Title:

Image interpretation above and below the object level

Authors name and affiliation:

Guy Ben-Yosef[1,2,3] and Shimon Ullman[2,3]

1. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
2. Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 7610001, Israel
3. Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

Abstract:

Computational models of vision have advanced in recent years at a rapid rate, rivaling in some areas human-level performance. Much of the progress to date has focused on analyzing the visual scene at the object level – the recognition and localization of objects in the scene. Human understanding of images reaches a richer and deeper image understanding both 'below' the object level, such as identifying and localizing object parts and sub-parts, as well as 'above' the object levels, such as identifying object relations, and agents with their actions and interactions. In both cases, understanding depends on recovering meaningful structures in the image, their components, properties, and inter-relations, a process referred here as 'image interpretation'.

In this paper we describe recent directions, based on human and computer vision studies, towards human-like image interpretation, beyond the reach of current schemes, both below the object level, as well as some aspects of image interpretation at the level of meaningful configurations beyond the recognition of individual objects, in particular, interactions between two people in close contact. In both cases the recognition process depends on the detailed interpretation of so-called 'minimal images', and at both levels recognition depends on combining 'bottom-up' processing, proceeding from low to higher levels of a processing hierarchy, together with 'top-down' processing, proceeding from high to lower levels stages of visual analysis.

Keywords:

Visual recognition, Visual interpretation, Social Interactions, Interaction Recognition, Minimal images

1. Introduction

Substantial progress has been made in recent years in visual recognition, mainly at the level of recognizing individual objects. However, image understanding goes beyond individual objects, and requires understanding both below and above the object level. Below the level of individual objects, image understanding requires the recognition of object parts and fine-level details, which may be impossible to recognize on their own, such as a door handle in a full car image, or a belt-buckle in a full person's image. Above the object level, image understanding includes dealing with complex configurations, and interactions between objects, including interactions between agents (e.g., 'hugging') or between an agent and an object (e.g., 'playing the violin'). Common to scene understanding both above and below object recognition is the use of semantic structural representation, including relations between internal parts of a single object (e.g., Fig. 1B), as well as relations between multiple objects in a complex

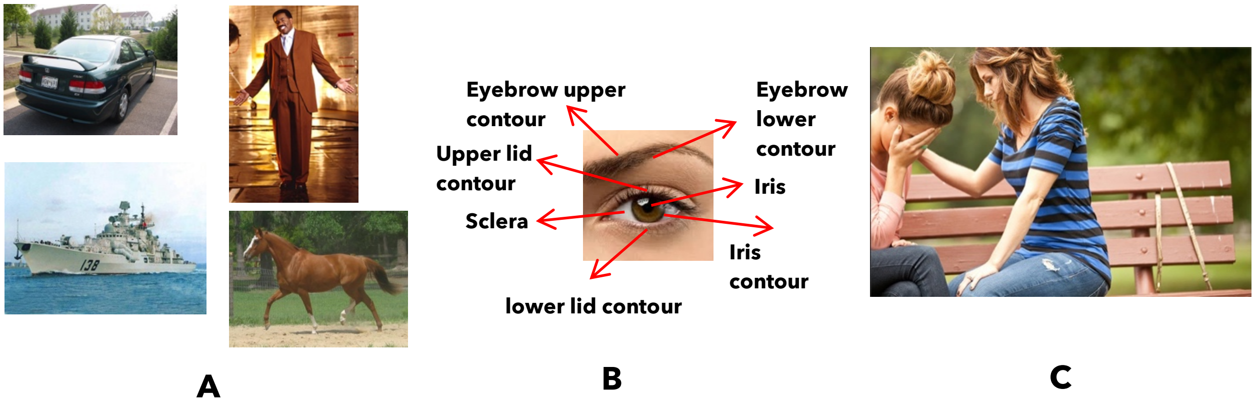

Figure 1. Below and above object recognition. **(A). Object recognition:** labeling images by basic level object classes (Rosch et al., 1976), e.g. cars, people, horses, and ships. Current computational models deal successfully with this level of recognition (e.g., the ImageNet data set, Deng et al., 2012). **(B). Image understanding below object recognition:** In addition to the object label, humans can identify a large number of semantic features and parts within an object image. In the image of a person's face, local features such as eyebrow, lid, or iris, are identified by humans, among many others. **(C). Image understanding above object recognition:** Humans can identify object configurations and interactions between objects and agents. In this image, humans can identify objects (two human bodies, a bench) and their parts, but also understand aspects of the configuration, e.g. an interaction of consolation.

scene (Fig. 1C). As will be discussed below in sections 2,3, such structural representation is a fundamental aspect of human visual understanding at all levels.

The process of acquiring semantic structure from raw sensory input (pixels) is termed here 'image interpretation', and it involves a mapping between image pixels to familiar components and relations in our world. Semantic components of interest in the world may be small details such as a crease in a shirt or a thin ring on a finger, or complex multi-object configurations such as an orchestra or a chessboard, and the interpretation process needs to span all levels. The term 'image understanding' as used here depends on the image interpretation process, but it can be more abstract, in the sense of using concepts which go beyond components of the physical world and relations between them, for example, goals, moods, judgments such as 'dangerous' and others. In Fig. 1C, for instance, the interpretation process can identify certain image structures as corresponding to human bodies, or parts such as face of fingers. It can also identify relations between body parts, such as 'touching' 'covering face', and the like. Image understanding will depend on results of this interpretation process, but will include higher, more abstract aspects, such inferring the 'consolation' interaction in the image. In the next sections, we describe our recent modeling studies of human interpretation processes, which are below (Sec. 2) and above (Sec. 3) object recognition. We conclude in Sec. 4 by discussing approaches and future directions in the study of human understanding of complex scenes.

2. Image understanding below object recognition

When looking at an object image, humans can identify not only the object label (or class), but also a set of semantic features and relations corresponding to the object's internal parts (e.g., as in Fig. 1B). This capability of humans is a part of image understanding below the object level, and the process of finding the parts and relations from pixels is called here 'full object interpretation'. This local level of image interpretation is discussed below in the context of so-called 'minimal images'.

The process of full object interpretation is difficult to replicate in computational models, since an object may contain a large number of identifiable components in highly variable configurations. We approach the modeling of this process 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 limited 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 (Ullman et al., 2016), but it becomes more tractable, since the number of semantic recognizable components is highly reduced, making effective interpretation more feasible (Ben-Yosef et al., 2018). 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. 2B, placing a limit on the locality of the interpretation process.

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. Finally, as discussed further below, 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. Based on these considerations, we present in the next section a model for local image interpretation, which is applied to local regions that are small, yet interpretable on their own by human observers.

In performing local interpretation, a question that naturally arises is: how should an object image be best divided into local regions? The approach we take in our studies 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 study of minimal recognizable images (Ullman et al., 2016). We briefly describe below how these minimal images were obtained, and then describe a model for their interpretation.

2.1 Minimal recognizable and interpretable configurations

A minimal image (also termed Minimal Recognizable Configuration, or MIRC) is defined below 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 by either cropping at one corner, or reducing resolution of the original patch. A recognizable patch is identified as a minimal image if none of its 5 descendants reach recognition criterion (50%). The process is illustrated in Fig. 2A. A search started with images from different object classes, 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 on the top row of Fig. 2B.

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. 2B, bottom row, together with their respective recognition rates. The small changes between minimal vs. sub-minimal configurations that cause large drop in recognition are used below to identify features and relations used in the interpretation model. Ullman et al. (2016) 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 et al., 2007) and recent deep network models (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015). As was shown in Ben-Yosef et al. (2018), the full interpretation model can provide at least a partial explanation to this sharp drop in recognition.

