution over the corresponding tags, a new tag that has low likelihood under this posterior is likely to be declared as irrelevant by an indicator variable. This model can also infer tags for images when they are missing. Quite a different approach to modelling textual and visual information is presented in [5]. They use a kernel canonical correlation analysis model (CCA) that attempts to learn the latent subspace that connects visual features with unaligned text (e.g. web pages with images). They then use this learned subspace for scene classification tasks.

Fei-Fei and Li [16] also present a model where the scene and object levels are classified in the same framework, but are linked through a higher "event" level, such as a particular sporting event. For example, the objects in an image may be a person, skis, etc., and the scene may be of a snowy mountain. Naturally, these are both related to a "skiing" event, which is simultaneously inferred. These higher level contextual relationships were shown to aid image classification. Similarly, it has been found in works such as [17,18] that knowledge of scene-type context can aid object recognition. In [17] the authors use a hidden Markov model (HMM) to classify a scene, and give certain objects a priori more probability of being detected conditioned on the scene type. For example, it is more likely you would find a coffee machine in a kitchen. Similarly, certain objects commonly co-occur, and so detection of one object (street) may be used to aid detection of another object (building), as demonstrated by [19]. They use tree-like models to infer the contextual and spatial relationships of, and between, labelled objects to aid inference in unlabelled test sets. It is worth noting at this juncture that while object discovery can be an important part of scene understanding, emphasis has usually been placed on scene recognition in the scene understanding literature. Object recognition and discovery performance is usually presented in a qualitative fashion. Conversely, scene recognition is not given much attention in the object recognition and discovery literature, which focuses on quantitative measures of recognition and object purity.

Many models for scene understanding explicitly model the spatial layout of scenes [14,15,19,20], or may make use of non-parametric processes or random fields to enforce segment-label contiguity [4,21–24]. In [14] the authors present a supervised model, the context-aware topic model (CA-TM), that is similar to hierarchical Bayesian models like [13], but it also learns the absolute (as opposed to relative) position of objects within a scene type. For example, it learns that sky objects are at the top, and buildings are at the sides, of street scenes, etc. Hence it takes advantage of both scene and spatial context for classification and object recognition. It can be both supervised at the scene and, optionally, at the object level. A model with a similar concept is presented in [20], however their emphasis is on object detection/discovery rather than scene recognition. The models presented in this article do not explicitly model scene spatial layout, however, the image descriptors used encode this information, see Section 5 for details.

Recently [6] combined Beta-process sparse-code dictionary learning, topic modelling and image classification in one generative framework. Essentially this framework models images from the pixel level to scene level. This is quite an impressive feat, and results in a *very* complex model. This model can also be used for unsupervised image clustering, but not necessarily object detection/segmentation. It can also use image annotations where available. While the classification results are impressive, each iteration of learning (Gibbs sampling) takes on the order of minutes, when it is usually milliseconds or seconds for other models. A similar concept is presented in [7], however they use a Bayesian co-clustering framework to incorporate semantic knowledge from image labels for visual dictionary learning. They can directly relate image features to semantic concepts, and show better performance than [6].

Scene understanding is a very active area of research, however much of the literature is concerned with weakly supervised, semi-supervised (a few strong labels) or supervised approaches to image understanding. Some of the aforementioned models can be used in a fully unsupervised, visual data only setting, though they may operate in a reduced capacity. For instance, the model in [13] loses its ability to perform scene recognition/discovery and reverts to just clustering segments when image labels are not present. Also [4,6] can be used as unsupervised models when no annotation data is available, however they were not rigorously tested in such situations. The only publication, to the authors' knowledge, that presents a model exclusively designed for unsupervised scene understanding is [25]. This model is also reviewed and used for comparison in this work.