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
Available online 04 November 2017

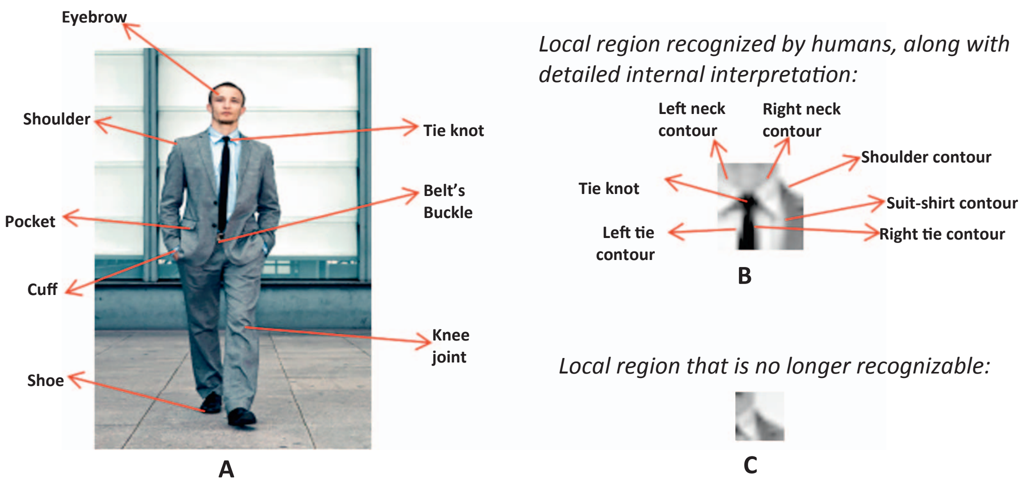

Fig. 1. (A). Humans can identify a large number of semantic features and parts in an object image. In the image of a walking person, features like the suit's pocket, tie's knot, left shoe, or the right ear, are easily identified by humans, among many others. **(B).** A detailed interpretation of a small 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

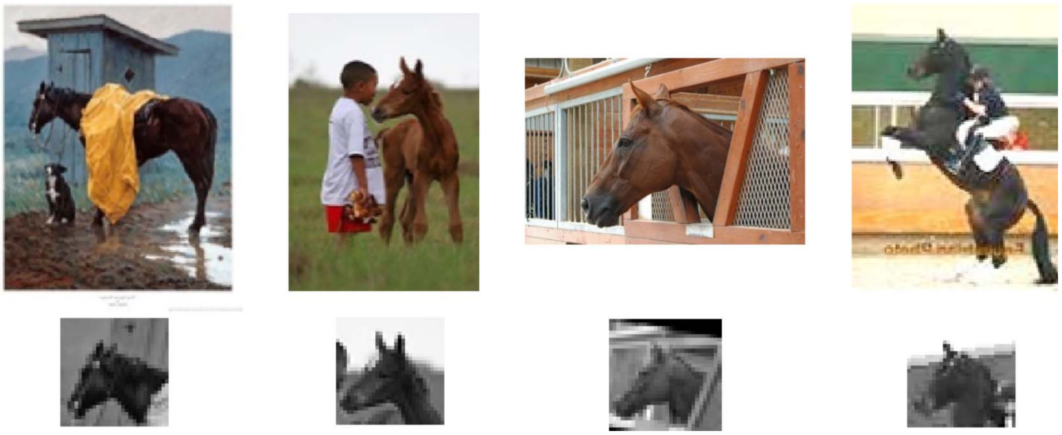

Fig. 2. Complete horse images taken from ImageNet object recognition benchmark (Deng et al., 2012), and a small recognizable region that is interpretable (similar to Fig. 4A), next to each complete horse image illustrating the reduced variability in small recognizable region vs. the complete object image.

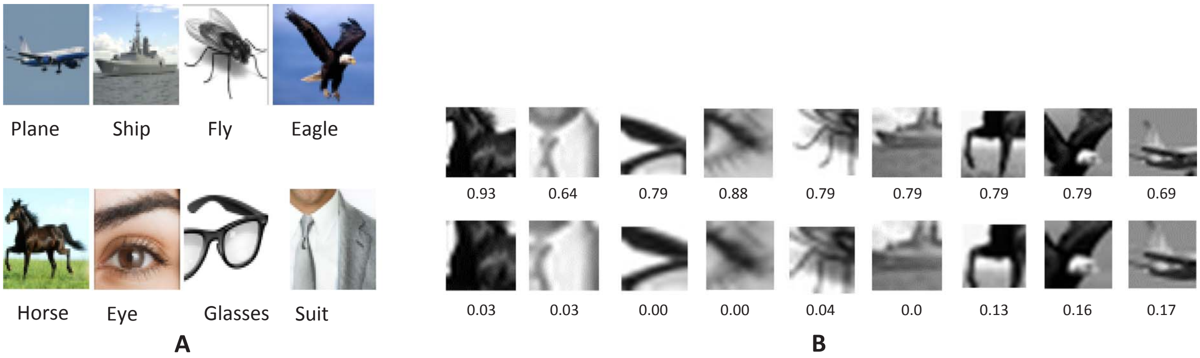

Fig. 3. Minimal configurations adapted from Ullman et al. (2016). **(A).** The search for minimal images started from different object images (8 shown here), each composed of 50 × 50 image samples. **(B).Top row:** minimal images discovered by the search. **Bottom row:** sub-minimal configurations, which are slightly reduced versions of the images on top. Numbers below each image show correct recognition rate by 30 human observers. Small changes to the local image at the minimal configuration level can have large effect on recognition. A data set of such pairs is used below for modeling the interpretation of local regions.

patch is identified as a 'minimal configuration' if none of its 5 descendants reach recognition criterion (set to 50%, results are insensitive to this setting). A search started with images from different object classes (Fig. 3A), and identified their minimal configurations over all possible positions, sizes and resolutions. Each subject saw a single patch only from each original image, requiring over 15,000 subjects. Testing was therefore done online using Amazon's Mechanical Turk platform (MTurk), combined with laboratory controls. At the end of the search, each object class was covered by multiple minimal configurations at different positions and sizes. Minimal configurations were on average about 15 image samples in size; some contained local object parts, others were more global views at a reduced resolution. Examples of identified minimal configurations are shown in the top row of Fig. 3B.

A notable aspect of the results for the purpose of the current study is the presence of a sharp transition for almost all minimal configurations from a recognizable to a non-recognizable minimal image: a surprisingly small change at the minimal-configuration level can make it unrecognizable. Examples are shown in Fig. 3B, bottom row, together with their respective recognition rates. The loss of recognition when the image is sufficiently reduced and features are removed is expected, but the sharp drop at the minimal level is remarkable, and consistent across many examples. It was used below to identify informative properties and relations for the interpretation process. It was also found that the large gap in human recognition rate between minimal and sub-minimal images is not reproduced by current computational models of human object recognition (Serre, Oliva, & Poggio, 2007) and recent deep network models (Krizhevsky, Sutskever, & Hinton, 2012; Simonyan & Zisserman, 2015). As shown below (Section 5.2), the full interpretation model can provide at least a partial explanation to this sharp drop in recognition.

1.3. Recognition and interpretation

With respect to local interpretation, recognition tests of minimal images showed that although the minimal images are 'atomic' in the sense that their partial images become unrecognizable, humans can consistently recognize multiple semantic features and parts within them. It was noted (Ullman et al., 2016) that recognition and interpretation of minimal images go hand in hand in the sense that under the tested conditions (unlimited viewing time), when subjects correctly recognized a minimal image, they were also able to provide an internal interpretation of multiple internal components. Since in minimal images all the available information is, by definition, crucial for recognition, we propose in the model below that all the interpreted components of minimal images also contribute to their recognition. As described further below (Sections 4.3,5.2,6.4), in the model, the full interpretation process contributes to accurate recognition, since a potential false detection can be rejected if it does not have the expected internal interpretation.

For the purpose of modeling human visual interpretation, our initial focus is on the interpretation of minimal images, for the following reasons. First, they provide a useful test set for the model: since they are interpretable by humans, a theory of human image interpretation should be applicable to such configurations. Second, we use minimal and sub-minimal pairs with a large gap in recognizability and interpretability as a source for inferring useful features for the interpretation of minimal images (Section 4). Before describing the model, we briefly describe past work related to visual object interpretation.

2. Related work on visual object interpretation

Visual recognition can take place at different levels of details, from full objects and their main parts, to fine details of objects' structure. In modeling human visual perception, as well as in computer vision, much of the work to date has focused on relatively coarse levels, rather than full object interpretation considered here. For example, a leading biological model of the human object recognition system, the HMAX model (Riesenhuber & Poggio, 1999; Serre et al., 2007) produces as its output general category labels of full objects, rather than a detailed interpretation. Other biologically inspired models of recognition use features based on unit responses along the ventral cortical hierarchy (e.g., Murphy & Finkel, 2007; Rodríguez-Sánchez & Tsotsos, 2011), but their focus is again on shape representation and object recognition rather than full interpretation. Some models of human vision, such as the Recognition by Components (RBC) model of human object categorization (Biederman, 1987), deal with both objects and parts, but the parts are limited to a small number of 3-D major components, and do not provide a detailed object interpretation.

A model for human image interpretation (Epshtein, Lifshitz, & Ullman, 2008) was shown to provide partial image interpretation by a combination of bottom-up with top-down processing. The model uses a hierarchy of informative image patches to represent object parts at multiple levels. The current model also uses a combination of bottom-up and top-down processing, but it provides a significantly richer interpretation, and based on computational and psychophysical considerations, it uses an extended set of elements and relations. A preliminary version of the model was described in Ben-Yosef, Assif, Harari, and Ullman (2015). The current model extends the early version in the use of minimal images (rather than local image regions), in testing on multiple classes, and in comparisons with human vision.

In computer vision, there has been a rapid progress in different aspects of object and scene recognition, based primarily on deep convolutional neural networks and related methods (He et al., 2016;

Hinton, 2007; Krizhevsky et al., 2012; LeCun, Bengio, & Hinton, 2015; Simonyan & Zisserman, 2015; Yamins et al., 2014). Such methods have also been adapted successfully for image segmentation, namely, the delineation of image regions belonging to different objects. For example, recent algorithms (e.g., Chen et al., 2017; Long et al., 2015) can identify image regions belonging to different objects in the PASCAL (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010) or CoCo (Lin et al., 2014) benchmarks; however, they do not locate the precise object boundaries, and do not identify the object's semantic components.

A number of studies have begun to address the problem of a fuller object interpretation, including methods for part-based detectors, object parsing, and methods for so-called fine-grained recognition. Recent examples include modeling objects by their main parts, for example an airplane's nose, tail, or wing (Vedaldi et al., 2014), or modeling human-body parts such as the head, shoulder, elbow, or wrist (e.g., Felzenszwalb, Girshick, McAllester, & Ramanan, 2010; Girshick, Iandola, Darrell, & Malik, 2015). Related models provide segmentation at the level of object parts rather than complete objects (applied e.g. to animal body parts such as head, leg, torso, or tail, e.g., Azizpour & Laptev, 2012; Chen et al., 2017). Another form of interpretation has been the detection of key-points within an object, such as key-points of the human body (e.g., Andriluka, Pishchulin, Gehler, & Schiele, 2014; Chen & Yuille, 2014; Tompson, Jain, LeCun, & Bregler, 2014) and within the human face (e.g., Xiao et al., 2016; Yang, Luo, Loy, & Tang, 2015).

2.1. Structured models and top-down processing

The goal of interpretation models, such as those above, is to identify the semantic structure in an image region. The model is usually given during learning a set of training images together with their interpretation, i.e., a set of semantic elements within each image, and the goal of the model is to identify similar elements in a novel image. In a correct interpretation, the internal components are expected to be arranged in certain consistent configurations, which are often characterized in the model by a set of spatial relations between components. The task of producing the semantic interpretation can therefore be naturally approached in terms of locating within an image region a set of elements (primitives), arranged in a configuration that satisfies relevant relations. The term 'relations' also includes here properties of single elements (e.g., the curvature, location, or size of a contour), which can be considered as unary relations.

The model described in this work belongs to this general approach of structured models. There is a rich history to the use of structural models in the computational study of vision, including visual recognition and interpretation. Models differ in the shape components used to create structured configurations, the relations used to represent configurations, and the algorithms used to learn structures from image examples, and to identify similar structure in novel images.

Basic shape components used in past structural models include edge and boundary elements, including contours (e.g., Brooks, 1983), contour-pairs (e.g., Brooks, 1983; Ferrari, Jurie, & Schmid, 2010) and boundary fragments (e.g., Opelt, Pinz, & Zisserman, 2006); image patches, regions and their descriptors (e.g., Fei-Fei, Fergus, & Perona, 2006; Felzenszwalb & Huttenlocher, 2005; Hanson & Riseman, 1978; Todorovic & Ahuja, 2006; Zhu & Mumford, 2007) complete hierarchies of increasingly complex contour or region combinations and their descriptors (e.g., Fidler & Leonardis, 2007; Ommer & Buhmann, 2007; Siddiqi, Shokoufandeh, Dickinson, & Zucker, 1999; Zhu, Chen, & Yuille, 2009), obtained by grouping and segmentation processes; as well as 3-D surfaces and volumes (e.g., Hanson & Riseman, 1978; Marr & Nishihara, 1978).

The relations used in these models were mostly simple, in particular, the expected location within a reference frame and relative displacement (e.g., Chen & Yuille, 2014; Chen et al., 2017; Fei-Fei et al., 2006; Felzenszwalb & Huttenlocher, 2005; Felzenszwalb et al., 2010; Ferrari et al., 2010; Fidler & Leonardis, 2007; Ommer & Buhmann, 2007), but a few used more complex relations such as co-termination (Ferrari et al., 2010), parallelism of elements (Zhu & Mumford, 2007), and containment (Todorovic & Ahuja, 2006).

In terms of algorithms used to learn, and then identify, image structures, closest to our model are methods developed and used in the field of machine vision under the general term 'structured prediction' such as Structured Support Vector Machine (Joachims, Hofmann, Yue, & Yu, 2009), and Conditional Random Field (Lafferty, McCallum, & Pereira, 2001), combined with deep network algorithms (e.g., Chen & Yuille, 2014). These models are given the set of possible relations to use, and they learn the specific parameters from examples. Other methods used in related past models include probabilistic graphical models (e.g. Epshtein et al., 2008; Fei-Fei et al., 2006; Jin & Geman, 2006), and stochastic grammars (e.g., Zhu & Mumford, 2007; Zhu et al., 2009). A basic distinction between methods is that some rely on purely bottom-up processing (e.g. Felzenszwalb & Huttenlocher, 2005; Krizhevsky et al., 2012; Riesenhuber & Poggio, 1999), while others combine bottom-up with top-down processes (e.g. Epshtein et al., 2008; Fei-Fei et al., 2006; Zhu & Mumford, 2007).