Minimal configurations are minimal in the sense that when further reduced, humans can no longer recognize them. Still, when humans recognize minimal configurations, they can also identify internal

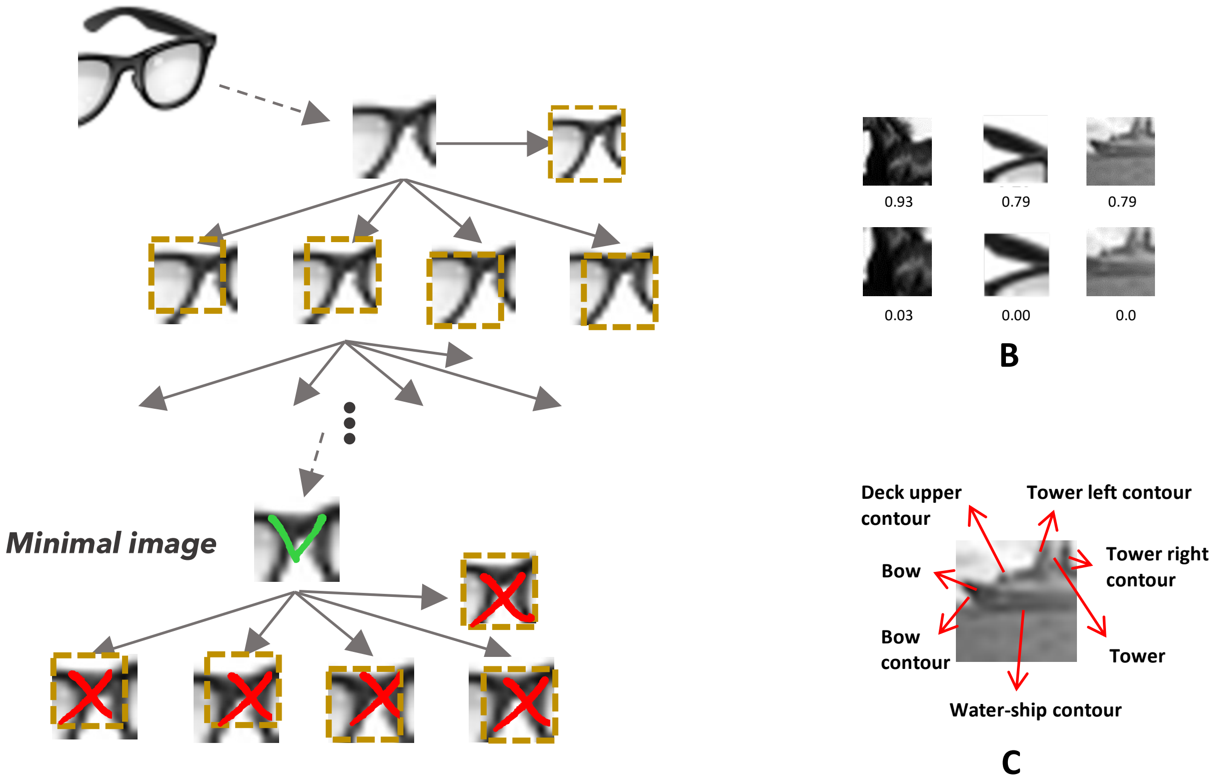

Figure 2. Minimal recognizable and interpretable configurations. **(A).** A search for minimal images started from a fully-viewed object image, which was gradually reduced in small steps, by slightly cropping corners or slightly reducing resolution. A minimal image is an image region that is recognizable on its own (green check mark), but is no longer recognizable when reduced further (red X mark). **(B).** Minimal (Top row), and their slightly reduced sub-minimal versions (bottom row) images. 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. (Adapted from Ullman et al., 2016). **(C).** Full human interpretation of minimal images. Arrows point to the parts and features that humans can reliably identify in a minimal image. (Adapted from Ben-Yosef et al., 2018).

parts and components in them (Ullman et al., 2016). Further tests (Ben-Yosef et al., 2018) have shown that the number of recognizable parts in minimal images is small (example in Fig. 2C), and that humans can consistently identify internal components in a large set of tested minimal configurations. Naturally, humans cannot identify any of the internal parts in the slightly reduced, but non-recognizable sub-minimal configurations. These results provide an empirical indication that in the human visual system recognition and interpretation go hand in hand, and that recognition is combined with the understanding of internal structures.

2.2 Object interpretation in related work

Image object interpretation 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 recognition, 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, in the Recognition by Components (RBC) model of human object categorization (Biederman, 1987), objects are represented in terms of a small number of 3-D major parts. A leading biological model on 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.

A model for human image interpretation (Epshtein et al., 2008) was shown to provide partial 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 model below 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.

In computer vision, there has been rapid progress in different aspects of object and scene recognition, based primarily on deep convolutional neural networks and related methods (Hinton, 2007; LeCun et al., 2015; Yamins et al., 2014; Krizhevsky et al., 2012). 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., Long et al., 2015; Chen et al., 2017) can identify image regions belonging to different objects in the PASCAL (Everingham et al., 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 et al., 2010; Girshick et al., 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., 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., Chen & Yuille, 2014; Cao et al., 2017) and within the human face (e.g., Yang et al., 2015).

The goal of interpretation models, such as those above, is to produce the semantic structure in an image region. The model is usually constructed during learning by supplying 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 properties of single elements (e.g., the curvature, location, or size of a contour), which can be considered as unary relations.

A number of algorithms have been developed and used in the field of machine vision under the general term 'structured prediction' to deal with problems related to the learning and discovery of image structures, such as Conditional Random Field (Lafferty et al., 2001), or Structured Support Vector Machine (Joachims et al., 2009). These models are given the set of possible relations to use, and then learn the specific parameters from examples. In terms of properties and relations, in most visual models that deal with image structures, such as the ones above, part properties (unary relations) are limited to local, deep CNN-based features, and binary relations are limited to relative displacements of components (parts or keypoints). As elaborated below, results of the present modeling show that the capacity to provide full interpretations requires the use of features and relations, which go beyond those used in most current recognition models.

2.3 A model for full interpretation of minimal object images

To study the process of human object interpretation, a model for full interpretation of minimal images was developed in (Ben-Yosef et al. 2015;2018), and below we briefly describe the design and results of this model. The interpretation scheme has two main components: in the learning stage, it learns

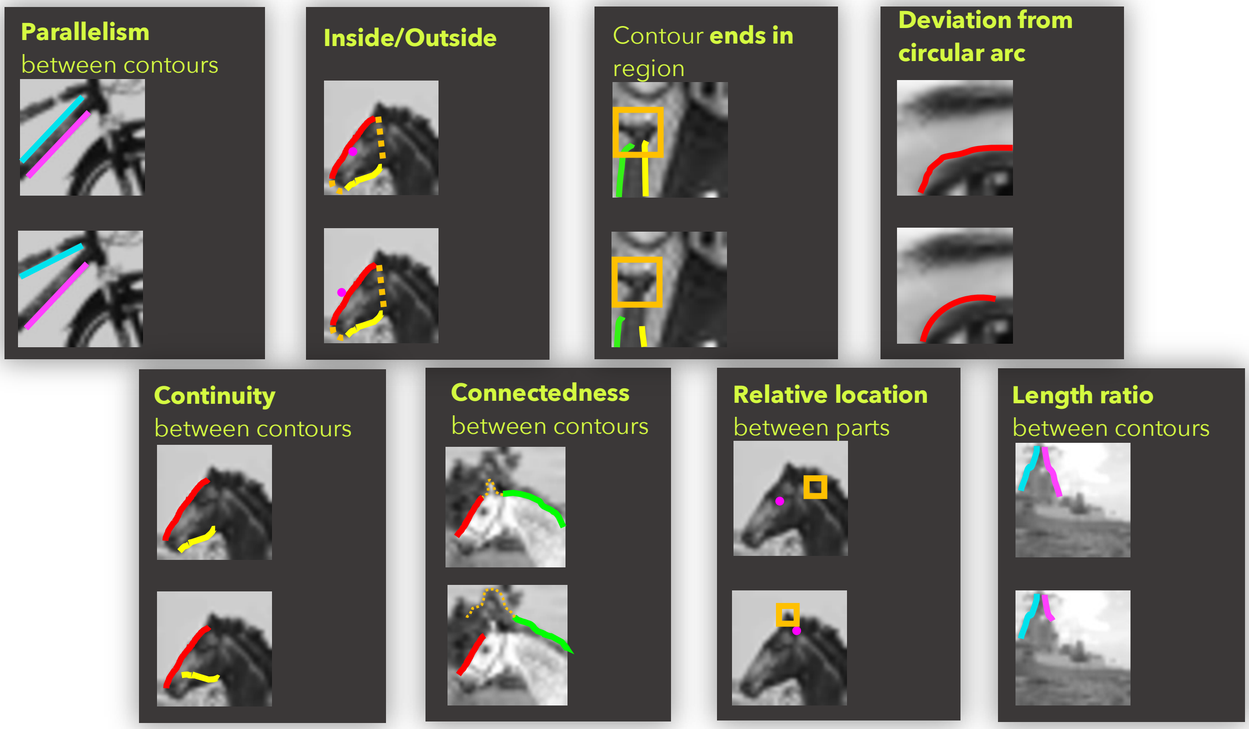

Figure 3. Useful features and relations for full interpretation. Each panel shows a relation when exists (top) and when not exist (bottom). Further details about the way they were inferred and implemented are in Ben-Yosef et al., 2018.

