Cognition 171 (2018) 65–84

Contents lists available at ScienceDirect

Cognition

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

Original Articles

Full interpretation of minimal images

Guy Ben-Yosef[a,b,c], Liav Assif[a], Shimon Ullman[a,c,*]

[a] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 7610001, Israel
[b] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[c] Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Keywords:
Image interpretation
Minimal images
Parts and relations
Top-down processing

ABSTRACT

The goal in this work is to model the process of 'full interpretation' of object images, which is the ability to identify and localize all semantic features and parts that are recognized by human observers. The task is approached by dividing the interpretation of the complete object to the interpretation of multiple reduced but interpretable local regions. In such reduced regions, interpretation is simpler, since the number of semantic components is small, and the variability of possible configurations is low.

We model the interpretation process by identifying primitive components and relations that play a useful role in local interpretation by humans. To identify useful components and relations used in the interpretation process, we consider the interpretation of 'minimal configurations': these are reduced local regions, which are minimal in the sense that further reduction renders them unrecognizable and uninterpretable. We show that such minimal interpretable images have useful properties, which we use to identify informative features and relations used for full interpretation. We describe our interpretation model, and show results of detailed interpretations of minimal configurations, produced automatically by the model. Finally, we discuss possible extensions and implications of full interpretation to difficult visual tasks, such as recognizing social interactions, which are beyond the scope of current models of visual recognition.

1. Introduction

Humans can recognize in images not only objects (e.g., a person) and their major parts (e.g., head, torso, limbs), but also multiple semantic components and structures at a fine level of detail (e.g., shirt, collar, zipper, pocket, cuffs etc.), as in Fig. 1A. Identifying detailed components of the objects in the image is an essential part of the visual process, contributing to the understanding of the surrounding scene and its potential meaning to the viewer (Section 6.1). Although this capacity is of fundamental importance in human perception and cognition, current understanding of the processes involved in detailed image interpretation is limited.

From the modeling perceptive, existing models cannot deal well with the full problem of detailed image interpretation, and, as discussed below, the limitations are of fundamental nature. Computational models of object recognition and categorization have made significant advances in recent years, demonstrating consistently improving results in recognizing thousands of natural object categories in complex natural scenes (Section 2). However, existing models cannot provide a detailed interpretation of a scene's components in a way that will approximate human perception. For example, for a given image such as Fig. 1A, existing models can correctly decide if the image contains a person (e.g., Csurka, Dance, Fan, Willamowski, & Bray, 2004; Simonyan & Zisserman, 2015), and can locate a bounding box around the body (e.g., Dalal & Triggs, 2005; Girshick, Donahue, Darrell, & Malik, 2014). At a more refined level, current algorithms can provide an approximate segmentation of the body figure (e.g., Long, Shelhamer, & Darrell, 2015), and can locate image region containing the main body parts, such as the torso region, the face, or the legs (e.g., Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2017; Vedaldi et al., 2014), or keypoints at the joints (e.g., Chen & Yuille, 2014; Wei, Ramakrishna, Kanade, & Sheikh, 2016). However, existing computational models cannot achieve the accuracy and richness of the local interpretation of image components perceived by a human observer (e.g., as in Fig. 1B).

To clarify the terminology, by the term 'visual interpretation' we refer to a mapping between entities in the images and entities in the world (such as objects, object categories, object parts at different levels, and other physical entities). For instance, within a face image, a particular image contour may correspond to, say, the mouth's upper lip. The contour is an image component, the upper-lip is a semantic component in the outside world, and the interpretation process maps

* Corresponding author at: Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 7610001, Israel.
E-mail address: shimon.ullman@weizmann.ac.il (S. Ullman).

http://dx.doi.org/10.1016/j.cognition.2017.10.006
Received 20 January 2017; Received in revised form 4 October 2017; Accepted 6 October 2017