2.2. Focus of the current model compared with past models

None of the past models mentioned above implemented semantic interpretation at the level considered in this paper, but several, in particular models using hierarchical representations and grammar models, incorporate descriptions at multiple levels and could possibly be extended to include semantic parts at all levels (e.g. Marr & Nishihara, 1978; Ommer & Buhmann, 2007; Zhu & Mumford, 2007; Zhu et al., 2009).

In formulating a structural model for image interpretation, the main aspects to consider are the components used to describe the structure, relations between them, and methods for learning the underlying image structure and finding similar structures in novel images. Our main focus in the current work is on the identification of informative components and relations, using the set of minimal images combined with sub-minimal images and hard-negative examples. In most past visual models that deal with image structures, informative relations have been limited to a small set of simple relations, in particular relative displacement between components. As elaborated below, results of the present modeling show that the capacity to provide full interpretation requires the use of features and relations, which go beyond those used in most past models.

In comparing the present model with alternative approaches, of particular interest are comparisons with recent bottom-up, network-based models (Sections 5.2,6.4). With the recent success of such models in various visual tasks, comparisons are useful to explore the possible limitations of purely bottom-up methods, and the potential contribution of structural models, which often naturally employ, as the current model does, a combination of bottom-up and top-down processes.

Network models are a useful form for formulating models of visual processes, but our results suggest that exclusively feed-forward network models are unlikely to be sufficient for detailed image interpretation. Recent deep nets modeling include extensions beyond feed-forward processing, such as recurrent and LSTM nets (e.g., Denil et al., 2012; Mnih, Heess, & Graves, 2014, for recognition, or Xiao et al., 2016, for facial landmark detection), but not for detailed local interpretation. It will be of interest to study in the future network models, which can efficiently incorporate the computations used by the current model for that task of detailed image interpretation (Section 6.5).

3. Model description

Our interpretation scheme has two main components: in the learning stage, it learns the semantic structure of an image region in a supervised manner, and in the interpretation stage, it identifies the learned structure in similar image regions. These two stages are described in the rest of this section (combined with the appendices, which supply more technical details).

3.1. Learning setup

The learning stage derives the semantic structure of an object region based on positive examples coming from class images, and negative examples derived by the system from similar but non-class images. We first describe how these training examples are obtained, and then how the region's semantic structure is learned from them.

Positive examples are supplied manually during a preparation stage as a set of image regions with their interpretation, namely, the semantic elements that should be identified and localized. Since the goal is to model humans' ability to obtain a detailed local interpretation, the target set of semantic primitives to identify was collected for different minimal images using human observers. The semantic features to be identified by the model, e.g. 'ear', 'eye', 'tie knot' etc., were features that human observers label consistently in minimal images, verified using a Mechanical Turk procedure (see examples in Fig. 4, top row, and Appendix A for procedure details). The average number of consistently identified elements within a single minimal configuration was 8. To capture the recognized internal components fully as perceived by humans, the primitive elements in the model were divided into three types: two-dimensional (2-D) regions, 1-D contours, and points (0-D). Example sets of primitives for modeling the interpretation of minimal images are shown in Fig. 4, bottom row. For instance, a point-type primitive may describe the eye in the horse head model (Fig. 4A), and a contour-type primitive describes borders such as the borders of the tie in the man-in-suit (Fig. 4B). Larger semantic features marked by observers such as the ship's 'bow' region or the tie's 'knot', were marked as region primitives (outlined squares in Fig. 4, bottom row). The three types of primitives are also supported by psychophysical and physiological studies (e.g., Attneave, 1954; Pasupathy & Connor, 1999).

Given the semantic elements identified by humans in a minimal image of class C (e.g., a horse-head), we prepared a set of annotated images, in which the semantic components (denoted P_C below) were marked manually (with automatic refinement). Examples for such annotations are shown in Fig. 5A. The unsupervised learning of components and relations is considered briefly in the final discussion (Section 6.2).

Having a set of interpretation examples, the learning process next searches automatically for negative interpretation examples – these are non-class images that are potentially confusable with class images. The procedure for identifying so-called 'hard negatives' (e.g., Azizpour & Laptev, 2012; Felzenszwalb et al., 2010) (detailed in Section 4.3), starts from a large set of random non-class examples, and then iterates over two steps: applying interpretation, and finding non-class examples with high interpretation score (which is produced by the interpretation algorithm), then adding them to the training set and re-training the interpretation model.

3.2. Learning the semantic structure

For a minimal configuration C, we define its semantic structure S_C as a pair of two sets: the set of semantic components P_C mentioned in Section 3.1 (also called below the 'primitives'), and a set of relations between primitives, denoted by R_C, namely

$$S_C = < P_C, R_C >$$

We include properties of a single primitive as a relation with a single argument. A basic problem at this stage is therefore to learn a set of relations that are useful for identifying configurations, namely, which appear in the positive class examples, and distinguish them from configurations found in the similar but non-class negative examples. The relevant relations for a given image are selected automatically during learning from an initial candidate set of potentially informative and useful relations to compute (see Section 4 on how this set was obtained). For instance, whether the relation 'containment' between pairs of primitives should be included in R_C, all potential pairs of primitives are examined, using the positive and negative examples, to test if one primitive is consistently contained within the other. (See Appendix B.1 for how the contribution of a relation to the final interpretation was measured.) Each of the relations used in the interpretation scheme is given an index, e.g. the relation 'containment' may have the index '4'. Following selection, the set of all informative relations identified in a given minimal image C are represented by the vector R_C. Each element in R_C specifies a relation, and its relevant components. For example, the 3rd component of R_C (i.e., $R_C^{(3)}$) could be the triplet (4, 5, 7). This triplet means that relation 4, which is 'containment', holds between components 5 and 7 in the local image model, specifying that component 5 in the local model should be contained inside component 7. Similarly, the element in position 4 in R_C (i.e., $R_C^{(4)}$) can be a 'straightness' (unary) relation of primitive index 2, etc. Relations in our model could be either binary, e.g., 'containment', or represented by a scalar, e.g., the property 'location', specifying the location of a component within the local image.

A detailed description of the learning model and procedure based on 'structured learning' framework (e.g. Shalev-Shwartz & Ben-David, 2014) is given in Appendix B. For a novel image, the vector representation R_C of the image structure is derived as described in the

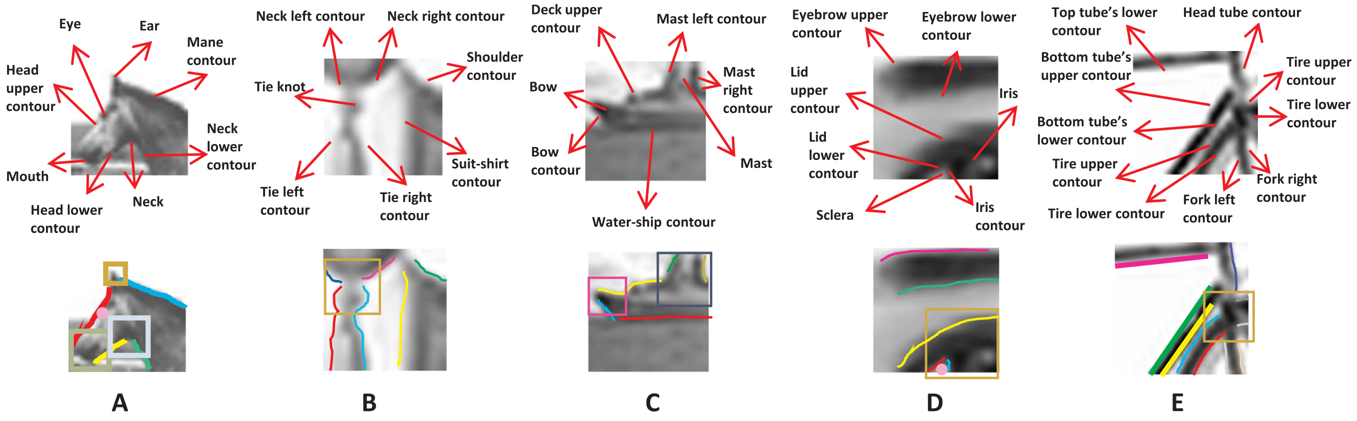

Fig. 4. Human interpretation of minimal configurations. **(Top row).** All components that were identified consistently by human observers (Appendix A). **(Bottom row).** In the interpretation model the components are represented by three types of primitives: points, contours, regions, together with relations between them. For each column, the identified components on the top panel are plotted in different colors on the bottom panel, and by either a point, a contour, or a region (an outlined square).

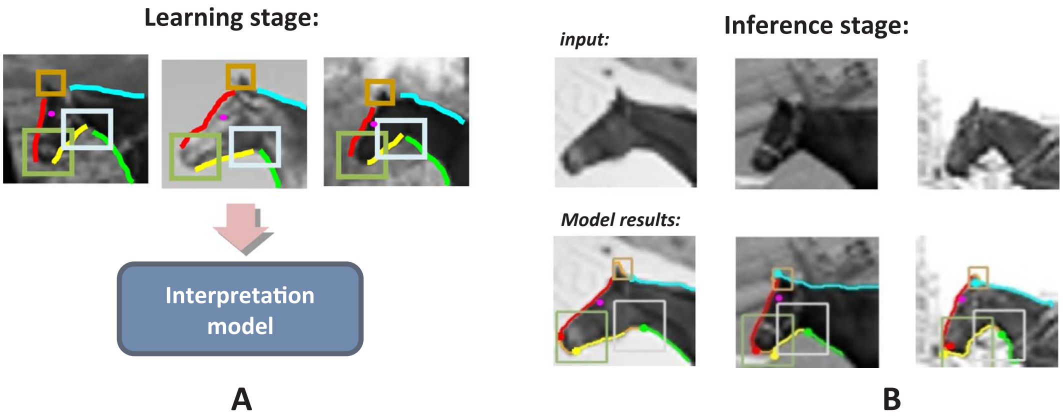

Fig. 5. Stages in the interpretation scheme, with horse-head as an example. **(A).** Point, contour, and region primitives that represent the identified parts (cf. Fig. 4) are annotated in training examples (several shown here), and are used to learn an interpretation model, which combines the primitives with relations between them. **(B).** Results of the interpretation model for 3 novel examples of the horse head minimal configuration.

next section, and then used for final interpretation decision.

3.3. Interpretation of a novel image

In this section we assume that a local image region has been identified as a likely candidate of a particular object or object part, and the current task is to produce an internal interpretation of the candidate region, and make a final decision about its identity. More details of the algorithm are given in Appendix B, and we also describe later (Section 6.4) how the initial detection and full interpretation are integrated together in a combined scheme of a bottom-up stage identifying likely candidates (e.g. by a DNN classifier trained for the task), followed by a top-down interpretation and validation stage.

The interpretation process starts with a candidate region and its proposed classification (e.g., that it contains a horse-head). The process then uses the learned model of the region's structure to identify within the region a structure that best approximates the learned one. This process proceeds in two main stages. The first is a search for local primitives, namely points, contours, and regions in the image, to serve as potential candidates for different components of the expected structure. The second stage searches for a configuration of the components that best matches the learned structure.

The first stage identifies in the image candidate primitives for the model components, as described in Appendix B.1. This process includes local edge detection and grouping edges into local contours. Properties and relation between components are then computed as described in Appendix C. The second stage searches for the best configuration of candidates, by a structured prediction algorithm described in Appendix B.1 and B.2. To match a given image configuration to the learned structure, we compute the relations in R_C for this configuration, and then use a compatibility scoring function based on a random forest classifier (Breiman, 2001, Appendix B), which produces a number that evaluates the degree to which the configuration is a correct interpretation of the input image. The interpretation scheme finally selects the highest-scoring configuration, and returns both this configuration (termed 'interpretation') and its score (termed 'interpretation score'). A search among multiple configurations is feasible due to the small number of primitives in the local region. This overall process is illustrated in Fig. 6. A detailed description of the scoring procedure and the optimization part (i.e., finding the most compatible configuration) is given in Appendix B.

4. Useful relations for interpretation

Producing an interpretation of an image region requires the localization of its participating components, and verifying their correct configuration. The model verifies the structure using inter-elements relations, and a natural question is therefore which relations are useful in modeling local semantic structures. The visual system is known to be sensitive to a range of spatial properties and relations between components such as curvature, straightness, proximity, relative displacement, collinearity, inclusion, bisection, and others, which have been studied both perceptually and physiologically (see review in Section 4.1 below). It is unknown, however, which relations play a significant role in the task of visual interpretation. In this section we describe the methods we used to identify informative relations for interpretation, which were then included in the set of interpretation relations used by the model.

In contrast with the richness of relations that can be efficiently perceived by the visual system (Section 4.1), the majority of models for image recognition and interpretation have been based on a limited number of basic relations. Recognition models based on deep networks obtain high performance in basic categorization, but when the task requires a more detailed interpretation, e.g. identifying keypoints in human pose estimation, performance often improves by explicitly incorporating inter-element relations, in particular relative displacement and orientations, using e.g. CRF models (Chen & Yuille, 2014; Chen et al., 2017; Wei et al., 2016). We next examined the set of relations which are informative for the full interpretation of local images.

The availability of minimal images allowed us to examine whether basic relations used in previous schemes are sufficient for producing an accurate interpretation by the interpretation model. Minimal configurations are by construction non-redundant visual patterns, and therefore their recognition and interpretation depend on the effective use of all the available visual information. It consequently becomes of interest to examine the performance of a model that uses a limited set of relations when applied to the interpretation of minimal images. To this end, we constructed a version of the interpretation scheme, where the set of relations was limited to displacement and proximity relations. Performance for this version proved insufficient compared with human interpretation (see more details in Section 5). This limitation motivated the search for additional informative relations, which were shown to improve the interpretation of minimal images. It is worth noting that since minimal images contain small sets of components, it becomes more feasible to use in the model inter-element relations that are more complex and more computationally demanding than used in past models.