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.

The semantic features to be identified by the model (e.g., ‘ear’, ‘tie knot’, etc.) were features that human observers label consistently in minimal images, identified by using an MTurk procedure (the average number of consistently identified elements within a single minimal image was 8). The semantic features were then represented by three types of visual primitives: points (e.g., a horse eye), one-dimensional contours (for borders, e.g., a tie border), and two-dimensional region primitives. Given these semantic elements, we prepared a set of annotated images, in which the semantic components were marked manually on multiple examples of the minimal image, and then used in a structured learning framework based on a random forest classifier. The learning scheme computes a set of relations between elements in the structure for both positive and negative examples, and then learns the contribution of each relation to the identification of valid interpretations. A critical component in this scheme is therefore the types of relations that were used to identify correct local structures.

At inference time, the interpretation process starts with a candidate image region and its proposed category (e.g., that it contains a horse-head). The process then used the learned model of the region’s internal 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 the local primitives, including points, contours, and region parts in the image, to serve as potential candidates for the different components of the expected structure. The second stage searches for a configuration of the components that best matches the learned structure.

2.4 Structural representations for full object interpretation

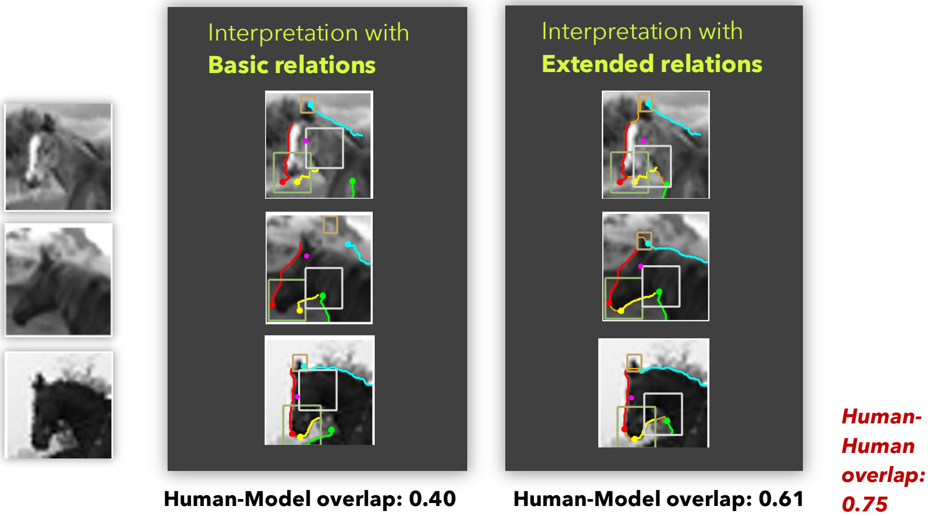
Interpretation with
Basic relations
Interpretation with
Extended relations
Human-Human overlap: 0.75
Human-Model overlap: 0.40
Human-Model overlap: 0.61

Figure 4. Evaluating predictions of the interpretation model. **Left column:** novel examples of minimal images of the horse-head type. On these examples, we show the predictions of two versions of our interpretation model, and compare between them. **Middle column:** a version with the 'basic set' of relations. **Right column:** a version with an 'extended set' of relations, which includes connectivity, continuity, inside/outside, 'Ends in', and others. The average overlap between the predicted interpretation and true human interpretation (shown at the bottom for each version), shows a significant improvement when the interpretation includes the extended set over basic set, but still a significant gap between the interpretation model and human interpretation.

The model described above belongs to the general approach of structured vision models. There is a rich history to the use of structural models in the computational study of vision, including visual recognition and interpretation (e.g., Felzenszwalb et al. 2010; Ferrari et al., 2010; Zhu & Mumford, 2007; Chen et al., 2017; Arnab et al., 2016). Models differ in the shape components used to create structured configurations, the relations used to represent configurations (including attributes of a single element as unary relations), and the algorithms used to learn structures from image examples, and to identify similar structure in novel images. The relations used in these models were mostly simple, including unary features of part resemblance (based on CNN features), and binary features of relative displacement (e.g., Felzenszwalb et al. 2010; Chen et al., 2017). We term these features and relations the 'basic relations'. As described below, our modeling showed that the usual set of basic relations is insufficient for interpretation, and minimal images were used to identify additional relations, which contribute to correct interpretation.

In the human and primate vision literature there has also been an extensive body of work on relations between elements in the visual field. These studies have shown sensitivity of the visual system to known principles of perceptual organization such as proximity, similarity, connectivity, symmetry and continuity between visual elements, and also to parallelism, curvature, convexity, co-linearity, co-circularity, connectedness of contours, and inclusion between elements (see review in Ben-Yosef et al., 2018). The availability of minimal images (sec. 2.1) allowed us to examine whether local appearance and basic relations are sufficient for producing an accurate 'full' interpretation by our 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 (e.g., the basic relations above) when applied to the interpretation of minimal images, and compare to interpretation produced by a model with a richer set of relations.

In the recognition of minimal images, the sharp drop in human's ability to recognize and interpret a minimal configuration when the image is slightly reduced, provided a tool for identifying useful relations for modeling human interpretation. A minimal image was compared with its similar, but unrecognizable sub-image, to identify either a missing component (e.g., a contour part) or a relation (e.g., between two contours parts), which were present in the minimal image but not in the sub-minimal configuration. For each candidate component or relation, we tested its consistent effect on other pairs of minimal and sub-minimal images, and we evaluated its statistical contribution to the interpretation process, by adding it to the set of relations, training a new interpretation algorithm, and measuring the difference in interpretation performance, with and without this relation. A list and illustrations of the most contributive relations are in Fig. 3 (hereinafter, the 'extended' set of relations), further details in (Ben-Yosef et al., 2018).

Interpretations produced by the model were compared with the ground truth annotations supplied by human annotators. To assess the role of the extended relations derived from minimal and sub-minimal pairs, 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 (namely, the basic relations and the relations in Fig. 3). Fig. 4 shows examples of the interpretations produced by the model with the basic and extended sets 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 120 examples of minimal images or more (480 examples for the horse-head minimal image). We automatically matched the ground truth annotated primitives to the interpretation output by the so-called overlap index (Intersection over Union of two regions. See Tan et al., 2006). Our results show a significant improvement in interpretation results when using the extended set of relations, but still a significant gap between the model and human interpretation. As an example, for the horse-head model in Fig. 4, the average overlap was 0.40 for the basic set, 0.61 for the extended relations set, and 0.75 overlap between different two human annotators (which served as an upper bound for comparing interpretations).

2.5 A two-stream view for recognition and interpretation in the human visual system

So far we presented a model for the interpretation of minimal images, and discussed the types of features that it uses and the type of predictions that it can provide. In this section, we discuss how the interpretation model can be used to predict human recognition, including the sharp drop in recognition at the minimal image level. As a first step, we tested a baseline recognition model based on deep convolutional networks, including multi and binary classification networks, which were pre-trained on ImageNet but fine-tuned for recognition of minimal images (Ullman et al., 2016; Ben-Yosef et al., 2018). For example, a binary classification network in Ben-Yosef et al. (2018) was trained to recognize a horse-head minimal image, based on the VGG19 network model (Simoniyan and Zisserman, 2015). It was trained on 120 minimal image examples of a horse-head (the positive train set), and 200,000 examples from non-horse images of the same size as the horse-head minimal image (the negative train set). Experimental results showed that the network was unable to replicate human behavior in two aspects: (i) it was often confused by examples that look similar to the horse-head, but were not confusable for humans (termed 'hard negatives'), and (ii) it could not predict the sharp gap in human recognition between the minimal and sub-minimal images.