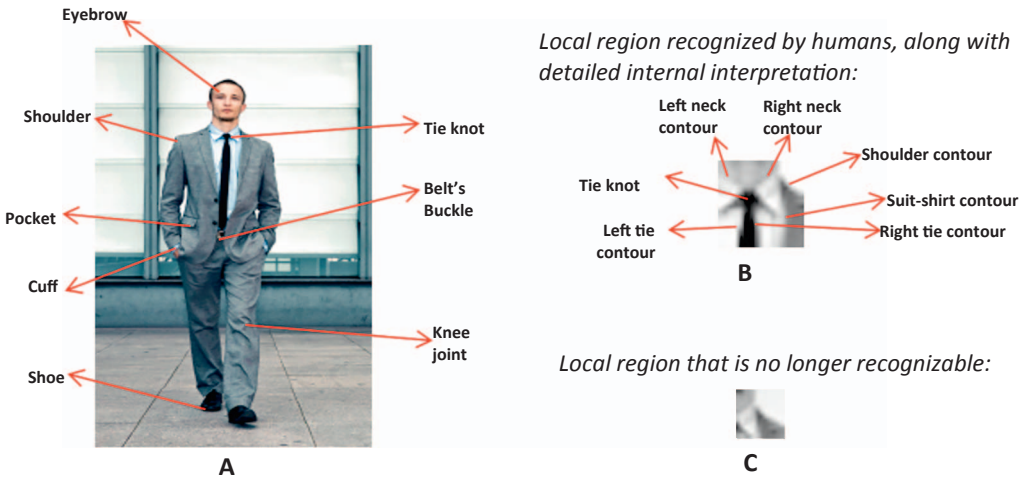

Fig. 1. (A). Humans can identify a large number of semantic features and parts in an object image. In the image of a walking person, features like the suit's pocket, tie's knot, left shoe, or the right ear, are easily identified by humans, among many others. **(B).** A detailed interpretation of a small image region, as identified by human observers. In small local regions, the number of semantic components is significantly smaller than in full images, and variability is reduced. **(C).** When the local region becomes too limited, human observers can no longer recognize and interpret its content when presented on its own (Ullman et al., 2016).

between the two.

1.1. Local image interpretation

Producing a detailed interpretation of an object's image is a challenging task, since a full object may contain a large number of identifiable components in highly variable configurations. We approach this task by decomposing the full object or scene image into smaller, local, regions containing recognizable object components. There are several advantages to perform the interpretation first in local regions, and then combine the results. First, as exemplified in Fig. 1B, in such local regions the task of full interpretation is still possible (Torralba, 2009; Ullman, Assif, Fetaya, & Harari, 2016), but it becomes more tractable, since the number of semantic recognizable components is highly reduced. As will be shown (Section 5), reducing the number of components plays a key factor in effective interpretation. At the same time, when the interpretation region becomes too limited, observers can no longer interpret or even identify its content, as illustrated in Fig. 1C (Ullman et al., 2016). The goal of the model is therefore to apply the interpretation process to local regions that are small, yet interpretable on their own by human observers. A second advantage of applying the interpretation locally is that variability of configurations taken from the same object class, but limited to local regions, is often significantly lower compared with complete object images. For example, the full horse images in Fig. 2 (taken from the 'horse' category in ImageNet, Deng et al., 2012, a common benchmark for evaluating object recognition models) are quite different from each other, but can become significantly more similar at the level of local regions. This well-known advantage of local regions, which has been used in part-based recognition models, is extended below to define minimal recognition configurations. Finally, as will be discussed in the next section, the image of a single object typically contains multiple, partially overlapping regions, where each one can be interpreted on its own. Due to this redundancy, performing the interpretation locally and then combining the results increases the robustness of the full process to local occlusions and distortions.

1.2. Minimal configurations

In performing local interpretation, how should an object image be divided into local regions? The approach we take in this study is to develop and test the interpretation model on regions that can be interpreted on their own by human observers, but at the same time are as limited as possible. We used for this purpose a set of local recognizable images derived by a recent study of minimal recognizable images (Ullman et al., 2016). We briefly describe below how these images were obtained, and then explain the reasons for using these local images in developing and testing the interpretation model.

A 'minimal configuration' (also termed Minimal Recognizable Configuration, or MIRC) is defined as an image patch that can be reliably recognized by human observers, which is minimal in the sense that further reduction by either size or resolution makes the patch unrecognizable. To discover minimal configurations, an image patch was presented to observers: if it was recognizable, 5 descendants were generated: four by small (20%) cropping at one of the corners, and one by reducing resolution (by 20%) of the original patch. A recognizable

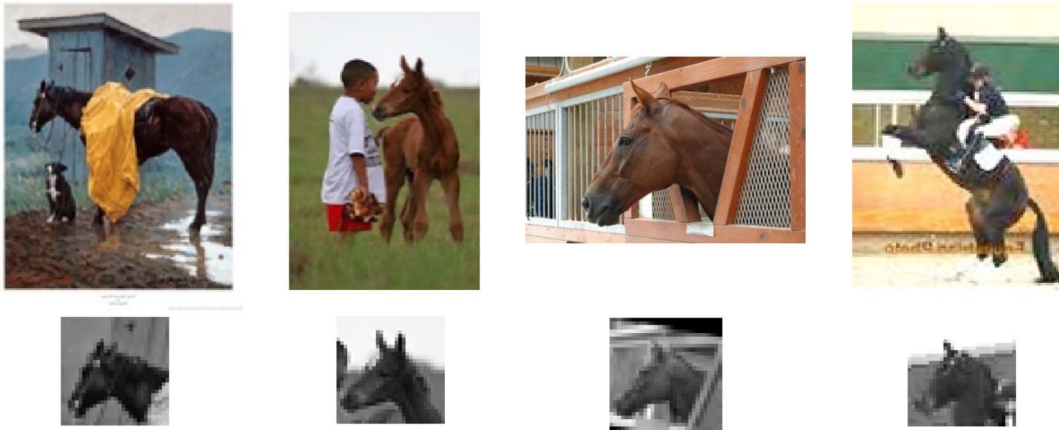

Fig. 2. Complete horse images taken from ImageNet object recognition benchmark (Deng et al., 2012), and a small recognizable region that is interpretable (similar to Fig. 4A), next to each complete horse image illustrating the reduced variability in small recognizable region vs. the complete object image.