We describe in Sections 4.1–4.4 below the process of identifying informative relations for the interpretation process. Previous psychophysical and physiological studies have proposed a number of relations that the visual system is sensitive to. These provided an initial set of candidate relations, and each relation was evaluated by measuring its

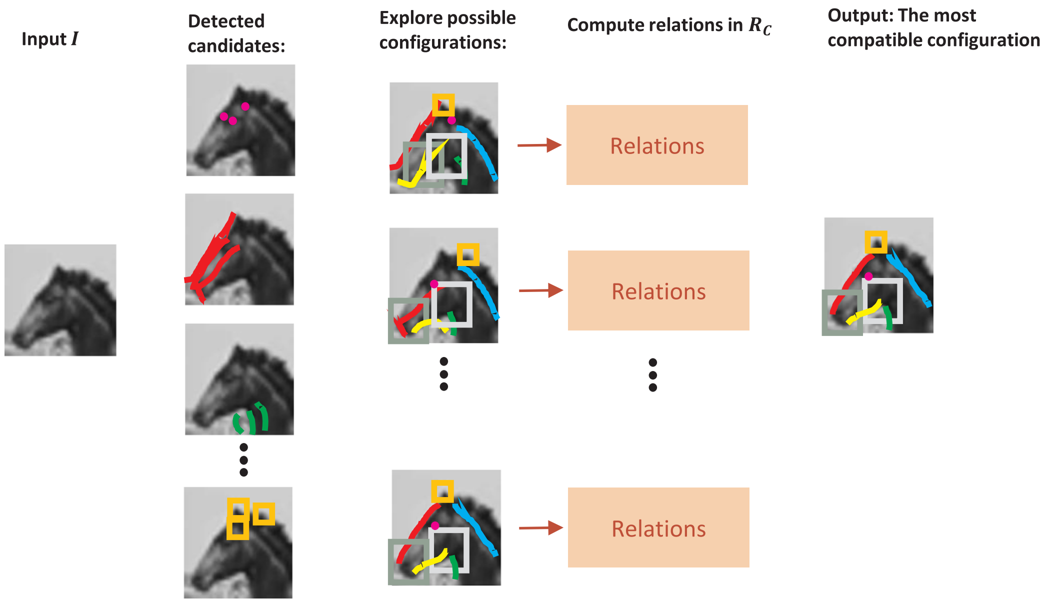
Input I
Detected candidates:
Explore possible configurations:
Compute relations in R_C
Output: The most compatible configuration
Relations
Relations
Relations

Fig. 6. An overview of the interpretation process of a novel image. From left to right, **Input image. Detected candidates:** of the primitive components, examples for 3 candidates of each primitive are shown. **Configurations:** examples of possible configurations of detected primitives (denoted by π in Appendix B); the one at the bottom is the optimal one. **Computing relations:** compute the relations in R_C for each candidate configuration (the vector $\phi_S(I,\pi)$ in Appendix B). **A compatibility score:** a scoring function ($g(\phi_S(I,\pi);w)$ in Appendix B) is computed for each configuration. The configuration with highest score is returned as final interpretation; the highest score is returned as the 'interpretation score'.

contribution to the interpretation model applied to a test set of minimal configurations, combined with sub-minimal configurations (Section 4.2) and hard-negative examples (Section 4.3). We finally describe the relations that were found to be informative for learning interpretation. We also consider (Section 6.2) how a more complete set of informative interpretation relations could be learned and refined over time.

4.1. Relevant visual relations in past literature

The study of relations between elements in the visual field dates back at least to the Gestalt school and its principles of perceptual organization (Wertheimer, 1923). These principles were based on relations that group visual elements together to be perceived as coherent units, and included proximity, similarity, connectivity, symmetry, and continuity between dots, contours, or regions. Psychophysical experiments since have shown that the human visual system is effortlessly sensitive to a range of spatial properties and relations between visual elements. Such relations include: parallelism and symmetry (e.g., Feldman, 2007; Machilsen, Pauwels, & Wagemans, 2009; Stahl & Wang, 2008), curvature and convexity (e.g., Foster, Simmons, & Cook, 1993; Murphy & Finkel, 2007; Rodríguez-Sánchez & Tsotsos, 2011), connectedness of blobs (e.g., Palmer & Rock, 1994), and connectedness of contours (e.g., Elder, Krupnik, & Johnston, 2003; Elder & Zucker, 1996; Jacobs, 1996; Jolicoeur, Ullman, & Mackay, 1986), continuity of contours (e.g., Kanizsa, 1979), co-linearity (e.g., Field, Hayes, & Hess, 1993) and co-circularity (e.g., Parent & Zucker, 1989) of contours, relative length of lines and contours (e.g., Saarela, Sayim, Westheimer, & Herzog, 2009), bisection (e.g., Westheimer, Crist, Gorski, & Gilbert, 2001), and inclusion (Ullman, 1984).

For many of these relations, it remains unclear whether they are being formed at early stages of visual perception in a bottom-up manner (e.g., Field et al., 1993; Kanizsa, 1979; Parent & Zucker, 1989) or at later stages, applied in a top-down manner to early visual representations (e.g., Jolicoeur et al., 1986; Roelfsema, Lamme, & Spekreijse, 1998; Ullman, 1984). It is also still unclear which of the relations perceived effortlessly by humans play also a direct role in recognition and interpretation. The computational test described below evaluated directly the contribution of different relations to the interpretation of minimal image. To search for informative relations for interpretation, we started with a list of visual relations identified in past studies listed above, called the 'candidate relations', and tested their contribution to the interpretation process applied to minimal and sub-minimal images and hard-negative examples, as discussed next (Sections 4.2–4.3).

4.2. Useful relations from minimal vs. sub-minimal images

The sharp drop in humans' ability to recognize and interpret a minimal configuration when the image is slightly reduced (Ullman et al., 2016), provided a tool for identifying useful relations for modeling human interpretation. A minimal image was compared with its similar, but unrecognized sub-image, to identify either a missing component (e.g., a contour), a missing region feature, or a relation (e.g., connected contours that become unconnected), which were present in the minimal image but not in the sub-minimal configuration. Examples are illustrated in Fig. 7, where pairs of minimal vs. sub-minimal configurations are shown (columns 1–2), along with the sets of internal semantic components that were identified by human observers in the minimal images (column 3). By using the human annotations, we found if any components in the minimal image were missing in its sub-minimal image. Using the set of candidate relations, we identified relations that are satisfied in the minimal but not the sub-minimal image. The missing component or relation may not be unique, and in such cases we evaluated a number of alternatives. The examples in Fig. 7 include the existence of the left-side tie contour (7A), connectedness of the two horse muzzle contours (7B), high-curvature meeting of contours (7C), and characteristic texture in the water region (7D). These features have been shown in the past to be informative for recognition (e.g., Foster et al., 1993; Murphy & Finkel, 2007; Pasupathy & Connor, 1999; Rodríguez-Sánchez & Tsotsos, 2011).

We next evaluated for each of the missing components or relations, how consistent it is among other examples of minimal images, and how informative it is for the interpretation process, using our full data set of training examples. We start by testing for consistency in the set of minimal and sub-minimal pairs of the same class namely, finding additional pairs separated by the same component or relation (Fig. 7, columns 5–6). As an initial filtering stage, components or relations playing a role in at least 3 additional pairs were kept for the next stage, in which they were tested by their contribution to the performance of the interpretation algorithm. Each relation (similarly for candidate components) was tested by adding it to the set of relations (namely, to the relations R_C), training a new interpretation algorithm, and measuring the difference in interpretation performance with and without this relation.

In more details, to test how informative is a given relation to the

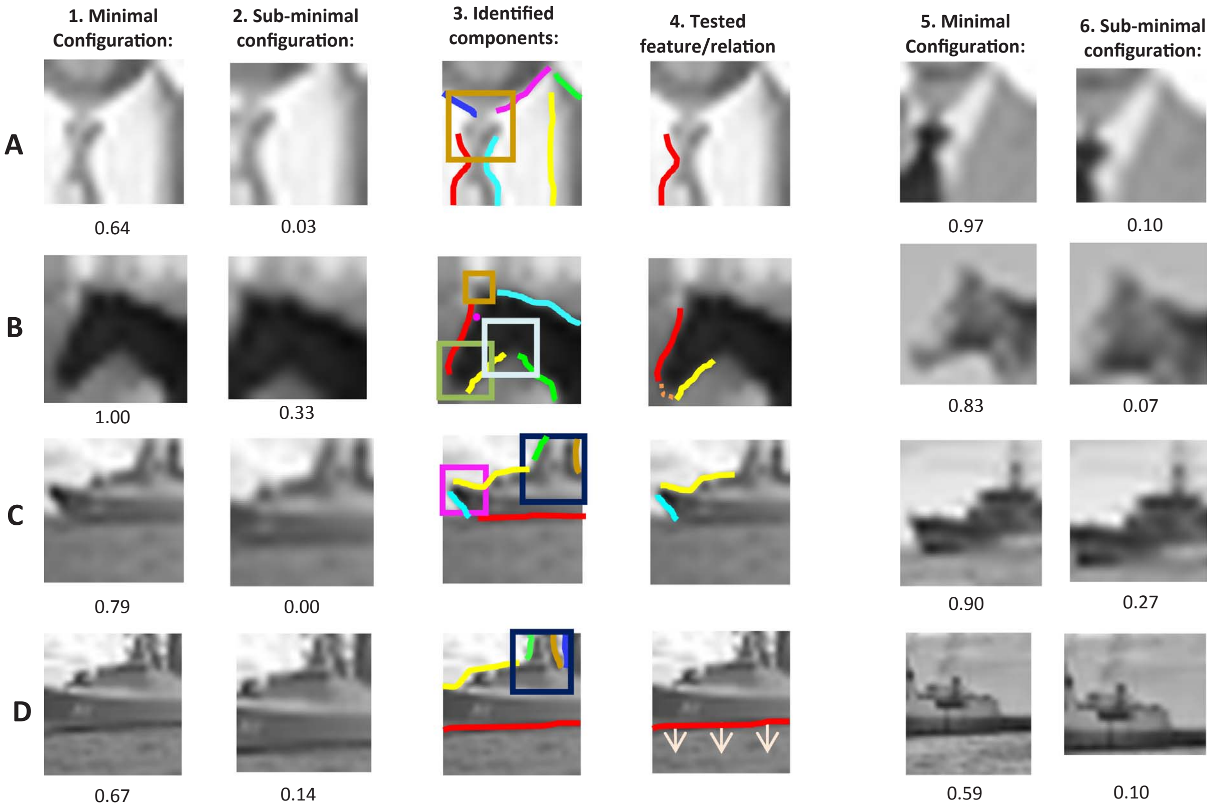
1. Minimal Configuration:
2. Sub-minimal configuration:
3. Identified components:
4. Tested feature/relation
5. Minimal Configuration:
6. Sub-minimal configuration:
A
0.64
0.03
0.97
0.10
B
1.00
0.33
0.83
0.07
C
0.79
0.00
0.90
0.27
D
0.67
0.14
0.59
0.10

Fig. 7. Inferring relations between internal components with large contribution to recognition and interpretation. Minimal and sub-minimal pairs (columns 1, 2, recognition rate shown below the images), are shown with internal components recognized by humans in the minimal images (column 3). To identify useful components and relations for interpretation, we compared the minimal and sub-minimal images. Using the identified components, we found if any component in (1) are missing in (2). Using the set of candidate relations, we identified relations that are satisfied in (1) but not in (2). The contribution of each missing component or relation was then evaluated using training examples (see text). When necessary, several alternatives were evaluated. Examples of informative components and relations are shown in column 4. Examples of additional MIRC/sub-MIRC pairs in the training set with the same missing component or relation, with its effect on recognition, are shown in columns 5, 6. Inferred components and relations illustrated in the figure are: missing contour element (in A), connectedness of two contours (B), contours meet at high curvature (C), and characteristic texture in a region (bounded by the red contour and image border) in (D).

interpretation process, we have trained and compared two alternative versions of the interpretation model. The first version, (termed ‘basic’), included a limited set of relations commonly used in the visual structure modeling literature (Section 2), namely, unary relations based on local texture and shape appearance, and binary ones based on the relative displacement of components. The basic model is then compared with a second interpretation model (termed ‘augmented’), where the basic set of relations is augmented with the relation we wish to test. Performance of both models was evaluated on a data set, which included for each of the minimal images in Fig. 4, a set of 120 positive examples, and 8000 negative (non-class) examples, split between training and validation sets. Performance of the two models was compared by classification by the random forest classifier (using the Out-Of-Bag test for strength of random forest features, Breiman, 2001, Appendix B), to assess the contribution of each new relation. Relations that improved random forest classification average precision by 1% or more (found in pilot experiments to be significant), were incorporated in our final extended set of relations. The extended set was subsequently used in the overall evaluation of the model, applied to the interpretation and recognition of minimal and sub-minimal images (Section 5). Fig. 7 illustrates the process for example relations, which were found to be informative for interpreting the corresponding minimal images.

4.3. Useful relations from ‘hard negative’ examples

In addition to the sub-minimal images test discussed above, which compared images from the same class, a complementary source for identifying useful relations for full interpretation is a comparison of minimal configurations with ‘hard’ non-class examples, which are difficult in the sense that they are confusable with true class examples by current computational models (a deep net model, Simonyan & Zisserman, 2015, and a human recognition model, Serre et al., 2007). Such a comparison can identify components and relations that are informative for human recognition and interpretation, but are missing from current models. We describe next how hard-negative examples were generated and how they were used to identify useful relations for interpretation.

To identify hard negative examples for a given minimal image, we trained a deep CNN-based classifier. We used the 19-layer CNN model described in Simonyan and Zisserman (2015), adjusted to recognize minimal images as follows: we fine-tuned the network to classify regions at the size of minimal images, and then used an intermediate layer (conv3_4, the 8th convolutional layer) as a descriptor for a final SVM classifier. We chose this intermediate layer based on its classification performance, compared to other network layers, and to complete end-to-end fine-tuning of the network.). We trained the classifier using 120 examples of the minimal image (see Section 5 for how these examples were obtained), and a large set of negative examples (200,000 local regions cropped and rescaled from various non-class images). We then applied the classifier on a validation set (equal in size to the training set), and finally retained the 4000 non-class image regions with the highest detection scores. These are the hard-negative examples, used in the search for informative relations. Similar to the use of sub-minimal images described above, the search proceeds along the following steps.