Next, we examined whether the interpretation model described above can explain the sharp drop in recognition between minimal images and similar, but slightly reduced sub-minimal images. The results regarding human interpretation of minimal images suggest that humans combine the recognition of local image regions with the interpretation of their internal structure (Sec. 2.1). As a result, a false detection by the recognition model can be rejected if it does not contain the internal structure expected by the interpretation process. We therefore tested whether an integrated scheme, which combines recognition and interpretation, will also exhibit the sharp transitions found in human recognition. In the combined scheme, we used the confidence score provided by the interpretation model, for making the final recognition decision. In this manner, high-confidence interpretation, for instance in interpreting a horse-head image, is required for a positive recognition decision.

The interpretation model was trained with the same training set used for the VGG19 binary classifier (in the interpretation model, the positive examples were also annotated with the different parts). The match to human recognition on novel examples was then compared between the two schemes: with and without the interpretation model. The comparison showed that the recognition results using the interpretation model are much closer to human behavior on the set of confusable examples (i.e., hard negatives; see more details in Ben-Yosef et al., 2018). Furthermore, the interpretation model could predict the human recognition gap between minimal and sub-minimal images, and replicated the sharp drop in recognition when minimal images are reduced. The reason for the sharp drop is likely to be that even a small reduction of a minimal images can cause components of the internal structure (e.g., a horse's ear), as well as some pairwise relations, to be disrupted. The conclusion from these experiments is that the interpretation features discussed in Sec. 2.4 are not only useful to predict human interpretation, but also to predict human recognition of minimal images.

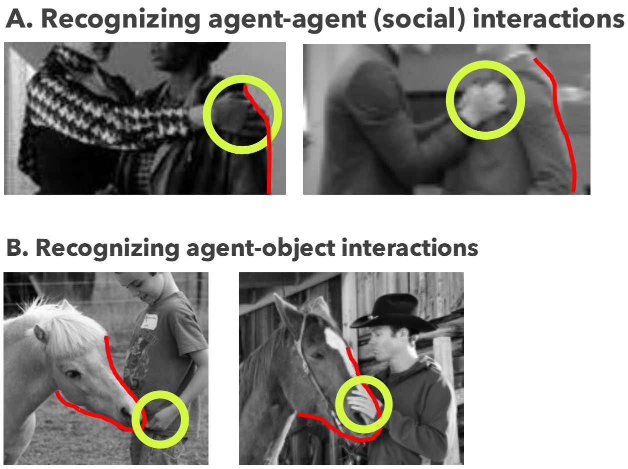
A. Recognizing agent-agent (social) interactions
B. Recognizing agent-object interactions

Figure 5. The role of interpretation in recognizing social interactions. (A). The location and shape of the hand of one person touching is essential to distinguish between recognition of a friendly or more aggressive type of social interaction. (B). The exact location of the hand relative to the horse head contours and parts is critical to judge if the interaction is 'feeding a horse' or 'petting a horse'.

The results and conclusions above lead us to suggest a two-stage view for recognition and interpretation in the human visual system. The first stage is based on a hierarchical feed-forward process in the visual ventral pathways, which may be roughly similar to the way that existing deep convolutional networks are operating (Reisenhuber & Poggio, 2001; Hinton, 2007; Yamins & Dicarlo, 2016). Results of the first stage then trigger a second stage, which performs the full interpretation process. The computational model suggests that the interpretation task relies on more complex and higher-order features compared with the first stage, to achieve fine localization of internal parts as well as their inter-relations. Computations performed by the second stage include in the model relations such as connectedness and enclosure. In the model, these computations are applied selectively at relevant image locations, rather than the parallel processing across the image used by convolutional deep networks. We suggest that in human vision this stage is likely to be applied at least in part by top-down processes, where object models stored in higher-level areas direct the application of the required processes to selected image locations. On this view, human object recognition is followed by an interpretation process at selected locations, and the interpretation process is also used to resolve ambiguous examples and reject false detections by the initial stage.

3. Recognition above the object level

The tasks of visual recognition and image understanding extend 'above' the object level, to include meaningful configurations of objects, agents, and their interactions. In this section, we discuss some aspects of this complex task. In particular, we focus on the problem of recognizing different types of interactions between two objects, two agents, or an agent and an object. Examples of agent-object interactions are transitive actions such as 'holding a book', 'playing the violin', or 'smoking a cigarette'. Examples of agent-agent interactions we consider include 'hugging', 'shaking hands' or 'helping'. The interaction of 'stealing' is an example for a configuration involving both agents and an object. Humans can not only understand the type of interactions from images, but also their tone and manner. For example, humans can tell if a violin player is holding the instrument correctly or not, or if two people are having a warm or a more formal hug (Fig. 6B), and the like.

Meaningful configurations can include complex interactions, involving multiple objects and agents, but the focus here will be on pairwise agent-object and in particular agent-agent interactions. The recognition of the type and tone of such interactions often depends on detailed analysis of subtle cues, in particular at the locations of contact between the interacting agents and objects. The fine localization of parts within the interacting objects, and the understanding of relations between these parts, are critical to judge the nature of the interactions. For example, the difference between the social interaction images in Fig. 5A depend on details of the shape and contact between agents and objects. As another example, a hand placed on a horse's mouth can tell us that the interaction is 'feeding a horse' (Fig. 5B). However, if the hand is placed slightly above the mouth, then we are more likely to understand the interaction as 'petting a horse'. As these examples illustrate, recognizing the type and tone of interactions often depends on a detailed interpretation of the participating agents and objects, with focus on the locations of contact between them. Detailed local interpretation discussed in the previous section is therefore also a key element for understanding interactions between objects and agents. In the sections below, we focus therefore on the use of a detailed local interpretation in the recognition of interactions. We further chose to focus in particular on social interactions, for several reasons: understanding social interactions from visual input is an important cognitive task, it is highly challenging from a modeling standpoint, and the ability to perform this interpretation task automatically is at present severely restricted.

In a recent study, we have started to develop parts for a computational model for interpreting social interactions between agents. In approaching the problem, we used a similar approach to that of Sec. 2.2, namely, focusing on the minimal interaction configurations and their interpretation in terms of parts and relations. The approach and results are discussed in the next sections. We begin in section 3.1 with a brief list of related computational work. Section 3.2 describes our psychophysical data and computational models for the interpretation of social interaction images. Section 4 discusses the relevance of the proposed framework using detailed local interpretation to the understanding of interactions in full-scale real-world images.

3.1 Image understanding above object recognition in recent computational work

Recognizing interactions between objects and agents is an active research area in current computer vision. As in the recognition of objects (e.g., ImageNet by Deng et al., 2012), the dominant approaches for recognizing interactions are based on training with 'big-data', and an effort to collect large datasets of interaction images and videos is currently under way (e.g., Stanford40 by Yao et al., 2011; HICO by Chao et al., 2015; Visual Genome by Krishna et al., 2017; the Kinetics dataset by Kay et al., 2017; AVA by Gu et al., 2018). Identifying interactions is also a major component of related computer vision challenges, such as image captioning (Vinyals et al., 2017) and visual question answering (Antol et al., 2015), which are based on combined text and visual data (e.g., Antol et al., 2015; COCO-VA by Lin et al., 2014; Visual7w by Zhu et al., 2016). Despite these efforts, performance of existing algorithms for interaction recognition is low (Gu et la., 2018) and significantly behind the performance of algorithms for object recognition. Specifically, current machine understanding of interactions between multiple objects in a complex scene -- even when these scenes are highly constrained -- is limited (e.g., Li et al., 2009; Johnson et al., 2017). We next turn to briefly review related work in our specific area of focus, namely, modeling the recognition of social interactions in images.

