Pattern Recognition Letters 48 (2014) 42–48

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Tower of Knowledge for scene interpretation: A survey ☆

Mai Xu *, Zulin Wang, Maria Petrou †

School of Electronic and Information Engineering, Beihang University, 37 Xueyuan Road, Beijing 100191, China

ARTICLE INFO

Article history:
Available online 24 February 2014

Keywords:
Tower of Knowledge
Scene interpretation
Computer vision

ABSTRACT

The past few decades have witnessed a wealth of promising work in making machines interpret the scenes around us. However, scene interpretation is still in its infancy, in comparison with human cognition. As such, human language, a highly developed output of human cognition, can be seen as an important cue towards scene interpretation. We survey in this paper Tower of Knowledge (ToK) approaches, which take advantage of human language, for scene interpretation. The core of ToK approaches is a multi-layer architecture, namely ToK architecture, aiming to establish the information flow of scene interpretation. In general, ToK architecture can be applied in scene interpretation by exploiting its either vertical or horizontal connections. First, we focus on the approaches with respect to the vertical connections in ToK architecture. In such approaches, the optimal label is assigned to each identified object in a scene, on the basis of verifying whether the object has the right characteristics to fulfil the functions a label implies. Second, we discuss the approaches on utilising the horizontal connections of ToK architecture to interpret a scene, according to the asymmetric spatial relationships of the objects. In retrospect of what has been achieved so far, we finally outlook what the future may hold for ToK.

1. Introduction

Computer vision as an ongoing research field is exciting and challenging today [1]. Its purpose is to make computers understand and describe the world from pictures or sequences of pictures, big enough to model a whole city, small enough to segment the cells of the human body. One of most important tasks of computer vision is scene interpretation, which aims at assigning labels or semantic representations to regions of a scene. The past few decades since the 1960s have seen extensive research efforts in this area. Now, it is surely clear that the development of a 3D/2D scene interpretation system is an immense research field, with its engineering applications in the vision systems of intelligent robots, the unmanned vehicles, etc. However, until most recently, it has not been possible to build a computer system to automatically interpret the whole world around us, despite computers and imaging systems having been improved greatly in the performance, with cheaper prices.

Towards the totally automatic scene interpretation, a wealth of statistical learning approaches have been the focus of much research [2–4], in particular using graphical probabilistic models, such as Bayesian networks (BN) [5–7] and Markov random fields (MRF) or conditional random fields (CRF) [8–10], for scene interpretation. For example, Fei-fei and Perona [6] and Li and Fei-fei [7] developed a type of Bayesian hierarchy model for learning and categorising objects, scenes and events. Beyond Bayesian approaches, MRF [8] was utilised to assign labels to specific regions of an image, with a proposed statistical model on exploring the spatial relationships between objects in a scene.

From 2008 onwards, researchers have made much progress in scene interpretation from the framework aspect. Prior to 2008, sub-problems of segmentation, semantic labelling and geometric reasoning were normally taken into account in isolation. After that, combining such sub-problems together in a unified framework, for scene interpretation, has attracted extensive research interests, e.g. cascaded classification model (CMM), either without feedback [11] or with feedback [12]. Besides, Shotton et al. [10] proposed a novel discriminative model on the basis of CRF, for simultaneously achieving segmentation and labelling results on natural scenes. Gould et al. proposed a scene decomposition approach [13] to integrate both scene appearance and structure into a region-based method with a unified energy function, for considering geometric and semantic consistency of regions in scene interpretation. Beyond segmentation, such an approach can work out semantic labelling and geometric reasoning simultaneously for a scene. Instead of relying on a global probabilistic model, stacked hierarchical labelling [14] was developed to train a hierarchical inference procedure, i.e. training a sequence of simple subproblems of

☆ This paper has been recommended for acceptance by Edwin Hancock.

* Corresponding author. Tel.: +86 18611630728; fax: +86 1082314663.
E-mail address: MaiXu@buaa.edu.cn (M. Xu).
† Deceased.

http://dx.doi.org/10.1016/j.patrec.2014.02.009
0167-8655/

segmentation, semantic labelling and geometric reasoning, in order to precisely interpret a scene.

Most recently, there has been a lot of work focusing on co-occurrence statistics in scene interpretation. For example, a novel framework [15] was developed to learn co-occurrence statistics between various label classes of the same region in images, in order to obtain a kind of many-to-many relationships (e.g. "roads are horizontal") for the semantic labelling in scene interpretation. Another type of object co-occurrence statistics [16] has been utilised in scene interpretation, benefitting from figuring out which classes (such as chair or motorbike) are likely to occur in the same image together. Besides, there exist a large number of other advanced scene interpretation approaches [17–22].