We start with the ‘basic’ interpretation model as defined in Section 4.2 and iterate over the following procedure:

(i) Keep the k hard-negative images that received the highest interpretation score (since images later required MTurk tests, we used the limit k = 40).
(ii) Confirm (using MTurk testing) that these negative examples are not confusable for human observers. (Examples that were also difficult for humans were removed from the set in practice, no more than 2 examples were removed at this stage).
(iii) Compare the interpretation produced by the model for the images collected in (i), to human annotations of the corresponding minimal image examples. As in Section 4.2, identify components or relations (from the list of candidate relations) present in the positive examples but not in the hard negatives.
(iv) For each such a relation, test its contribution to the interpretation model by the difference in random forest classification with and without this addition, as in Section 4.2.
(v) Once relations from all hard negative images were tested, and the contributing subset was added to the relations set, train a new version of the interpretation model and repeat the search from step (i), to discover additional relations from hard negatives of the new version.

We iterated this procedure until no new contributing relations were found (at most 3 iterations were needed per class).

Fig. 8 illustrates examples of hard negatives discovered and used to identify informative relations, and the process of finding these relations. Examples include 'highly-straight' contours (typical for man-made objects) in the horse head (e.g., the red and yellow contours in Fig. 8A), the connectedness of horse head contours through the ear region (red and cyan contours in Fig. 8B), sharp corners at the tie knot's (cyan and magenta contours, connected inside the brown square in Fig. 8C), and coherent visual texture (or intensity level) between the two shirt parts (the area that is left to the red contour and the area that is bounded by the contours in cyan and yellow in Fig. 8D).

4.4. The final set of relations

The final set of relations, obtained by comparing MIRCs to both sub-MIRCs and hard negatives, includes unary relations (properties), binary relations, and relations among three or more primitives. Relations in this set are composed of basic relations as listed in Section 4.2, extended with candidate relations which proved to contribute to the recognition and interpretation accuracy by the computational experiments in Sections 4.2 and 4.3. Relations in the model range from low-complexity ones such as computing relative location between primitives, to higher complexity procedures such as computing the continuity, bridging (connectedness), or parallelism of contours. Table 1 lists relations with the highest contribution, as measured in Sections 4.2 and 4.3. Below, we refer to the set of relations in Table 1 as the 'extended set' of relations, and we distinguish it from the 'basic set' of relations, namely relations 1, 4 and 5 in Table 1, which are based on local appearance, location, and displacement, and were wieldy used in previous visual structure approaches (see Section 2). Technical details for implementing the relation procedures, including their higher order versions that include more than two primitives, are discussed in Appendix C.

We further compared the relative contribution of individual binary vs. unary relations to successful interpretation, by comparing the model performance with and without the relation in question (Section 4.2, Appendix B.1). By this measure, binary relations contribute more on average to reducing ambiguity than the unary ones; however, some unary relations have high contribution as well (e.g., 'straightness', 'intensity minimum/maximum').

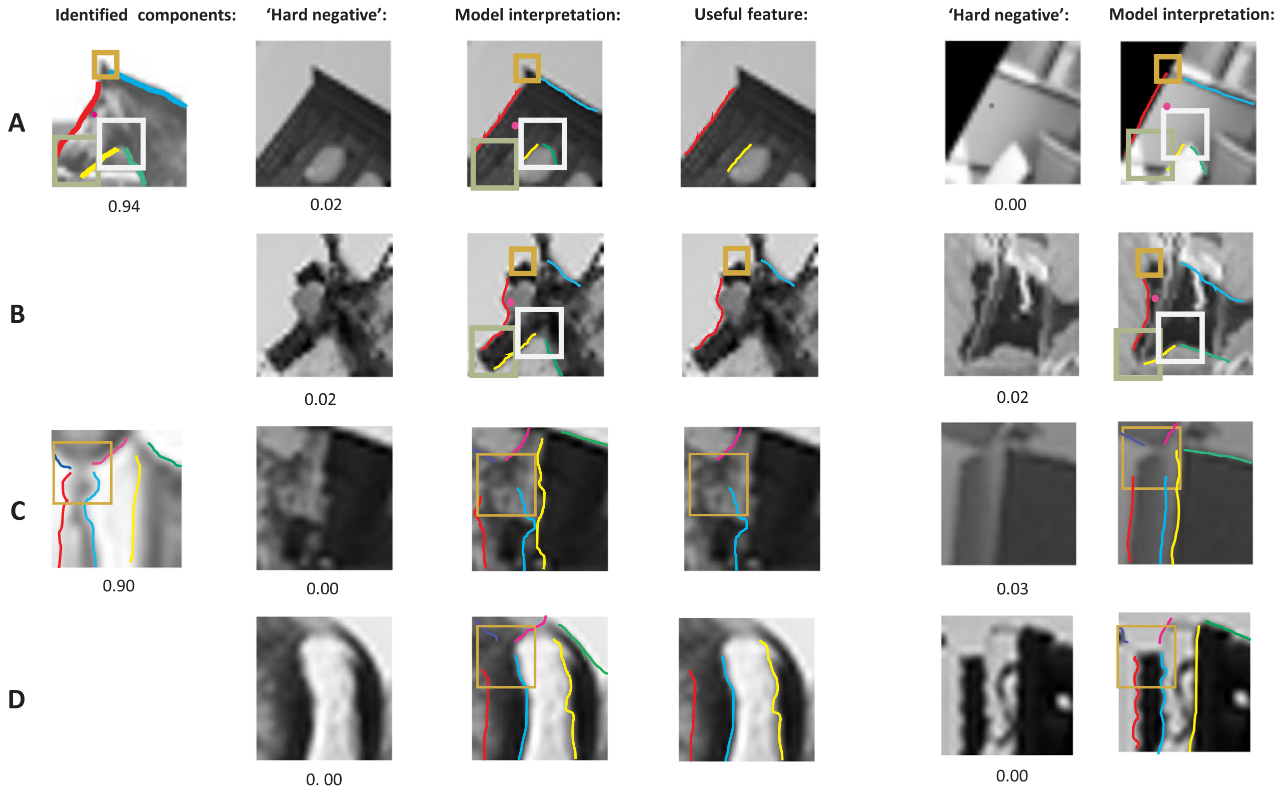

Fig. 8. Useful relations for interpretation extracted from 'hard negative' examples. Columns show (left to right): minimal images with their human interpretation, non-class examples with high detection score and their human recognition rate, interpretation applied to the negative example by the model. Differences in components or relations are identified and evaluated, see text. Column 4 shows relations found to be informative for the interpretation model. They include: high straightness of two contours, typical of man-made objects (in A), connectedness of two contours through the ear region (in B), connectedness of two contours through a tie knot region (in C), coherent texture between the two shirt parts, see text (in D). The identified relations were used to reject hard negatives, examples in the last two columns.

Table 1
Relations that were found informative for the learning process, by the method and criteria in Sections 4.2 and 4.3. See implementation details for relation procedures in Appendix C. Relations 1, 4 and 5 form the 'basic set' of relations (widely used in previous visual structure approaches, see Section 2), relations 1–14 are part of the 'extended set' of relations.

	Relation operands	Description
1	All primitives	**Location and relative location:** for all primitives, and for all pairs of primitives in the structure
2	Point	**Strength of intensity maxima/minima**, center-surround filter responses at a point location
3	Contour	**Deviation from line/circular arc:** in particular for man-made objects
4	Contour	**Visual appearance along contour** distribution of visual appearance/texture features along contour
5	Region	**Visual appearance inside a region** distribution of visual appearance/texture features in a region
6	Contour, Contour	**Relative location of contour endings:** between endings of two different contours
7	Contour, Contour	**Continuity:** smooth continuation between two given contour endings
8	Contour, Contour	**Length ratio** between two contours
9	Contour, Contour	**Parallelism** between two contours
10	Region, Region	**Coherent visual appearance** similar appearance/texture features in region i and in region j
11	Contour, Point	**Cover of a point by a contour:** if a contour i covers a point j. For 'cover' refer to Appendix C
12	Contour, Region	**Contour ends in a region:** if a contour i ends in a region j
13	Point, Region	**Containment**: if point i is inside region j
14	Contour, Contour, Region	**Contour Bridging:** Testing whether two disconnected contour elements can be bridged (linked in the edge map)

5. Experimental evaluation

So far, we have identified the useful components and relations when tested individually. We next combined all of them in the full interpretation model (as described in Section 3) and tested its performance. The full set of relations for the trained model was composed of the extended set of relations listed in Table 1. To evaluate the full interpretation model, we performed experiments to assess (i) the interpretation correctness on novel images, (ii) the ability of the interpretation model to predict human recognition at the level of minimal image, and (iii) the contribution of informative relations included in the model to human recognition, using modified minimal images.

Training of the model was obtained as described in Section 3, with annotated examples of minimal images, and non-class (negative) examples. To get positive class examples for the minimal image we wanted to model (e.g., 'horse-head'), we collected fully-viewed object images from known data sets (Flicker, Google images, ImageNet), and manually extracted from each image a local region at the position and size similar to the discovered minimal image (Ullman et al., 2016). The minimal image examples used for training were in slightly higher resolution than the minimal images found in Ullman et al., 2016 (image resolution was increased by factor of 1.5), since we found that using this scale during training improved the model results when applied to novel images.

To have ground truth for the interpretation, two human subjects provided annotation of the set of primitives for all examples (one annotator used for ground truth, the other for measuring consistency, details in Appendix A). Negative (non-class) examples for training were collected automatically from cropped windows in non-class images at similar size to the minimal image. To get hard negative examples, we trained a deep CNN classifier (Simonyan & Zisserman, 2015), as described in Section 4.3, and collected images that received high recognition scores. We next turn to describe our three testing procedures, in Sections 5.1–5.3 below.

5.1. Comparing model output to human interpretation

The interpretations produced by the model were compared with the ground truth annotations supplied by the human annotators. Since the model is novel in terms of producing full interpretation, it cannot be compared directly with any existing alternative models. However, we made our set of annotations publicly available, and the current model provides a baseline to also evaluate future results. To assess the role of the extended relations derived in Sections 4.2 and 4.3, we compared results from two versions of our model, which differed in the relations included in the model: one using only the basic, and the other using the extended set of relations.

Fig. 9 shows examples of the interpretations produced by the model with the extended set for novel test images. To assess the interpretations, we matched the model output to human annotations for multiple examples. Our training set contained 120 positive examples, and 25,000 negative examples for each interpretation model. Our test set contained 480 examples for the horse head minimal image (Fig. 4A), 330 examples for the man-in-suit minimal image (Fig. 4B), and 120 of the eye (Fig. 4D) and the bike (Fig. 4E) minimal images (with same resolution as the training set). We automatically matched the ground truth annotated primitives to the interpretation output by the so-called Jaccard index, (Tan, Steinbach, & Kumar, 2006), which is a commonly used similarity measure for comparing automatic detection results (high Jaccard means similar interpretations). This index compares the similarity of two regions, by the area of the regions' intersection divided by area of their union, and was adapted to compare the accuracy of detecting region, contour, and point primitives, as illustrated in Fig. 10, and explained in more details in Appendix D. Table 2 shows results for the basic and extended relation sets, as well as agreement between different human annotators, which can serve as an upper bound for

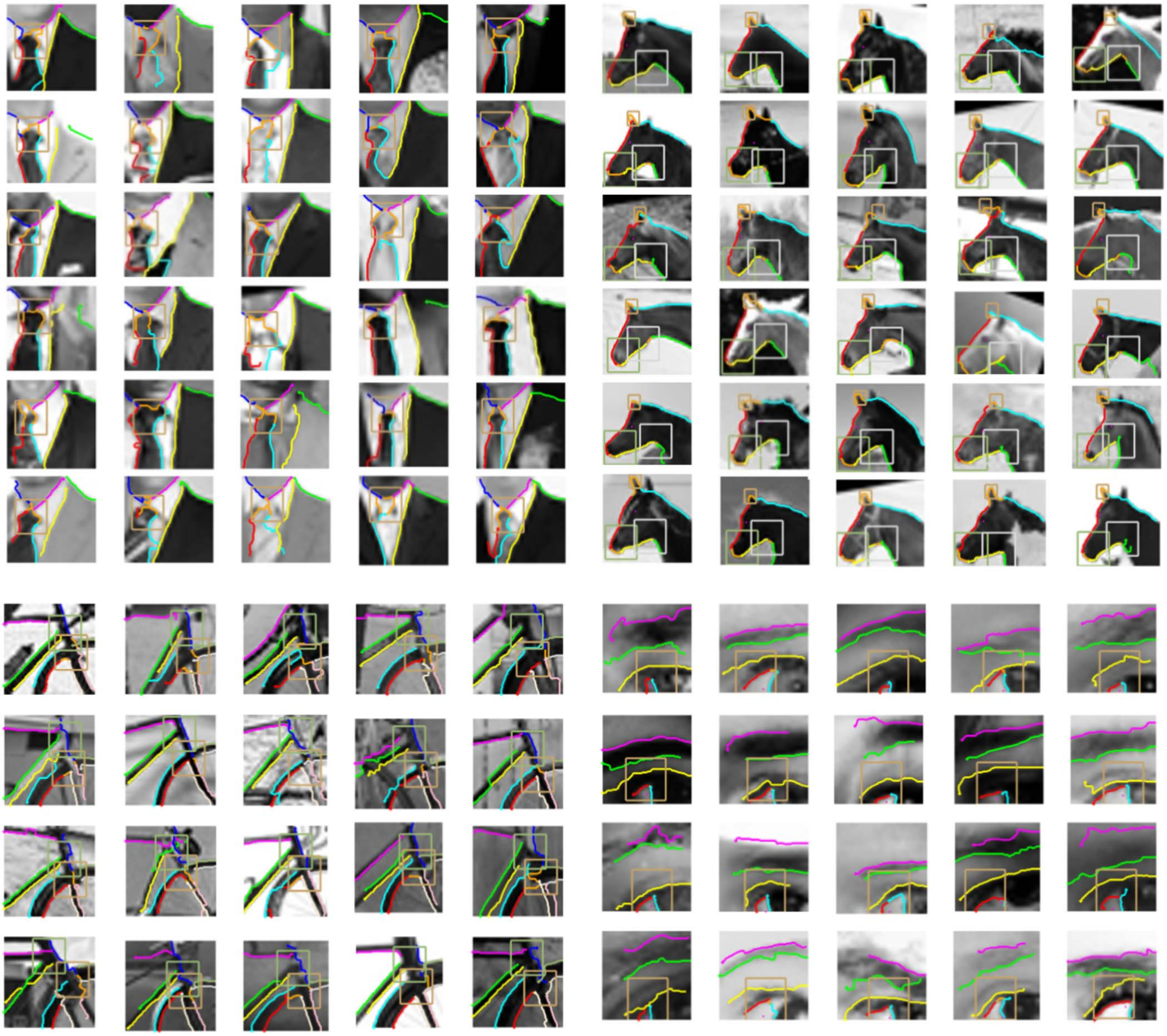

Fig. 9. Interpretation results for minimal images belonging to (counter-clockwise) a horse-head, a man in a suit, a bike, and an eye.

comparing interpretation performance. Interpretation using the extended set of relations was significantly closer to the 'ground truth' human interpretation compared with the use of basic set of relations ($P < 4.99 \times 10^{-11}$ for all primitives in 4 classes, n = 33, one-tailed paired *t test*). However, the agreement between the model and ground truth interpretations was still lower than the agreement between different human interpretations ($P < 1.14 \times 10^{-13}$ for all primitives in 4 classes, n = 33, one-tailed paired *t test*).