3.1.1 Visual understanding of social interactions in previous work

Early research on visual understanding of social interactions is rooted in the field of social and psychological sciences, studying the different types (e.g., Leary's circumplex, Leary, 1958), and physical characteristics (e.g., the distance between two individuals, Hall, 1966) of social relations. This analysis of interactions includes so-called proxemics (spatial aspects of interacting humans), and interaction taxonomies such as the Wiggins circumplex (Wiggins 1979), which applied Leary's Circumplex to the interaction and relations domain. The field of social interactions has also included developmental studies on infants and children. For example, Hamlin and Wynn (2011) have demonstrated recognition of social

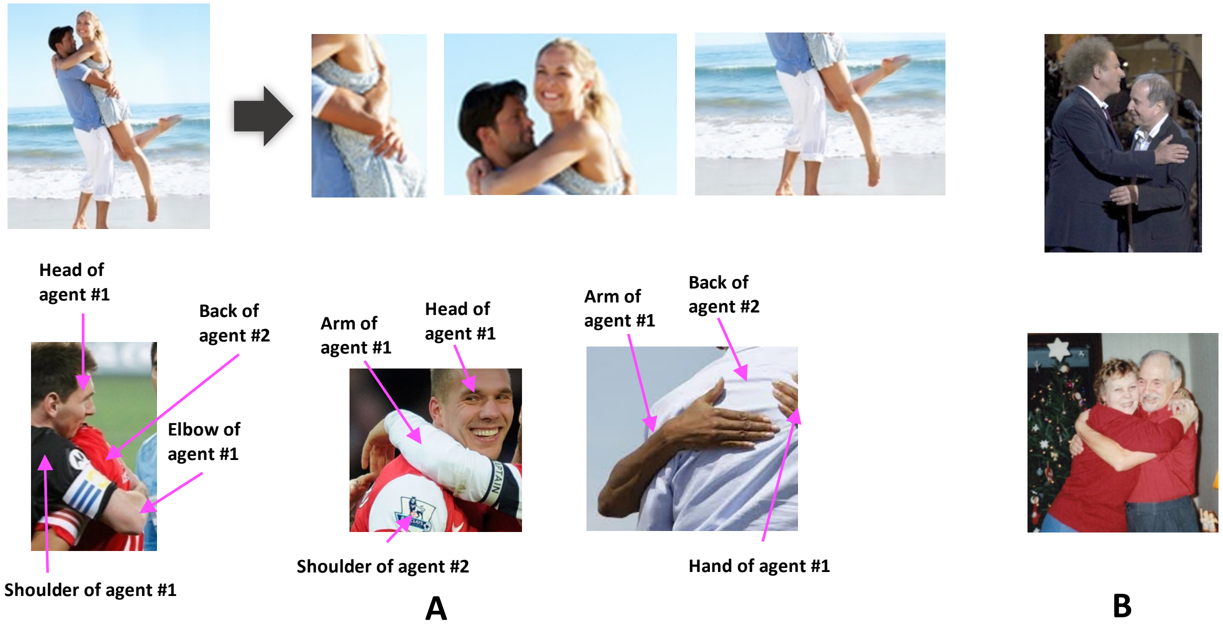

Figure 6. **(A).** ‘hugging’ can be recognized independently from local body parts of the interacting agents. **Top row:** Different image regions taken from the same scene, each of which is sufficient for humans to recognize a ‘hug’. **Bottom row:** Local image regions from different scenes in which humans can recognize a ‘hug’. Each region contains different sub-configurations of body parts. In each of these sub-configurations humans can also identify and localize the different body parts, and their relations, the process described in this paper as ‘image interpretation’. (B). Different tone of interactions. On the top panel humans report a cold formal hug, while on the bottom panel humans report an intimate warm hug. Detailed interpretation can play critical role in determining the type and tone of social interactions.

interactions in infants around the age of six months, such as a preference by the infants towards an agent seen helping another agent, over an agent hindering others. Other studies have shown that a perception of social dominance starts developing during the first year of life (Mascaro and Csibra 2012, Thomsen et al 2011). Such studies underscore the basic importance and the natural capacities of recognizing the type and tone of social interactions between people.

Brain studies have further highlighted the role of visual understanding of social interactions in the primate cortex. For instance, activations in human brain regions (in the posterior superior temporal sulcus, pSTS) were reported when subjects viewed interacting humans, but not when viewing non-interacting humans, for stimuli composed from moving point-light representing human figures (e.g., Centales et al., 2011; Isik et al., 2017). Similarly, a recent study in macaques found regions of the frontal and parietal cortex that responded exclusively to movies of monkey engaged in social interactions, but not to movies of monkeys conducting independent actions or of interactions between inanimate objects (Sliwa and Freiwald, 2017). These studies reveal the existence of cortical machinery that is dedicated to visual analysis of social interactions.

In terms of computational modeling coming from cognitive studies and machine vision, only a limited number of studies have addressed the problem of visual recognition of social interactions. Most of these studies relied primarily on spatiotemporal patterns in video sequences, unlike humans who can also reliably perceive social interactions in still images. Early methods for recognizing interactions were based on characterizing low-level visual features in interaction videos (e.g., Patron-Perez et al., 2012). More recent methods are based on finding body parts and modeling relations between the agents. Examples include localization of agents’ body pose (e.g., Yang et al., 2012), or face pose (Tanisik et al., 2016), e.g., by deep CNN features, and features based on distance between agents (Patron-Perez et al. 2012, Yang et al. 2012). Kong and Fu (2016) have further used the localization of body components and their relations for the recognition of social interactions, by modeling a set of spatiotemporal relations between body parts. Similar to these studies, but with a significantly richer set of body features and relations, our work uses full interpretation of the interacting agents, in order to achieve correct and robust interaction recognition.

3.2 Full interpretation of social interaction images

To deal with the extreme variability of images within a given interaction category such as ‘hug’, we first used the minimal-images approach described above (Sec. 2.1 and Sec. 2.3), in order to identify reduced configurations, which still provide sufficient support for correctly recognizing the interaction. Such minimal interaction images are useful to identify the visual components and relations, which are crucial for making the correct interpretation. Our study suggests that an image of interacting agents (e.g., ‘hugging’) contains multiple informative sub-configurations, where each one of them is sufficient for humans to recognize the interaction (Fig. 6A). Different configurations typically include different body parts e.g., a hand of an agent and the back of another, arms of the two agents, etc. Such sub-configurations can be clustered into several different ‘templates’ of the interaction category, which are defined by the parts that they contain. Since the number of parts in these sub-configurations is small, their variability is considerably reduced compared with fully-viewed images. Identifying these configurations individually, and then combining them together, can lead to a flexible and robust recognition of social interactions. We next describe psychophysical results of studying minimal images for social interactions.

3.2.1 A set of minimal images for social interactions

Minimal images become particularly useful at the limit of recognition, where further reduction of the image makes them unrecognizable. We applied the minimal images approach to find the most limited configurations from which humans can still recognize social interactions. We used a psychophysical study to identify the minimal recognizable configurations in social interaction images; these are local image regions in which the interaction type is recognizable, and which further reduction by either size or resolution turns them unrecognizable. To identify the minimal interaction configurations, we used a similar search procedure to the one in Sec. 2.1. The search started from a fully viewed interaction image, such as a ‘hug’, which was reduced in small steps, by cropping corners or reducing resolution. At each step, human interaction recognition was tested via MTurk. A minimal interaction configuration is an image region from which the interaction type is reliably recognized, but any further reduction in size or resolution makes the image un-recognizable (a recognition criterion set at 50% correct recognition by the MTurk subjects was used). Examples are shown in Fig. 7 for minimal interaction configurations.

The search started from various interaction images (e.g., two people ‘hugging’, ‘fighting’, ‘toasting’, ‘board playing, etc., examples in Fig. 7A), and each interaction image was used to identify a number of different, partially overlapping, local configurations (minimal images). Subjects were presented with images from different social interaction categories, as well as individual object images for control. Each image was presented to 30 different subjects, and each subject saw a single image from each interaction class. Overall, we had approximately 7000 different subjects participating in the study. More details about the psychophysics procedure are in Appendix A. For the discovered minimal configurations, we also tested the internal semantic components that humans can recognize in them (namely human interpretation of minimal hugging configurations; See partial lists of such components in Fig. 7B). The minimal interaction configurations varied in the body parts they contained. For example, one minimal ‘hug’ configuration included the agents’ faces and arms, while another contained only torsos and arms (without faces). Overall, our search generated minimal configurations coming from 8 different social interaction classes. The average size of the minimal interaction images was ~30^2 image samples.