However, the above scene interpretation approaches rely heavily on the availability of sufficient training data. In fact, the scene interpretation systems of computers, despite their immense computational capability, are still far from that of human beings. In some sense, the ease, with which a human recognises the windows, doors and balconies of a building, is inspiring for computer vision. It is therefore natural that we should try to design and build a scene interpretation system that is capable of recognising and interpreting its surrounding environment, with some logic rules gained from the cognition system of human vision. The research on inserting human logics in scene interpretation has been around for a couple of decades [23–26]. For example, Han and Zhu [26] proposed to construct a simple attribute grammar, in which six rules of layout of the rectangular surfaces are discovered, for interpreting man-made scenes. However, it is limited on the objects with their shapes being rectangle.

Although applying human logics is promising, it is intractable to develop a fully automatic scene interpretation system, since the study on how the human brain works is still in progress [27]. Fortunately, Petrou found out [28] that there exists an important cue from human language for applying human logics in scene interpretation, i.e. Tower of Knowledge (ToK)[1] [29]. For several years, Petrou and Xu have been dedicated to the research on ToK [29–34] for the specific task of interpreting building scenes. In essence, ToK is a multi-layer architecture, encapsulating the causal dependencies among the labels and functionalities of objects in a scene and their corresponding descriptors.[2] Moreover, the development of 3D reconstruction also facilitates the implementation of ToK, as more detailed descriptors of the objects, such as the object depth, can be obtained from their 3D models. Therefore, ToK was successfully applied in 3D scene interpretation system of buildings.

In this paper, we comprehensively survey and outline the various contributions related to ToK. The rest of this paper is arranged as follows. In Section 2, we briefly introduce the architecture of ToK. For utilising ToK in the real scene interpretation system, some key approaches on exploiting the vertical and horizontal connections of units in ToK architecture are reviewed in Sections 3 and 4, respectively. Section 5 summarises the work beyond ToK. Finally, we conclude this paper and look into the future research directions of ToK in Section 6.

2. The architecture of Tower of Knowledge

The ToK architecture is designed so that it models the acquired knowledge in the form of logic rules containing the answers to the questions **"what?"**, **"why?"** and **"how?"** [29]. For example, in choosing the label "balcony" for an object, the following fragment of conversation between the various modules of information flow is envisaged:

- *"What is this?"*
- *"It is a balcony."*
- *"Why?"*
- *"Because it is attached to a building and people can stand in it."*
- *"How?"*
- *"By offering enough space for a person to stand in and by being attached on a wall with an opening area to allow people to enter it from the building."*
- *"Is it really like that? Let me check."*

According to the above sequence, the architecture of ToK [29,30] for scene interpretation, in general, is shown in Fig. 1. To be more specific, as seen from the left hand side of Fig. 1, the ToK architecture is designed so that it encapsulates the acquired knowledge in the form of networks containing the answers to the questions "what?", "why?" and "how?". As the answers to these queries, the units belong to four levels: image, semantic, functionality and descriptor levels. The input to scene interpretation is the units at the image level, related to the image pixels or other features at low-level vision. The output of scene interpretation is the units at the semantic level, for answering "what is the object?" (e.g. "window"). The remaining two levels are those of functionalities and descriptors, as the *latent* units. The units of the functionality level, corresponding to answering question "why", are functionalities of the objects, such as "to look out" for confirming a "window". A functionality may be fulfilled, if the object has certain descriptors at the descriptor level, in order to answer question "how is it?". Indeed, for making sure if an object has certain functionality "to look out", the descriptor "glass-like" may be proper descriptor. The units at the descriptor level can interrogate the units of the image level and even sensors, to verify whether a required descriptor applies to an object.

In ToK architecture, the relationships of units in hierarchy encode the causal dependencies among semantic, functionalities and descriptors for an object, whereas the relationships of units at the same level indicate peer to peer contextual influences between each unit. Beyond the vertical interaction, the horizontal connections of units in ToK are also important for scene interpretation. For example, at the semantic level, the presence of a "balcony" strongly implies that the label of its neighbouring object might be "door". At the image level, the dependencies of pixels are modelled via MRF, and the pixels are then grouped together by graph cut approach [35], for image segmentation.

Then, the processing scheme of ToK architecture works in light of the information flow in Fig. 1, similar to the conversation fragment mentioned above. The label of an object can be determined by closing the feedback loop of vertical information flow in ToK architecture. Such an information flow starts from discovering the functions at functionality level, for each class of labels at semantic level, and then it works out seeking the right characteristics of the object at descriptor level, to fulfil the functions of the label for this object. Finally, the information flow is terminated in the feedback loop to confirm the label of the object, after finding out the characteristics of the object by interrogating the image (at image level). In addition, the horizontal information flow can be directly used to label an object according to its neighbouring objects, through exploiting the asymmetric dependencies of units at the same level.

In the following, we introduce some scene interpretation approaches working out the vertical connections and horizontal connections in ToK architecture, in Sections 3 and 4, respectively.

[1] Tower of Knowledge is one of most important works of Maria Petrou in the late year of her life, as it is the main part of her keynote talks during 2007 to 2011.

[2] It is inspired by human language in answering questions "what?" (object labels), "why?" (object functionalities) and "how?" (object descriptors).