5.2. Interpretation for predicting minimal and sub-minimal images

The link between interpretation and recognition, as discussed in Section 1.3, suggests that the interpretation score (which is a part of the model output) may be used as a part of the human recognition process at the minimal image level. In particular, it is interesting to compare the interpretation scores for minimal and sub-minimal images, to assess the usefulness of interpretation for recognition. In human perception, there is a sharp drop in recognition rates at the minimal image level: a small change to the image can have drastic effects on recognition rate (Section 1.2, Ullman et al., 2016). This sharp drop was not reproduced by computational models of recognition, and it therefore becomes of interest to examine whether the internal interpretation of minimal image may provide a basis for this perceptual sensitivity. It is possible, for example, that even small changes to a minimal images could disrupt the presence of key elements and their relations. To test this possibility, we measured the gap between human recognition rates for minimal and

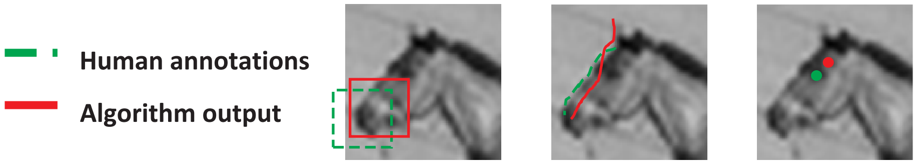

Fig. 10. Quantitative evaluation of the model interpretation results. We compared interpretation results to human annotations based on the Jaccard measure similarity criteria: for regions, contours, and points (see Appendix D for details).

Table 2
Accuracy of the interpretation results, comparing the basic model, extended model, and human annotators. Accuracy is measured by the average Jaccard index between the model interpretation and ground truth supplied by human annotations. For comparison, human accuracy is measured by the agreement, measured by the Jaccard index, between the human annotators.

	Basic	Extended	Humans		Basic	Extended	Humans
Horse-head				Man-In-Suit			
Ear region	0.11	0.37	0.60	Knot region	0.62	0.66	0.74
Mouth region	0.69	0.76	0.85	Left tie contour	0.48	0.55	0.72
Neck region	0.55	0.68	0.74	Right tie contour	0.47	0.53	0.72
Upper head contour	0.44	0.69	0.84	Suit-shirt contour	0.64	0.73	0.83
Mane contour	0.34	0.61	0.79	Shoulder contour	0.50	0.63	0.66
Lower head contour	0.46	0.66	0.79	Left neck contour	0.49	0.65	0.84
Lower neck contour	0.32	0.63	0.74	Right neck contour	0.39	0.49	0.77
Eye point	0.29	0.49	0.60	All primitives			
All primitives	0.40	0. 61	0.75		0.51	0.61	0.77
Eye				Bike			
Iris region	0.39	0.56	0.79	Fork region	0.72	0.73	0.80
Lower lid contour	0.47	0.62	0.73	Tire lower contour (left side)	0.68	0.75	0.86
Cornea contour	0.33	0.60	0.81	Tire lower contour (left side)	0.62	0.75	0.90
Upper lid contour	0.41	0.64	0.74	Bottom tube's upper contour	0.59	0.74	0.86
Lower eyebrow contour	0.51	0.64	0.83	Bottom tube's lower contour	0.54	0.70	0.87
Upper eyebrow contour	0.45	0.51	0.81	Top tube's lower contour	0.36	0.43	0.84
Sclera point	0.56	0.54	0.79	Head tube contour	0.49	0.60	0.85
All primitives	0.44	0.59	0.78	Tire upper contour (right side)	0.50	0.62	0.81
				Tire lower contour (right side)	0.53	0.57	0.78
				Fork left contour	0.60	0.68	0.82
				Fork right contour	0.59	0.71	0.83
				All primitives	0.56	0.66	0.84

sub-minimal images (via MTurk search on new image examples) and compared it to the gap predicted by two models: the current interpretation model, and a classifier based on deep convolutional networks (very-deep CNN, Simonyan & Zisserman, 2015), trained on minimal image examples, as in Section 4.3. Our test set included 12 examples of minimal images and 20 examples of sub-minimal images for each of two minimal image categories: the horse-head (Fig. 4A) and man-in-suit (Fig. 4B). The average gap measured between human recognition rates for minimal images and for sub-minimal images was 0.75 for the horse head, 0.74 for man-in-suit. This sharp gap in human recognition at the minimal image level was compared with the computational models as described next.

To compute the recall gap of models, the model's classification score was compared against an acceptance threshold, and scores above threshold were considered class detections. For each model, we set the acceptance threshold to match the human recognition rate. For example, for the man-in-suit, the average human recognition across all 12 examples was 0.88, and the model threshold was set so that 11/12 examples will be accepted (see Fig. 11A). Recognition rate for the sub-minimal images was then derived from the fraction of sub-minimal images exceeding the threshold, and the difference in recognition rates defines the model's recognition gap.

The scores of the CNN model for the minimal and sub-minimal images on the test sets are shown in Fig. 11A and C. The gaps computed for the horse-head and man-in-suit were 0.20, and 0.37, respectively, both considerably smaller than the human recognition gap. The second model tested the interpretation trained as in Section 5.1, with the extended set of relations. Interpretation scores are shown in Fig. 11B and D, along with the interpretation examples of minimal and sub-minimal pair from each category. The average interpretation gap was 0.75 for the horse-head and 0.76 for the man-in-suit, closely similar to the gaps measured for humans. The differences in recognition gap between the CNN and interpretation models were highly significant ($P < 2.44 \times 10^{-4}$ for horse-head, $P < 5.7 \times 10^{-3}$ for man-in-suit, $n = 20$, *Fisher's exact test*). The difference is likely to arise because the interpretation model incorporates class-specific properties and relations that are not included in the CNN model. We discuss this difference further in Sections 6.3,6.4 below.

5.3. *Testing predicted relations via intervention on minimal images*

The interpretation model includes informative relations between components, which were identified using the data sets of sub-minimal images and hard negatives. The model predicts that disrupting these relations should reduce the ability of human observers to recognize and interpret minimal images. To further verify the role of these relations, we used direct intervention (Pearl, 2009) on minimal images, testing whether removing specific relations from the minimal image will decrease human recognition. For this purpose, we created transformed versions of the minimal images, in which specific relations were selectively manipulated. The transformed versions were then tested psychophysically via the MTurk.

The transformations applied to minimal images included rendering sketches, including rendering k-color cartoons ($k \le 5$), and re-coloring a small set of pixels (number of re-colored pixels ≤ 4), examples in Fig. 12. Such sketches are typically highly recognizable, and are similar in terms of perceptual and brain responses to natural images (Walther, Chai, Caddigan, Beck, & Fei-Fei, 2011). To create sketches, we traced contours of the original MIRC image, either manually as in Fig. 12A, B (right column), and C, or semi-manually using straight lines, as in Fig. 12B, middle column. Cartoon sketches are similar, but using a small number of grey-levels (≤ 5) for the regions (12D). Re-coloring images were done with interactive graphics design tools (Irfan, Photoshop). For all sketches, we kept all contours or segments in the minimal image that are used as primitives in the interpretation model, and verified that the sketched images were still recognizable (e.g., Fig. A–D, middle column).

In the sketch images, a specific contour or a region can be selectively modified, with minimal or no change to other image parts. We created a modified version for each sketch, where selected contours or regions were changed based on the tested property or relation (e.g., Fig. A–D, right column). Since we know how a relation is computed in the model, we can change contours or regions such that this relation will no longer be detected. We then tested whether the specific disruption of a single relation will cause a significant drop in MIRC recognition as

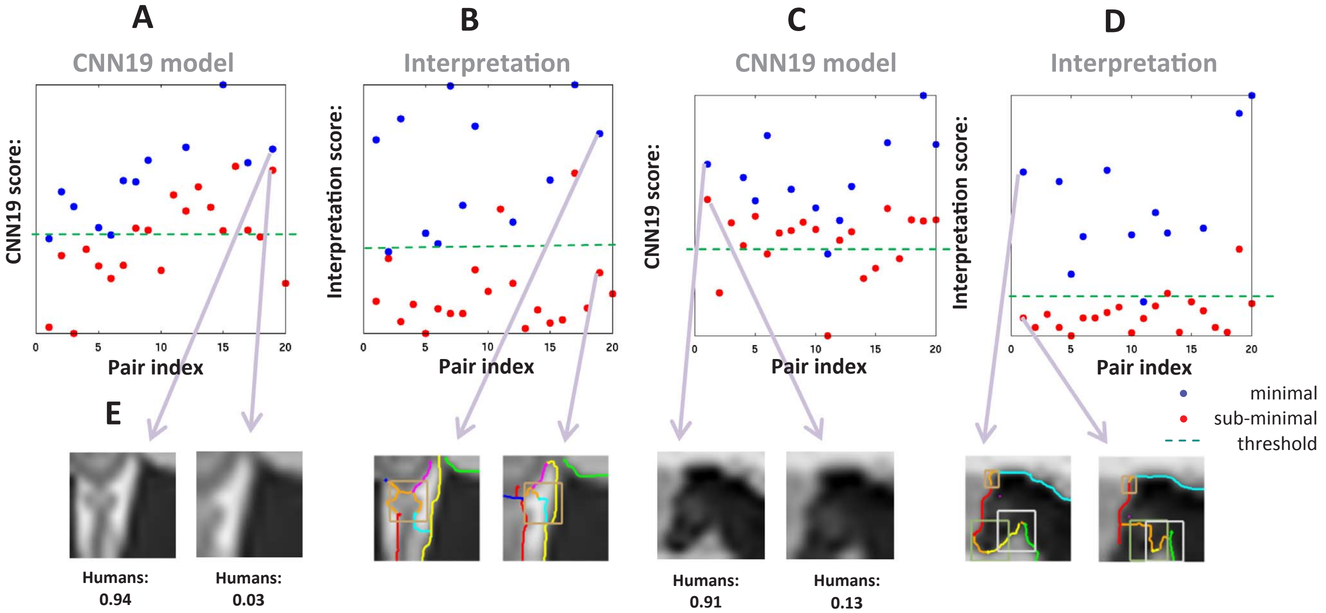

Fig. 11. Recognition of minimal and sub-minimal images. Recognition gaps between minimal and sub-minimal images were computed for two recognition models, a 19-layer feed-forward CNN classifier, and the interpretation model (trained on a similar set of examples, see details in Section 4.3 for CNN, and Section 5.1 for interpretation). Results were compared with the large gaps that characterize human recognition of minimal images. **(A).** The panel shows recognition scores (y-axis) of minimal images (blue dots) and sub-minimal images (red), as computed by the CNN classifier. Each column (marked 1–20 on the x-axis) shows scores of one pair of minimal and its sub-minimal image. (A single minimal image can have more than one sub-minimal image). Green dashed line represents the human recognition rate, and recognition gap for the model is derived from the number of blue and red points above the threshold green line (see text for details). **(B).** Similar to A, but computing the recognition gap for the interpretation Model. **(C and D).** Similar to A, B but applied to the horse-head images. The simulation results show that the sharp drop in human recognition between minimal and their sub-minimal images, is reproduces by the interpretation model but not by the deep CNN model. **(E).** Examples of minimal and sub-minimal pairs and their interpretation, produced by the model. The interpretation of the minimal images is more accurate compared with the sub-minimal ones. The arrows show the corresponding score of each image by the interpretation and CNN models. Note that the scores of the minimal and sub-minimal images are more separated by the interpretation model compared with the CNN model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

predicted by the model.

The tested relations were taken from the set of the most informative relations in the relations set (Table 1). For each tested relation, we first applied a manipulation which removes the relation from the model relations vector (the computed R_C) while keeping the rest of the relations intact (the model can provide interpretation for both natural and sketched minimal images).

Each relation was tested using five different pairs of manipulated

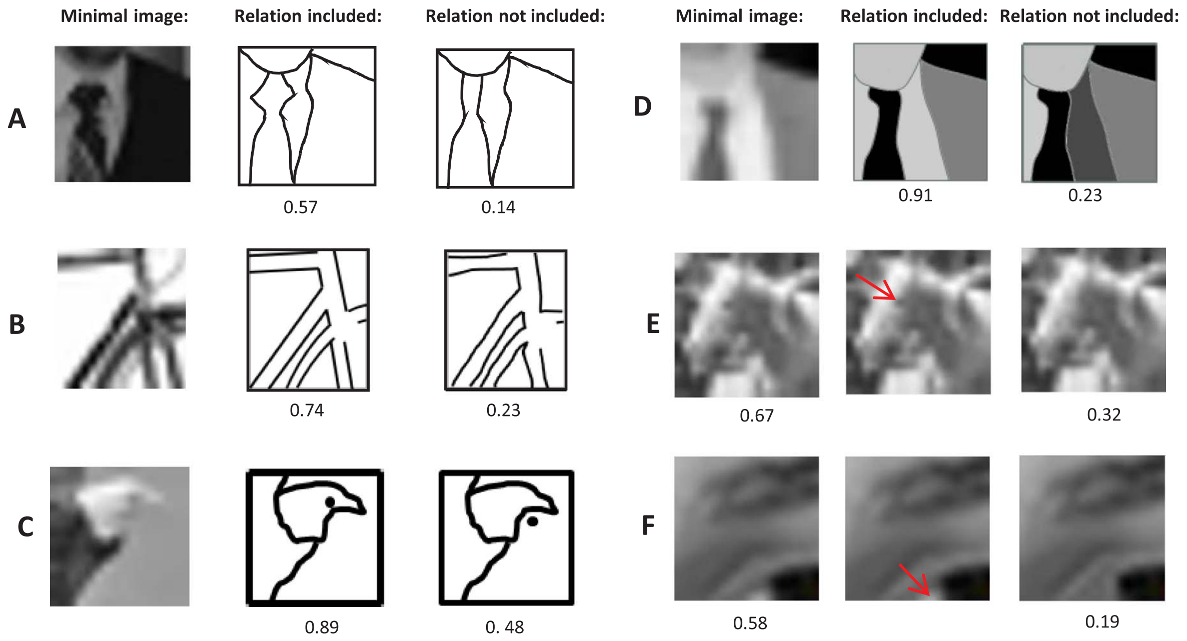

Fig. 12. 'Intervention': Testing informative relations via transformed minimal images. **(A–C).** Rendering sketches from images. **(D).** Creating k-color cartoons. **(E, F).** Re-coloring a small set of pixels (≤4, pointed by the red arrow) with the same color of their neighboring pixels. In a transformed image, a relation is removed to test its predicted role in human perception. Relations tested: sharp curvature in the tie contour (in A), high contour straightness (in B), containment of a point in bounded contours (in C), coherent color/texture in the two parts (in D), minimum intensity (in E), and maximum intensity (in F). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and non-manipulated versions, and the average human recognition drop for each relation was measured. Example results are shown in Fig. 12. Fig. 12A–D used sketches from minimal images. The sketched versions eliminated specific relations in the representations: sharp curvature (12A, the tie knot, cf. Fig. 8C), high straightness measure (12B, bike contours, cf. Fig. 8A), containment of a point in region (12C, bird's eye) and the coherent appearance (in intensity or texture) between two regions (12D, cf. Fig. 8D). In Fig. 12E and F, a local change was introduced to disrupt the model property of minimal (12E) or maximal (12F) local intensity. The change was induced by re-coloring 3–4 pixels, to match the average intensity of their neighboring pixels.