Similar to the minimal configurations from object images, we found that in minimal interaction images too, small changes in the image could cause a large drop in human recognition of the interaction

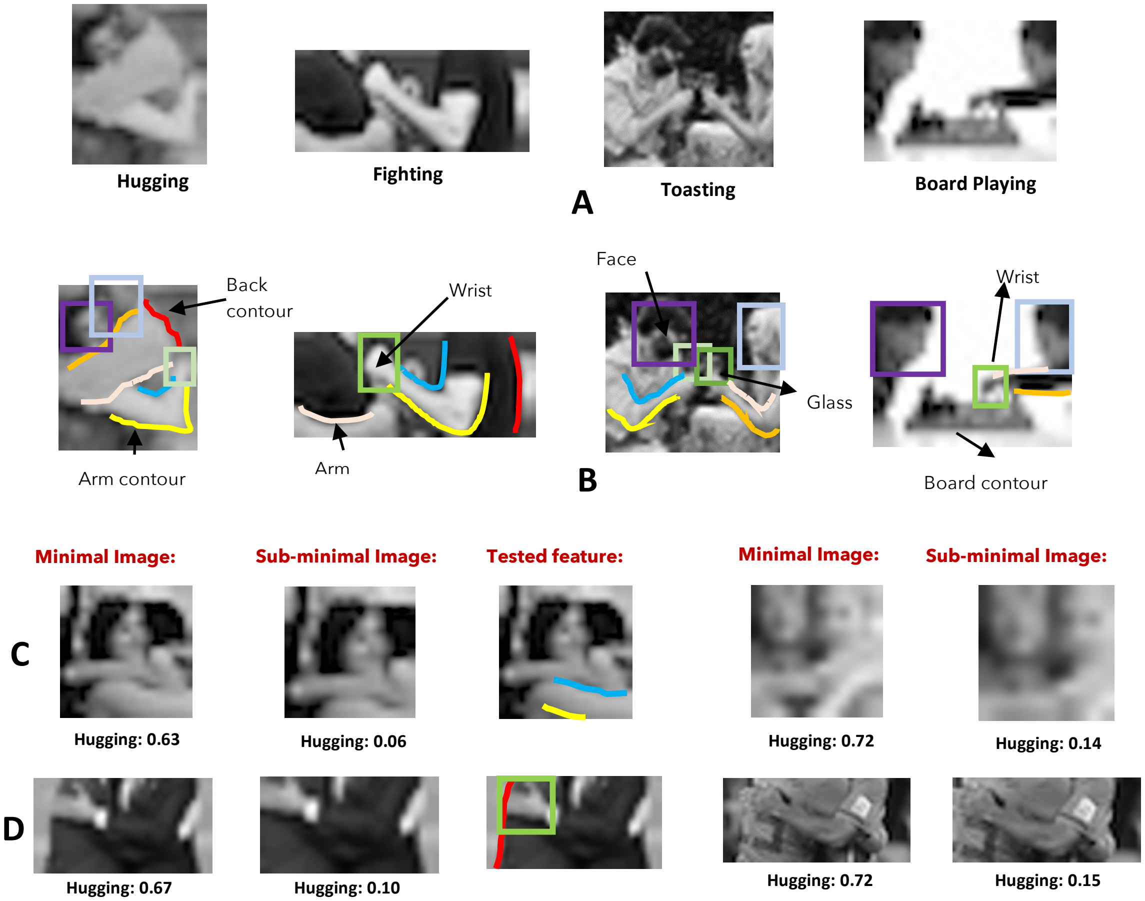

Figure 7. **(A).** Minimal interaction configurations for: 'hugging' and 'fighting' (agent-agent interactions), and for toasting and board playing (interactions involving two agents and objects.). **(B).** Human interpretation of the internal semantic components in the minimal configurations. **(C-D).** Inferring interpretation features from minimal and sub-minimal configurations. Columns from left to right: Minimal and sub-Minimal configurations, candidate critical interpretation feature, minimal and sub-minimal images with a similar loss of feature. Features included here are missing arm contours (in C) and 'touching' relation involving the hand of agent#1 and the back of agent#2 (in D).

type. Figs 8C-D show this characteristic for 'hug' minimal configurations. For each minimal image, our study identified a set of internal features that are used by the recognition process, together with informative properties and relations between adjacent components. The internal components and their spatial relations, identified in the psychophysical study, were next used in a computational model for full interpretation and recognition of interaction images, described in the next section.

3.2.2 A model for full interpretation of social interaction images

The model for the automatic interpretation of social interactions is based on the structured-prediction framework, discussed in Sec. 2.3, and in more detail in (Ben-Yosef et al. 2018). The model was trained to perform the interpretation of a single interaction type (e.g., 'hugging'), and a single interaction configuration (e.g., a configuration showing an arm and a back, as in the examples of Fig. 8A). The interpretation score that the model provides was also used for recognition, by comparing the model score to a decision threshold. To train the model, we collected multiple examples from the same minimal configuration (the positive set), as well as interpretations for all the examples provided by a human annotator. A negative set for the model was composed from multiple non-class examples of similar size to the positive images, containing various objects, non-interacting agents, or interacting agents from a different interaction class (e.g., 'fighting' examples were used in the negative set of a 'hugging' model).

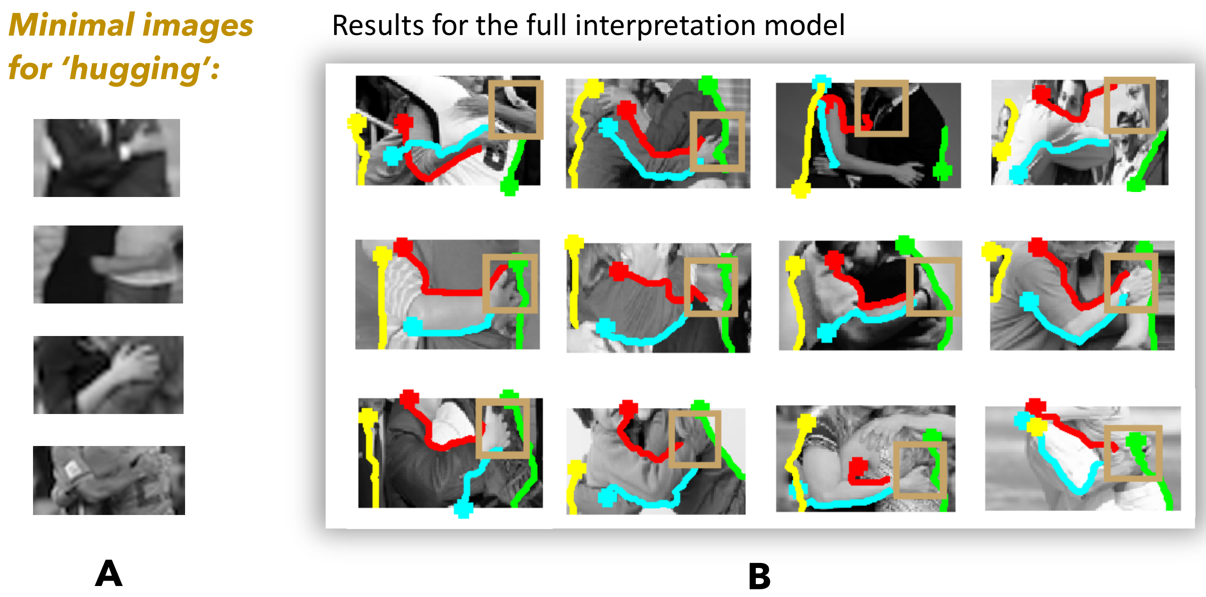
Minimal images for 'hugging':
Results for the full interpretation model
A
B

Figure 8. A model for full interpretation of minimal interaction configurations. **(A).** Four 'hugging' minimal images from the same type. A full interpretation model was trained on 120 examples of minimal images from this type, provided with human detailed interpretation of the internal parts in them, as well as negative examples from non-hugging images. **(B).** The model was tested on novel examples of minimal images of the same type as in (A),and returned the predicted interpretation for these examples. The results, few of them are shown here, show good match with human interpretation (see text for details and quantitative evaluation).