For all tested relations in Fig. 12, the manipulation resulted in a significant drop in human recognition rate. (For example, Fig. 12A, 5 image pairs, average drop = 0.41, $P < 2.46 \times 10^{-4}$, n = 5, one-tailed paired *t test*. In similar one-tailed paired *t tests* for Fig. B–F, $P < 0.0052$ for all cases). In summary, the results show a sharp drop in recognition following intervention to eliminate a relation predicted by the model to be highly informative for the interpretation of the relevant minimal image. This agreement between the model and human recognition supports the proposed role of the tested relations in human recognition and interpretation of minimal images.

5.4. Dealing with variability

The interpretation process identifies features in the image that correspond to semantic components in the stored MIRC model. This correspondence between model and image features may be disrupted in several ways (or their combination). It may not be one-to-one because the image either lacks a model feature, or it may have additional ones. Alternatively, a feature in the model may be replaced by a different one in the image. To be robust, the interpretation process is required to deal with such changes, which can be caused by natural image variations. Because of its local nature and the limited number of components, image variability in MIRCs is reduced to a minimum. In addition, the interpretation model can cope with significant image variability as described below. With respect to losing a feature, in the case of a MIRC, such a reduction will render it a sub-MIRC, which cannot be reliably recognized. This means that to recognize an object, we need at least one of its multiple MIRCs to be preserved. The constraint can be relaxed in a more complex scheme, where two or more MIRCs that are below recognition can be combined. This raises an interesting empirical question about human vision, testing whether the integrity of at least one MIRCs in an object image is a requirement for recognition. We suspect that this may be the case, but the question is open for empirical studies.

Additional features in the image, beyond the features included in the MIRC model, can be tolerated by the current scheme, since the algorithm searches for a configuration in the image that matches the model, without requiring to match all the image features. Additional features may also be approached by feature matching that allows some degree of many-to-many matches (Demirci, Shokoufandeh, Keselman, Bretzner, & Dickinson, 2006). With respect to replacing a model feature by a different image feature, the current scheme can tolerate significant changes between the model and image features. It can allow the replacement of a feature by another and can allow a feature to change within a broad range of parameters (e.g. a range of orientation, curvatures etc.) This is obtained in the model in two ways. First, the model learns to use abstract relations, which allow considerable variation. For example, the tie contour (Fig. 9) are required to be roughly parallel and end at the knot region, but are allowed to change considerably in orientation and be either straight or curved. Such abstract relations allow for variability of primitives' shape and appearance, and for some degree of many-to-one matching (e.g., the bike tubes in Fig. 9). Second, the random forest representation allows for several possible correct configurations of primitives and relations, which can be captured by its different trees. A configuration of primitive candidates is represented by a vector of candidates' properties and relations (the 'relations vector'), which is then given to each tree in the random forest (Appendix B). If different features, or a range of parameters were present during training, it becomes likely that different relations vectors, which represent allowed variations of the correct interpretation will get high score. The range of variations allowed in this manner may be more restricted than, e.g., in general stochastic grammars. When the range of variations are too large, then an additional MIRC model will be required to cover the full range.

6. Discussion and implications

In this work, we described a model for local image interpretation, applied to minimal recognizable images. The ultimate goal of full image interpretation is to recognize meaningful semantic components anywhere in the image, but we used minimal images for the development and testing of the model for two reasons. First, local interpretation reduces the number of components and the complexity of the model, and second, using a data set of minimal and sub-minimal images is useful for identifying informative components and relations, which play a part in the interpretation process.

The interpretation model was shown to produce reliable interpretation of local image regions. It also helps to explain the sharp drop in recognition between minimal and sub-minimal images, which is characteristic of human observers, but not reproduced by current bottom-up computational models. It will be interesting to further test in the future the agreement between human recognition errors of difficult images and errors made by recognition models, with and without an interpretation stage.

Similar to other cognitive and computational models, interpretation is defined in the model in terms of a local structure, composed of components, properties, and relations. Our empirical testing of properties and relations proposed in past studies (Section 4.1), showed that a number of them contributed to the performance of the model (Table 1). In comparison, restricting the relations to relative displacements between components ('basic' relations, 1, 4, 5 in Table 1), which are commonly used in computational models, proved insufficient for reliable interpretation. Consistent with this computational evidence, a subset of the relations used by the model were tested and found to directly affect human recognition, as human recognition of modified minimal images, where tested relations were excluded, dropped significantly. Taken together, the role of the components and relations incorporated in the interpretation model is supported by three complementary sources of evidence: their contribution to correct interpretation by the model, the effect they have on the sharp difference in recognition between MIRCs and sub-MIRCs, and the effects of their selective elimination from minimal images on human recognition of these images.

Future work in modeling the interpretation process should go beyond the interpretation of local regions discussed in this study, towards the interpretation of full, natural images. The interpretation of full images is likely to be goal-directed, namely, providing detailed interpretation of regions of interest, rather than uniformly across the image. Minimal images, at multiple scales, can provide a natural starting point for the fuller interpretation process, because they can be reliably recognized and interpreted on their own, independent of the surrounding context, and can subsequently help in further disambiguation and interpretation of nearby regions.

6.1. Detailed interpretation for complex visual tasks

Full interpretation of semantic components at the level produced by the current model can play a useful role for extracting meaning from complex configurations, arising in tasks such as recognizing actions or social interactions between agents. The reason is that the exact meaning of an image may depend on fine localization of object parts and the relations between relevant parts, as illustrated in Fig. 13. In particular,

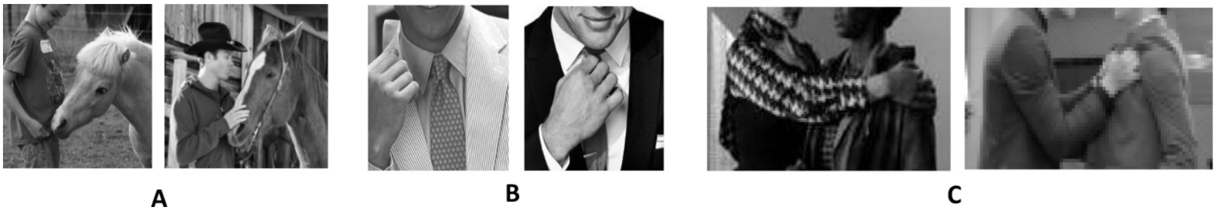

Fig. 13. Examples of fine interpretation in recognizing human actions and interactions. **(A).** Recognizing petting vs. feeding a horse (Yao et al., 2011) depends on the exact location of the human hand on the horse muzzle. **(B).** Whether the hand is touching the knot or not, determines the action of 'fixing a tie'. **(C).** The hands contact locations provide important cues for recognizing a 'hug' interaction between the agents.

the recognition of agents' social interactions by computational models has proven difficult, and is still a largely open problem. It will be of interest to extend in the future the current work, to study the role of detailed image interpretation in complex scenes, including the recognition of social interactions.

6.2. Learning relations

In the current model, relations between components of the local interpretation are used to identify the correct structure. There are two main questions regarding the relations used for the purpose of interpretation. The first is the full set of relations that are useful for the task, and the second is identifying informative relations for a particular local structure (e.g., horse-head). Since the set of so-called 'basic' relations proved insufficient, we evaluated a larger set of relations, using minimal, sub-minimal, and difficult non-class images. The resulting set is not necessarily complete, and future studies may identify additional relevant relations. In terms of the human visual systems, such relations could be in part pre-existing in the visual system, and in part learned from visual experience. Regarding the identification of informative relations for a novel class of images, the approach in the model was to examine the full set of possible relations, and identify the informative ones using positive and negative examples, where the negative examples came from high-scoring non-class examples. It will be of interest to examine in the future the possibility of replacing this search by network learning models, based on positive and negative examples, but without using an explicit set of possible relations. The issue of unsupervised learning of semantic components is left for future studies, we only note that some components may be learned based on their independent motion within the image (e.g. an eye or mouth within a face), or based on points of contact between an agent and an object (such as a cup-handle or door-knob).

6.3. Interpretation and top-down processing

Our model suggests that the relations required for a detailed interpretation are in part considerably more complex than spatial relations used in current recognition models (Section 2). Furthermore, the experimental results show that the relations used for interpretation are often class-specific, in the sense that the most informative relations for the interpretation of a given class often depend on the class. This does not mean that a given relation *R* is specific to a single class, but that it is typically used in the representation of some classes, and not others. This is illustrated in Table 3, which shows the most informative relations found by the model for the interpretation of 4 different classes of minimal images. Since the subsets of informative relations are class-dependent, it will be computationally efficient to compute the more complex relations selectively, in a class-specific manner, rather than computing all possible relations for all candidate classes. In addition, even when *R* is computed for a given class, it can be computed between some components, but not others. In such a scheme, the interpretation process will be naturally divided into two main stages. The first is a bottom-up recognition stage, similar to current feed-forward models. This stage will lead to the activation of one or several objects classes, but without detailed object interpretation. The activated classes will then trigger a top-down process for the computation of further class-specific components and relations required for a detailed interpretation. The interpretation will also be used for validation of the activated classes in the first stage, by rejecting bottom-up detections which do not have the expected interpretation. Future studies could explore this two-stage proposal further by psychophysical and physiological methods. For example, since the accurate recognition of minimal images depends in the model on its internal interpretation, the top-down component predicts that the reliable recognition and interpretation of minimal images will be a relatively slow process compared with a single feed-forward pass.

A successful recognition scheme should be able to tolerate natural variations in image transformations, such as changes in position, scale and orientation, combined with distortions and occlusion. In the two-stage model, invariance to image transformations is determined by both the first, bottom-up stage, and by the following top-down stage. Regarding translation and rotation, invariance in the model depends primarily on the bottom-up stage. Current bottom-up models can identify and localize candidate objects or parts at different positions, and can tolerate a range of rotations, which depends on the variability encountered in training. In terms of scale, motivated in part by the human variable resolution in image sampling and representation, our model analyzes the image at multiple scales, from which candidate MIRC regions are derived. Regarding occlusion, the minimality of MIRCs implies that both humans and the model will be affected by occlusion, since occlusion can turn a MIRCs into its sub-minimal version. However, large object occlusions will be tolerated, as long as at least one of the multiple MIRCs remains visible. Tolerance to deformations arises in the model from two sources: First, MIRCs are often limited to local object regions, which are inherently more tolerant to deformations than larger regions. Second, as mentioned above (Section

Table 3
Top 3 informative relations found for the different class models of minimal images.

Horse-head	Man-in-suit	Eye	Bike
Intensity minimum (at the eye point)	**Contour appearance** (along the tie)	**Deviation from circular** (lid upper contour)	**Parallelism** (tube contours)
Contour Bridging of the mane and mouth upper contours	**Region appearance** (suit region)	**Cover of point by contour** (sclera by lid contour)	**Continuity** (tire upper contours)
Contour Bridging (at the mouth)	**Contour ending in region** (tie contour in knot region)	**Relative contour endings** (lower lid and the iris contours)	**Region appearance** (wheel region)

5.4), the use of abstract properties and relations also contributes to tolerance in the face of image deformations.

6.4. Before and after MIRCs interpretation

The focus in this work is on modeling the full interpretation of minimal images on their own, but it is also of interest to consider briefly the broader process within which this interpretation process takes places. As described above, our model suggests a view in which the recognition and interpretation of a larger, natural image, with multiple objects and clutter, is a process that includes a bottom-up and a top-down stage. In our computational simulations, the bottom-up stage was modeled using existing CNN models (e.g. VGG-19, adjusted and fine-tuned to detect minimal images, same as in Section 4.3). These models can reliably recognize multiple objects, and locate them e.g., by bounding boxes (Girshick et al., 2014), or by approximate segmentation (Long et al., 2015). Such bottom-up models are not necessarily sufficient models for the early stages of human vision, but they produce adequate responses for our top-down interpretation stage. When trained for MIRCs recognition, their accuracy is well below human performance (Ullman et al., 2016), but they were sufficient in our simulations to identify candidate regions in the image for a given MIRC (Fig. 14A–E). When dealing with a full image, a natural next step will involve the selection of a MIRC (or a cluster of related MIRCs) by some attentional process, and applying the top-down stage. The outcome of the second stage is a detailed interpretation, combined with a validation of the proposed category for the region. To test such validation in the horse-head minimal image example, we have compared the precision-recall curve between a CNN-based classifier (very-deep CNN, Simonyan & Zisserman, 2015, trained as in Section 4.3) and the interpretation model. Our test set included 120 positive horse-head minimal image examples, and 200,000 negative examples taken from non-class images, similar to the ones used for training. We measured validation by the capability of interpretation to disambiguate 'hard negative' examples, namely non-class examples that received high CNN classification score. Since all positive examples were recognized by human observers, we set a threshold for CNN score that allows 100% recall, and applied interpretation to all positive and negative examples that passed this filter (all 120 positive example, 1943 negative examples). We compared the average precision for the CNN and interpretation on this set. The results shown in Fig. 14F, indicate improvement in MIRC recognition when applying Interpretation as a second stage following CNN detection.

Starting the process at the MIRCs level is potentially useful, because they are reliably recognizable on their own, and do not depend on additional supporting context to be recognized and interpreted (see Fig. 14E, as demonstration for a MIRC computation in a cluttered scene). We suggest that subsequent stages include at least two main components (and probably additional ones): integration and expansion. The integration process combines different MIRCs, in particular MIRCs that belong to the same object. An object in the image will typically be covered by multiple MIRCs at different locations and scales (Ullman et al., 2016), and their integration can lead to a robust and detailed representation of the object. This initial recognition and interpretation can next 'spread out' to surrounding regions, which are less recognizable on their own, but can use the context of the already-recognized MIRCs for disambiguation. The expansion can add information about the scene, for example, about the interaction of the object depicted in the MIRC with another object nearby (as e.g., in Fig. 14G–I).

6.5. Interpretation by network models

Recognition models based on deep convolutional networks have shown to produce high-accuracy results in object classification and promising results in related tasks, such as segmentation (e.g., Long et al., 2015). The current model combines network algorithms with

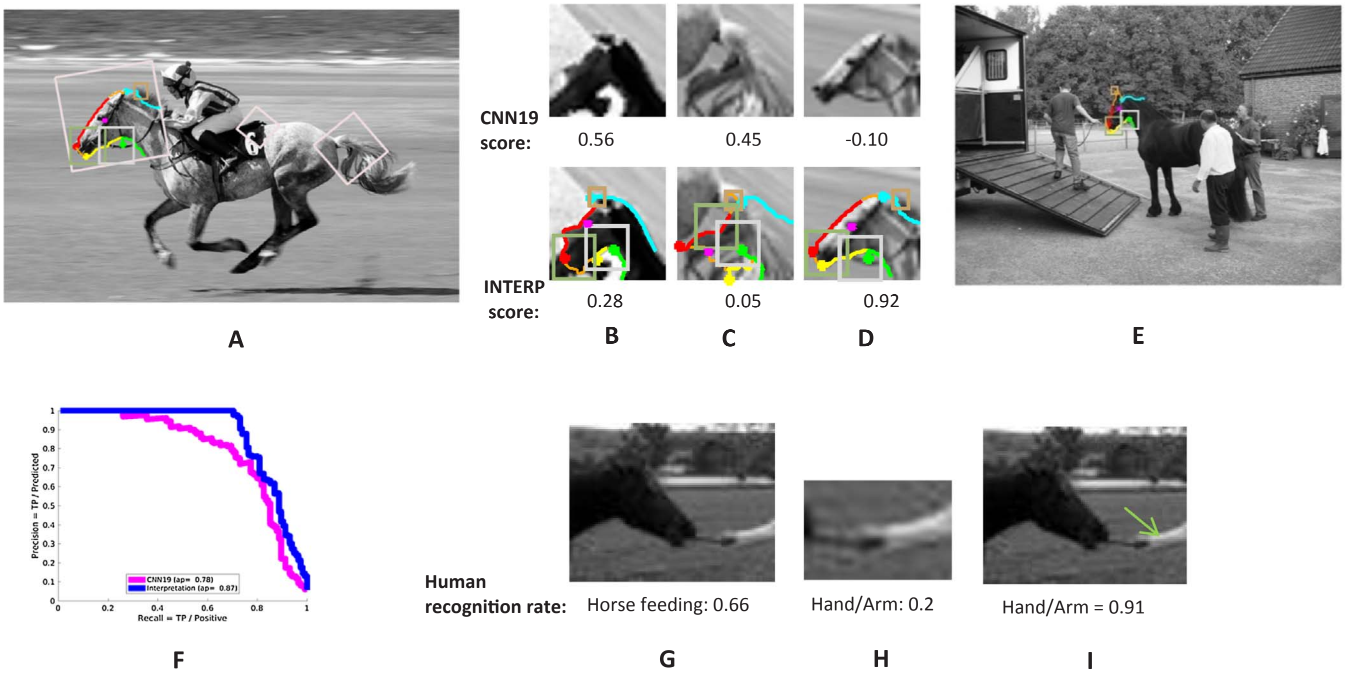

Fig. 14. Before and after Interpretation: Detection, Interpretation, Validation, and Expansion. **(A).Detection:** A bottom-up detector (here based on RCNN, details in Section 4.3) finds candidates for the horse-head MIRCs. Pink boxes denote top 3 candidates. **(B–D). Interpretation and Validation:** The interpretation applied to each candidate, to produce an interpretation combined with a confidence score. **(E).** Another example of a horse-head MIRC detection and interpretation in a cluttered scene. **(F).** Precision-recall curves for CNN19 alone (magenta, average precision (AP) = 0.78), and with interpretation confidence score (blue, AP = 0.87). Classification scores in (F) computed for the positive and negative horse-head minimal image examples in Section 5.1. **(G–I) Expansion:** The initial interpretation can next 'expand' to surrounding regions. In image region (G), slightly extended from the horse-head MIRC region, the expanded interpretation is to a region containing a portion of the human hand, which together with the horse-head is sufficient to recognize the human-horse interaction. In this image region humans can recognize the hand (I) and 'feeding' interaction (G), but the hand on its own is not recognized (H).

other methods to extract complex relations and identify the final structure. Similar combinations have been used recently by other models that extract complex structures (e.g. human pose, Chen & Yuille, 2014, combining CNN with a subsequent conditional random field stage; Lake, Salakhutdinov, & Tenenbaum, 2015, in the domain of written characters). We found that existing feed-forward network models have limited accuracy when applied to the interpretation of minimal images. Our evaluation trained a recent semantic segmentation network (Long et al., 2015) to identify interpretation components of minimal images. The accuracy of the resulting interpretation was closer to the 'basic' version of our interpretation model, compared with the full version of our model, which uses the extended set of relations (Section 5.1).

It is plausible, however, that extended network models, such as models using recurrence and memory, will cope more successfully with local interpretation. It will be of interest to develop such models in future work, and compare network structures that prove successful for local interpretation, as perceived by humans, and compare models with aspects of cortical circuitry in the visual system, e.g. in terms of using recurrent and feedback connectivity.

Acknowledgements

We thank Daniel Harari for sharing psychophysics data and for help with data collection, and anonymous reviewers for helpful comments and references. This work was supported by ERC Advanced Grant "Digital Baby", the EU's Horizon 2020 research and innovation program under grant agreement No. 720270, Israeli Science Foundation grant 320/16, and the Center for Brains, Minds and Machines, funded by NSF Science and Technology Centers Award CCF-1231216.

Appendix A. Psychophysics experimental methods

A.1. Labeling all semantic components in a minimal image

This experiment was used for identifying semantic elements, which humans can consistently identify in minimal images. Subjects ($n = 30$) were presented with a minimal image in which a red arrow pointed to a location in the image (e.g., the horse eye, or the center of the mouth region), and were asked to name the indicated location. Similarly, a contour was marked in red on the image, and subjects produced two labels for the two sides of the contours (e.g., tie and shirt). In both cases subjects were asked to also name the object they saw in the image (without the markings). To map the scope of 'full' human-level interpretation, we put the red arrows and contours at multiple image locations, and tested their consistent labeling. We considered a recognized component if more than 50% of human tags were consistent. Presentation time was unlimited, and the subjects responded by typing the labels. All experiments and procedures were approved by the institutional review boards of the Weizmann Institute of Science, Rehovot, Israel. All participants gave informed consent before starting the experiments.

A.2. Annotating point, contour, and region components in minimal image examples

Subjects (N = 2) were presented with examples of the semantic components found for a given minimal image by the experiment in Appendix A.1 (annotated by points, contours, and regions, as in Fig. 4B), and were asked to produce similar annotations in novel examples. Annotators were given partially overlapping sets of examples from each class, which together covered the complete training and testing sets. At least 50 examples from each class were annotated by two different subjects, and were used to test consistency in human annotations (see Table 2). The annotated images served as the 'ground truth' in evaluating the performance of the interpretation model (Section 5.1, and Table 2).

Appendix B. The learning model and procedure

B.1. A structured learning model based on random forest

The problem of local interpretation can be viewed as an instance of so-called 'structured learning' (e.g. Shalev-Shwartz & Ben-David, 2014). As described in Section 3.2, given a structure S_C consisting of a set of primitives P_C, and a vector R_C of relations between them, we wish to learn an interpretation function f_S that finds the structure S_C (denoted S below for simplicity) in an image I

$$f_S(I) = \pi$$

where I is the object image, and π is not just a class label, but a full assignment, which is in our case a mapping between components in the structure S and points, contours, and regions in the image I. We refer to π as an 'assignment', since it assigns to any primitive in the model S, a counterpart in the image, identified by π_i. π is then a vector $\pi = [\pi_1,\pi_2,...,\pi_N]$, where N is the number of primitives in the model S. For example, if the minimal image is the horse head, and the primitives set in S includes, among others, the horse eye (primitive index = 1, type = point), and the horse mane contour (primitive index = 5, type = contour), then, π_1 is a point in I assigned to the horse's eye, and π_5 is a contour in I assigned to the horse's mane.

It is common to express the function f_S using a (learnable) *scoring* function $g(I,\pi;w)$, which measures the compatibility between the model structure S, and the corresponding structure identified in the image. The additional variables w are parameters of the interpretation function, described below. $f_S(I)$ then takes the form:

$$f_s(I;w) = \underset{\pi}{\operatorname{argmax}}\{g(I,\pi;w)\}, \tag{1}$$

namely, given an image I (with parameters w already fixed), find the assignment π into I that has the highest compatibility with the model structure S. The goal of the function f_S is then to find the configuration of elements within the image I, which is as compatible as possible with the model structure S.

The function g in our interpretation measures the compatibility between properties and relations specified by the structure S of the model, and the same properties and relations computed for the corresponding image elements, identified by the assignment π. This compatibility is computed as follows. Given an assignment π of the model primitives P_C to the image I, we denote the results of measuring all the model relations R_C in the specific image I by the vector $\phi_S(\pi,I)$. Following the example in Section 3.2, position 3 in the vector $\phi_S(\pi,I)$ could be 'true' (or 1), indicating that primitive 5 is contained in primitive 7, and position 4 could be 0.9 indicating the degree of straightness for primitive 2.

The relations vector $\phi_S(I,\pi)$ is then used to measure the compatibility of the image structure with the model structure. This is obtained in our model by a random forest algorithm (Amit & Geman, 1997; Breiman, 2001), which is learned from training examples. A random forest is a non-linear model composed of a set of classification trees:

$$\{t_1,t_2,\ldots,t_j,\ldots\},$$

where t_j is the j-th tree in a forest. The parameters w in this model (in the definition of f_S and g) are the queries in the tree nodes, and a standard learning procedure for random forests (Breiman, 2001) is used to set these parameters based on training examples. Each tree is applied to the relations vector $\phi_S(I,\pi)$ to produce a decision whether the given assignment, represented by π, is consistent with a class structure or not (i.e., the relations vector $\phi_S(I,\pi)$ was classified as 1 or 0). Finally, the function g returns the average of all tree votes:

$$g(I,\pi,w) = \frac{1}{K}\sum_{j=1}^{K} t_j(\phi_S(I,\pi)), \tag{2}$$

where K is the number of trees in the forest. The assignment we seek is the one that maximizes the value of this expression, and the value of g for this assignment is the corresponding 'interpretation score'. An effective optimization search is described in Appendix B.2 below.

The random forest algorithm also provides a method for evaluating the individual contribution of each of the relations in the model to the learning process. This is obtained by removing a single relation in $\phi_S(I,\pi)$in all vectors in our data, and measuring the interpretation correctness (score) by the random forest with and without this relation. (Referred to as the 'Out of bag estimate' for strength of random forest features, Breiman, 2001). We used this method in Section 4 to derive a set of relations, which are useful for the interpretation process. 'Informative' relations in Section 4 are measured by the difference in the performance of the model (the interpretation score) with and without the relation in question.

B.2. Detecting primitive candidates and an effective optimization search

We describe below how we implemented the calculation of f_S (Eq. (2)), namely, derive the best assignment π for a given image I. Our implementation includes two stages: (i) finding k (k = 10) candidates for each primitive in S, and (ii) seeking the candidate combination that forms the best assignment. In more details, the two stages are

i. *Primitive candidates:* For primitives of type 'point' and 'region' we find candidates in a bottom-up manner: for 'point', we consider all pixels in the minimal image, and for 'region' we take all image windows of the region size in a 'sliding window' search. For type 'contour' we find the candidates in a top-down manner, as follows: We project ground truth annotated contours on an edge map (Arbelaez, Maire, Fowlkes, & Malik, 2011), to get edge contour fragments similar in their location and shape to the ground truth ones. We then used connected pairs of fragments (by the Kovesi edge linking toolbox, 2000) as candidates for the contour primitive. We rank all candidates of point, contour, and region types by their unary relations in R_C, and keep the top k for each primitive. Unary relations used for ranking include visual appearance of regions and contours (relations 4 and 5 in Table 1), and intensity minima/maxima of points (relation 2 in Table 1).

ii. *Finding the best assignment:* Given an image I, a trained model w, and a set of candidates for each primitive in P_C, we run over different configurations of candidates in a coordinate descent manner (Bertsekas, 1999). We start with a random configuration, and then optimize successively one candidate at a time. Specifically, the procedure is:

```
(1) Start with a random configuration of primitive candidates $\pi = [\pi_1,\pi_2,\ldots,\pi_i,\ldots,\pi_N]$.
(2) Repeat until g converges:
      For each primitive $i$, go over all candidates $\pi_i'$ and update:
        • $\pi' = [\pi_1,\pi_2,\ldots,\pi_i',\ldots,\pi_N]$.
        • $\pi \leftarrow argmax\{g(I,\pi,w),g(I,\pi',w)\}$
(3) Return $\pi$.
```

Such a procedure is guaranteed to converge to a local optimum (Bertsekas, 1999; a similar optimization search was used for Hopfield networks, Hopfield, 1982). Experimentally, because the search space in minimal images is limited due to small number of primitives, 3 initiations of the procedure were usually sufficient to get good convergence. The final assignment, together with contour grouping and bridging (Appendix C below), identify all the image components which correspond to the MIRC model.

Appendix C. Details of computing relation

Table 1 in Section 4.4 contains the extended set of relations used in our models. In this appendix, we add technical details about the computational procedures for computing the different relations. For all procedures described here, x, y represent the coordinates of the image plane. All procedures were implemented in MATLAB, code is available from the authors.