The interpretation model was based on a structured random forest algorithm (Breiman, 2001), similar to Sec. 2.3, with the structural features in the 'extended' set of Sec. 2.4, together with additional structural features, which were found to be useful for interpretation of minimal interaction configurations, using the same procedure used in Sec. 2.4 and (Ben-Yosef et al. 2018).

The new features and relations incorporated in the model for recognizing interactions were inferred from human interpretation and recognition of minimal and sub-minimal images for social interactions, as exemplified in Fig. 7C-D. A minimal image was compared to its slightly reduced, but unrecognizable sub-minimal configuration, and a feature of a part, or a relation between parts, which exist in the minimal but not in the corresponding sub-minimal image, was identified. For interactions, the search for structural features was extended beyond pure image relations (e.g., 'contour parallelism', 'inside/outside', as in Sec. 2.3 and Fig. 3), to include more properties and relations regarding the contact points of parts belong to the different agents. Such features coming from minimal and sub-minimal images included a unary feature of a closed hand configuration, or a binary feature of a hand 'touching' a person's back (Fig. 7D). Such relations are generic and can be used for interpretation of various interaction types. When a candidate feature or relation were identified by the difference between a minimal and its sub-minimal image, a computational test was applied for deciding whether to include it in the interpretation model. To perform the test, we compared the performance of the interpretation model using two versions of the model, one with the added feature and the other without it (using only the extended set described in Sec. 2.4). The new feature was added to the interpretation model only if it contributed to the model's performance (a similar to the paradigm used in Ben-Yosef et al., 2018).

3.2.3 Experimental evaluation

The interaction recognition model was trained and tested on a number of interaction configurations. Several examples of a minimal 'hugging' configuration are shown in Fig. 8A. The model was trained with a positive set including 120 examples, provided with full interpretation annotations by a human observer. A negative set for the training procedure included 5000 non-hug image regions, of the

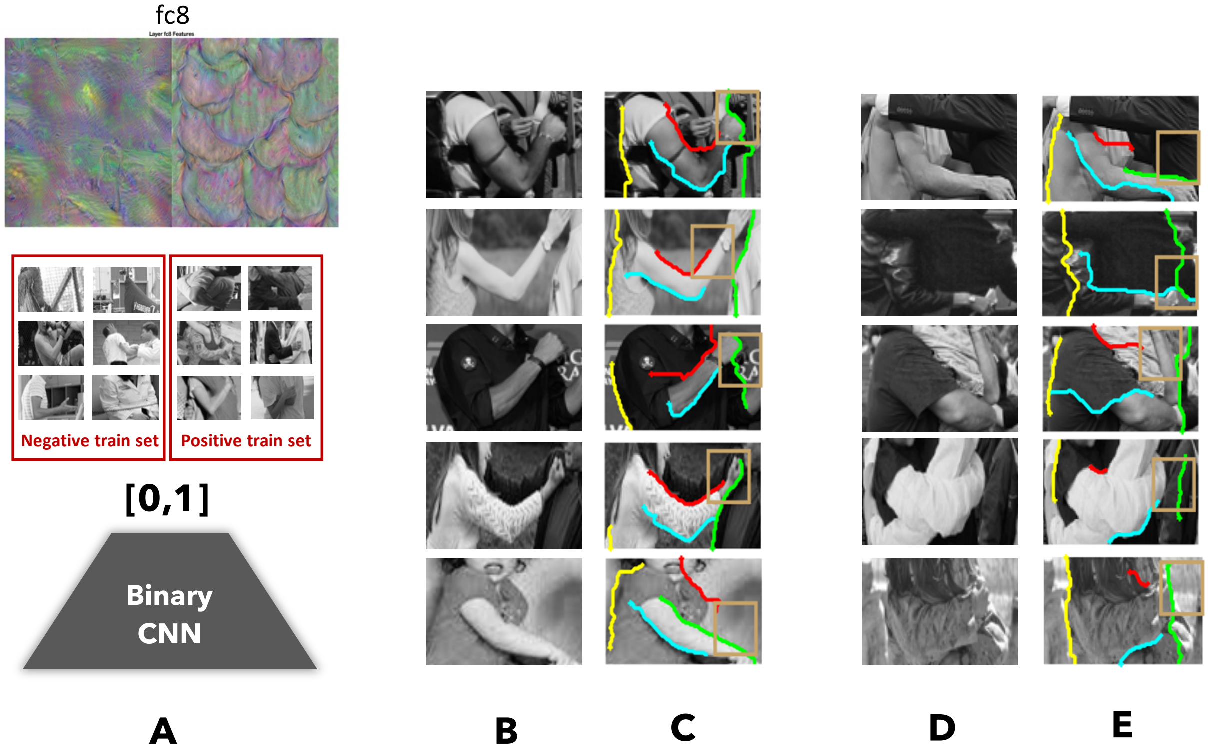

Figure 9. **(A).** A binary CNN classifier was fine-tuned by examples of minimal hugging images as positive set, and non-hugging examples as negative set. The top row shows visualization of the features activated by the classifier at the final stage of the net (here layer fc8 of AlexNet, Krizhevsky et al., 2012). Visualization produced by DeepDreamImage visualization (2017), showing feature activation for 'non-hug' in the left activation map, and 'hug' in the right activation map. The activation maps suggest that the network will output high activation score for an arm-like structure at multiple locations, but there is no indication of finer features and relations such as open-hand configuration, 'touching' relation, etc. **(B).** Non-class test examples that confuse the network and cause it to respond with high activation at the output layer ('hard-negative examples'). These examples were not confusable to the MTurk subjects. **(C).** The full interpretation model is applied to the hard-negative examples and returns low hugging scores since expected parts or relations do not exist. **(D).** Sub-minimal configurations taken from hugging images, but the MTurk users could not recognize as 'hug', which triggered high positive responses by the CNN. **(E).** The full interpretation model returns low hugging score for these examples, which is consistent with human recognition.

same size of the minimal image examples. The model was evaluated on a test set of 120 minimal images (see examples for interpretation results in Fig. 8B), and the overlap between the predicted interpretation by the model and human interpretation was measured. Table 1 shows the average overlap for two versions of our model, which are different by the structural features that they use: the basic set of features (Sec. 2.4), an extended set of features (Sec. 2.4) with the addition of features for contact points, as explained in Sec. 3.2.2.

The average overlap measured for the test examples of the two interpretation versions show that the addition of feature and relations beyond the basic set is useful to achieve accurate interpretation of social interaction images. Specifically, there was a significant improvement in interpretation results between the basic and extended models ($P<9.9*10^{-5}$, n=5, one-tailed paired t test). Examples for the predicted interpretation by the extended set are in Fig. 8B, showing how the model is able to generalize well the interpretation to novel instances of the local 'hugging' configuration.

To further explore the role of interpretation in interaction recognition, we next tested models for the recognition of minimal interaction images, on images which are different in both interaction type (namely, a hug or a non-hug image), as well as the tone of interaction, as reported by human observers via MTurk survey. On our collected set of local hugging configurations used above, human subjects were also asked to grade the tone of the hug, on a scale of 1 to 3, where 3 is 'an intimate, warm hug', 2 is 'a formal, neutral hug', 1 is 'a distant, cold hug'. To test recognition of the interaction type, the interpretation score was used as a measure for recognition. To test the recognition of interaction tone, we used the learned structural representation (the relations vector) of our interpretation model, together with an SVM classifier trained for the three levels of interaction tone.