Containment: Given a pixel point $[x,y]$ and a set of pixels comprising a region R, we return true if the point is in the region, i.e., $[x,y] \in R$. R can be either a single region primitive, or a region bounded by two (or more) contour primitives.

Contour ends in a region: Given an end point pixel $[x_1^C,y_1^C]$ of a contour C, and a set of region pixels R, we return 'true' if the end point is in the region, i.e., $[x_1^C,y_1^C] \in R$.

Parallelism: Given two contours, C_a and C_b, we compute a binary mask M:

$$M(x,y) = 1 \quad if \quad [x,y] \in C_a \; or \; [x,y] \in C_b$$
$$M(x,y) = 0 \quad otherwise.$$

We then compute the distance transform map (Maurer, Qi, & Raghavan, 2003) for M, denoted $DT\{M\}$, followed by a non-maxima suppression to get the ridges R of *DT{M}*. The ridges R is the set of pixels that are at equal distance from both contours. The two contours are considered parallel if the variance of R is close to zero. We exclude cases where the size of R is small. We thus return 'true' if $Var[R] < \varepsilon$, where ε is a threshold close to zero (we chose empirically $\varepsilon = 0.2$).

Continuity of contours: Given a contour C_a with one of its endings: $[x_1^{C_a},y_1^{C_a}]$, and a contour C_b with one of its ending: $[x_1^{C_b},y_1^{C_b}]$, we estimate the local orientations at the endings, namely $\theta_1^{C_a}$ and $\theta_1^{C_b}$, and use them to compute the completion path between $[x_1^{C_a},y_1^{C_a},\theta_1^{C_a}]$ and $[x_1^{C_b},y_1^{C_b},\theta_1^{C_b}]$ (Ben-Yosef & Ben-Shahar, 2012). We consider 'good continuation' between the two contours if the completed path does not contain inflection points. We return 'true' if the number of inflection points in the path equals to zero.

Bridging contours: Given a contour C_a with one of its endings $[x_1^{C_a},y_1^{C_a}]$, a contour C_b with one of its endings $[x_1^{C_b},y_1^{C_b}]$, and the image I from which the two contours are extracted, we test for an image contour connecting them. We compute the UCM map (an edge map, Arbelaez et al., 2011) for I and define a graph $G = < V,E>$, where V is the set all pixels in the UCM map, namely

$$v_i \in V:\ \ UCM(v_i) > \tau,$$

τ is a UCM threshold ($\tau = 0.1$), and E is a set of weighted edges. An edge $e \in E$ is put for each pair of pixels in V that are immediate image neighbors. The weight of an edge $e = \{v_i,v_j\}$ is defined as the difference in UCM levels between pixels:

$$w(e) = UCM(v_j) - UCM(v_i)$$

(The graph G is computed in a pre-process stage.) We return the shortest weighted path in G (if exists) between $[x_1^{C_a},y_1^{C_a}]$ and $[x_1^{C_b},y_1^{C_b}]$.

The bridging procedure was also extended in two versions: (i) finding a path in G that is the most consistent with the ways contours C_a and C_b are connected in positive train images, and (ii) finding a path in G that is constrained to pass through region primitive.

Visual appearance inside regions or along contours: Given a candidate image region R_a for a primitive R in the model, we match the distribution of the visual appearance features in R_a and in the training examples of R. Visual appearance features were 'visual words' features (Arandjelovic & Zisserman, 2013), and deep neural network features (top layer of a fully convolutional network, Long et al., 2015). For a contour candidate, we used a similar match of visual appearance features, this time along a thin region surrounding the contour. The visual appearance relations were used in our scheme as follows: suppose that human interpretation psychophysics suggests a region element (e.g., the tie knot region in Fig. 4B) as one of the primitives in the interpretation model. We then produce descriptors for both the visual words and deep CNN features, which serve as potential unary relations for this element. These descriptions have been used successfully in past models, and we found empirically that both can be useful for the current task. We then evaluate which of these unary relations is more informative for interpretation and use it as a part of the MIRC's model.

Coherent visual appearance: Given two candidate image regions R_a and R_b, we match the distribution of the visual appearance features in these two regions. Visual appearance features were 'visual words' features (Arandjelovic & Zisserman, 2013; Vedaldi & Fulkerson, 2008), and deep neural network features (Long et al., 2015). R_a or R_b could be either a single region primitive, or a region bounded by two (or more) contour primitives.

Cover of a point by a contour: Given a pixel point $[x,y]$ and a contour C, we project C on the X-axis of the image plane, and return 'true' if x is within the range of projection. We composed procedures for different directions of cover, namely for a contour covers a point from top or from bottom. Similar 'cover' procedures were also for the Y axis.

Appendix D. Evaluating similarity between elements: points, contours, and regions

This process was used for evaluating the correctness of the interpretation produced by the model (Section 5.1). For two regions, A and B, the standard Jaccard measure ($|A \cap B|/|A \cup B|$, Tan et al., 2006) was used. For two points, we construct a small square region around each point (size of 12% of the minimal image), and then evaluate the Jaccard index of these regions. For two contours, we used a simple extension of the Jaccard index to contours, by extending the contours into tube shaped regions (tube width was 4% of the minimal image) and measure the Jaccard index between these regions.

Appendix E. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2017.10.006.

References

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation, 9*(7), 1545–1588.

Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3686–3693). .

Arandjelovic, R., & Zisserman, A. (2013). All about VLAD. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1578–1585). .

Arbelaez, P., Maire, M., Fowlkes, C., & Malik, J. (2011). Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(5), 898–916.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61*(3), 183–193.

Azizpour, H., & Laptev, I. (2012). Object detection using strongly-supervised deformable part models. *Proceedings of the European conference on computer vision,* 836–849.

Ben-Yosef, G., Assif, L., Harari, D., & Ullman, S. (2015). A model for full local image interpretation. *Proceedings of the annual meeting of the cognitive science society,* 220–225.

Ben-Yosef, G., & Ben-Shahar, O. (2012). A tangent bundle theory for visual curve completion. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(7), 1263–1280.

Bertsekas, D. P. (1999). *Nonlinear programming.* Belmont: Athena Scientific1–60.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*(2), 115–147.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Brooks, R. (1983). Model-based 3-D interpretations of 2-D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5*(2), 140–150.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L. (2017). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE transactions on pattern analysis and machine intelligence (PP(99), pp. 1–1), doi: http://dx.doi.org/10.1109/TPAMI.2017.2699184.

Chen, X., & Yuille, A. L. (2014). Articulated pose estimation by a graphical model with image dependent pairwise relations. *Advances in Neural Information Processing Systems,* 1736–1744.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. *Proceedings of the workshop on statistical learning in computer vision, European conference on computer vision, 1*(1), 1–2.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 886–893). .

Demirci, M. F., Shokoufandeh, A., Keselman, Y., Bretzner, L., & Dickinson, S. (2006). Object recognition as many-to-many feature matching. *International Journal of Computer Vision, 69*(2), 203–222.

Deng, J., Berg, A., Satheesh, S., Su, H., Khosla, A., & Fei-Fei, L. (2012). Imagenet large scale visual recognition competition 2012 (ILSVRC2012). net.org/challenges/LSVRC/2012/.

Denil, M., Bazzani, L., Larochelle, H., & de Freitas, N. (2012). Learning where to attend with deep architectures for image tracking. *Neural Computation, 24*(8), 2151–2184.

Elder, J. H., Krupnik, A., & Johnston, L. A. (2003). Contour grouping with prior models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(6), 661–674.

Elder, J., & Zucker, S. (1996). Computing contour closure (1996). *European Conference on Computer Vision,* 399–412.

Epshtein, B., Lifshitz, I., & Ullman, S. (2008). Image interpretation by a single bottom-up top-down cycle. *Proceedings of the National Academy of Sciences, 105*(38), 14298–14303.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The Pascal visual object classes (voc) challenge. *International Journal of Computer Vision, 88*(2), 303–338.

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(4), 594–611.

Feldman, J. (2007). Formation of visual "objects" in the early computation of spatial relations. *Perception & Psychophysics, 69*(5), 816–827.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(9), 1627–1645.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision, 61*(1), 55–79.

Ferrari, V., Jurie, F., & Schmid, C. (2010). From images to shape models for object detection. *International Journal of Computer Vision, 87*(3), 284–303.

Fidler, S., & Leonardis, A. (2007). Towards scalable representations of object categories: Learning a hierarchy of parts. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–8). .

Field, D. J., Hayes, A., & Hess, R. F. (1993). Contour integration by the human visual system: evidence for a local "association field". *Vision Research, 33*(2), 173–193.

Foster, D. H., Simmons, D. R., & Cook, M. J. (1993). The cue for contour-curvature discrimination. *Vision Research, 33*(3), 329–341.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587). .

Girshick, R., Iandola, F., Darrell, T., & Malik, J. (2015). Deformable part models are convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 437–446). .

Hanson, A., & Riseman, E. (1978). Visions: A computer vision system for interpreting scenes. In A. Hanson, & E. Riseman (Eds.). *Computer vision systems* (pp. 303–334). New York, NY: Academic Press.

He, K., Zhang, Z., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). .

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences, 11*(10), 428–434.

Jacobs, D. W. (1996). Robust and efficient detection of salient convex groups. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(1), 23–37.

Jin, Y., & Geman, S. (2006). Context and hierarchy in a probabilistic image model. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2145–2152). .

Joachims, T., Hofmann, T., Yue, Y., & Yu, C. N. (2009). Predicting structured objects with support vector machines. *Communications of the ACM, 52*(11), 97–104.

Jolicoeur, P., Ullman, S., & Mackay, M. (1986). Curve tracing: A possible basic operation in the perception of spatial relations. *Memory & Cognition, 14*(2), 129–140.

Kanizsa, G. (1979). Organization in vision: Essays on Gestalt perception. Praeger Publishers.

Kovesi, P. D. (2000). MATLAB and Octave functions for computer vision and image processing. http://www.csse.uwa.edu.au/~pk/Research/MatlabFns/#match.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems,* 1097–1105.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the International Conference on Machine Learning,* 282–289.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science, 350*(6266), 1332–1338.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, CL. (2014). Microsoft coco: Common objects in context. *Proceedings of the European conference on computer vision* (pp. 740–755). .

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440). .

Machilsen, B., Pauwels, M., & Wagemans, J. (2009). The role of vertical mirror symmetry in visual shape detection. *Journal of Vision, 9*(12) 11-11.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London B: Biological Sciences, 200*(1140), 269–294.

Maurer, C. R., Qi, R., & Raghavan, V. (2003). A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(2), 265–270.

Mnih, V., Heess, N., & Graves, A. (2014). Recurrent models of visual attention. *Advances in Neural Information Processing Systems,* 2204–2212.

Murphy, T. M., & Finkel, L. H. (2007). Shape representation by a network of V4-like cells. *Neural Networks, 20*(8), 851–867.

Ommer, B., & Buhmann, J. M. (2007, June). Learning the compositional nature of visual objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 48.

Opelt, A., Pinz, A., & Zisserman, A. (2006). Incremental learning of object detectors using a visual shape alphabet. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3–10). .

Palmer, S., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin & Review, 1*(1), 29–55.

Parent, P., & Zucker, S. W. (1989). Trace inference, curvature consistency, and curve detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11*(8), 823–839.

Pasupathy, A., & Connor, C. E. (1999). Responses to contour features in macaque area V4. *Journal of Neurophysiology, 82*(5), 2490–2502.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019–1025.

Rodríguez-Sánchez, A. J., & Tsotsos, J. K. (2011). The importance of intermediate representations for the modeling of 2d shape detection: Endstopping and curvature tuned computations. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4321–4326). .

Roelfsema, P. R., Lamme, V. A., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature, 395*(6700), 376–381.

Saarela, T. P., Sayim, B., Westheimer, G., & Herzog, M. H. (2009). Global stimulus configuration modulates crowding. *Journal of Vision, 9*(2), 1–11.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences, 104*(15), 6424–6429.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Siddiqi, K., Shokoufandeh, A., Dickinson, S. J., & Zucker, S. W. (1999). Shock graphs and shape matching. *International Journal of Computer Vision, 35*(1), 13–32.

Simonyan, K., Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International conference on learning representations*.

Stahl, J. S., & Wang, S. (2008). Globally optimal grouping for symmetric closed boundaries by combining boundary and region information. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(3), 395–411.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining, Vol. 1*. Boston: Pearson Addison Wesley.

Todorovic, S., & Ahuja, N. (2006). Extracting subimages of an unknown category from a set of images. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 927–934). .

Tompson, J. J., Jain, A., LeCun, Y., & Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in Neural Information Processing Systems,* 1799–1807.

Torralba, A. (2009). How many pixels make an image? *Visual Neuroscience, 26*(01), 123–131.

Ullman, S. (1984). Visual routines. *Cognition, 18*(1–3), 97–159.

Ullman, S., Assif, L., Fetaya, E., & Harari, D. (2016). Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences, 113*(10), 2744–2749.

Vedaldi, A., Fulkerson, B. (2008). VLFeat: An open and portable library of computer vision algorithms. < http://www.vlfeat.org/ > .

Vedaldi, A., Mahendran, S., Tsogkas, S., Maji, S., Girshick, R., Kannala, J., ... Taskar, B. (2014). Understanding objects in detail with fine-grained attributes. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3622–3629). .

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of the National Academy of Sciences, 108*(23), 9661–9666.

Wei, S. E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4724–4732). .

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. *II. Psychological Research, 4*(1), 301–350.

Westheimer, G., Crist, R. E., Gorski, L., & Gilbert, C. D. (2001). Configuration specificity in bisection acuity. *Vision Research, 41*(9), 1133–1138.

Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., & Kassim, A. (2016). Robust facial landmark detection via recurrent attentive-refinement networks. *European conference on computer vision* (pp. 57–72). .

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, 111*(23), 8619–8624.

Yang, S., Luo, P., Loy, C. C., & Tang, X. (2015). From facial parts responses to face detection: A deep learning approach. *Proceedings of the IEEE international conference on computer vision* (pp. 3676–3684). .

Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L., & Fei-Fei, L. (2011, November). Human action recognition by learning bases of action attributes and parts. In *Proceedings of the IEEE international conference on computer vision*, pp. 1331–1338.

Zhu, S. C., & Mumford, D. (2007). A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision*, 2(4), 259–362.

Zhu, L., Chen, Y., & Yuille, A. (2009). Unsupervised learning of probabilistic grammar-markov models for object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1), 114–128.