The experimental results for recognition of the interaction type and tone were matched to human recognition measured via MTurk. We compared human judgment with our model and with a binary CNN model trained on the task of classifying 'hug' vs. 'non-hug'. The CNN classifier was based on the AlexNet (Krizhevsky et al., 2012) and the VGG19 (Simioniyan and Zisserman, 2015) network models. The training set for classifying the interaction type by both the network model and the interpretation model included 120 example of the minimal image as positive set, and 5,000 examples from non-interaction category as negative set (see positive and negative examples in Fig. 9A). The positive test set included 120 annotated minimal images, and negative set included 400 non-hug image regions (hard negatives) containing closely interacting agents that were not recognized as 'hugging' by human observers on the MTurk (examples are shown in Fig. 9B, together with the interpretation provided by the model with the extended relations set in Fig. 9C). The Average Precision (AP) was then computed for both the interpretation score (provided by the model with extended set), and a binary deep network classifier based on the VGG19 (Simioniyan and Zisserman, 2015) CNN model. (AP is an evaluation measure for scored retrieved results. Here both binary CNN and interpretation model retrieve the 'hugging' minimal configuration in novel images). A large improvement in AP was obtained between the results provided by the interpretation model (0.80) and the AP provided by the CNN model (0.69).

'Hugging' Components	Average Jaccard overlap	
	Basic set	Extended set
Agent 1, arm upper contour	0.27	0.48
Agent 1, arm lower contour	0.24	0.43
Agent 1, Back contour	0.39	0.59
Agent 2, Back contour	0.38	0.56
Agent 1, hand region	0.42	0.55
mean	0.34	0.52

Table 1. The overlap (Jaccard) index between human and model hugging interpretation. The overlap index is computed for each component, and also for the average overlap. The overlap was computed and compared for two versions of our model, with the basic and extended set of relations, for different configurations.

For testing predictions for the tone of interaction, we conducted a preliminary experiment using 50 examples of the local hugging configurations used above (Fig. 9A), for which humans gave consistent ratings about the tone of hugging interaction (each image rated by 20 different MTurk users). The interpretation model for hugging configurations (Fig. 8B) was applied on these examples, and then used to classify the image to one of the three tone categories. Our preliminary results show a good match to the psychophysics data, and motivate more computational experiments in this direction.

In summary, we presented in this section a novel interpretation scheme, applied to local image regions of interacting people. The scheme provides a computational model for identifying and interpreting such configurations of social interactions, and it also suggests a possible model for the interpretation process performed by humans. The scheme can identify complex interactions between agents, and can produce a full interpretation of internal components, in particular body parts of the interacting agents.

The method is based on the detection and interpretation of parts of the image that match a minimal configuration, from which a human observer can identify the interaction. These configurations of body parts are less variable than fuller configurations, and their interpretation helps focusing on the meaningful cues, which are often subtle and small in size. In a fully viewed image, more than one of these configurations may be found, and the scheme combines the interpretation of the component configurations. Future directions and extensions can include a range of social interactions, and

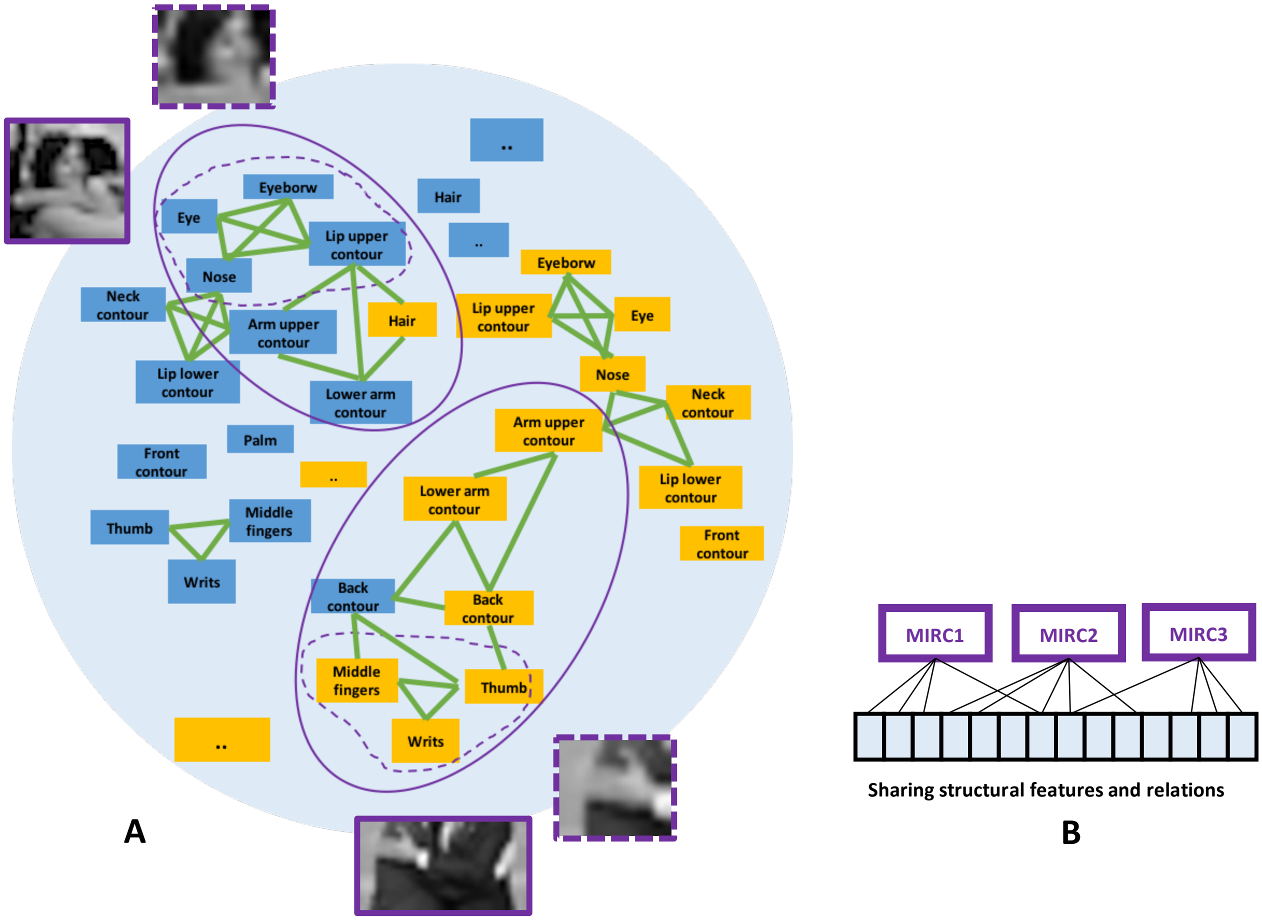

Figure 10. Expansion of minimal images' interpretation to surrounding regions in a visual scene. **(A).** In the suggested view, scene understanding begins with interpretation of local but sufficiently informative regions ('absolute minimal images'), from which the interpretation expands to larger regions, based on the visual task and goal. In the case of understanding social interactions, an absolute minimal image could be a hand or a face region of one of the interacting agents, while the extended regions include body parts from both agents (and form the 'interaction minimal images'). Here we plot body parts of two interacting agents (from agent#1 in blue, from agent#2 in orange), we well as few inter-relations between body parts (in green connectors). The two solid ellipses correspond to two interactions minimal configurations (which correspond to the images shown with solid-line image border), and the two dashed purple contours correspond to the two absolute minimal configurations (and referred to the images with dashed-line image border). **(B).** Different minimal configurations (either 'absolute' or 'interaction') may overlap in the parts and relations they use (their structural features). A mechanism of sharing structural features enables a more efficient learning of interpretation procedures for minimal images.

interactions of more than two agents. In addition, the general interpretation process described here could be applied to images beyond social interactions, in particular, interactions between agents and objects.

4. From understanding minimal images to the understanding of larger scenes

Vision is a process of recovering knowledge about the surrounding world (semantic information) from images. Humans can extract semantic information from images at a broad range of scales, from small parts of objects to configurations of multiple objects and agents. The studies reported here suggest that the ability to recognize and understand fine local objects structure on the one hand, and to recognize interactions between agents and objects on the other, share a common process of detailed local interpretation. Using psychophysical and computational studies based on the perception of minimal images, we proposed a model for local image interpretation. This model uses a structural description of a

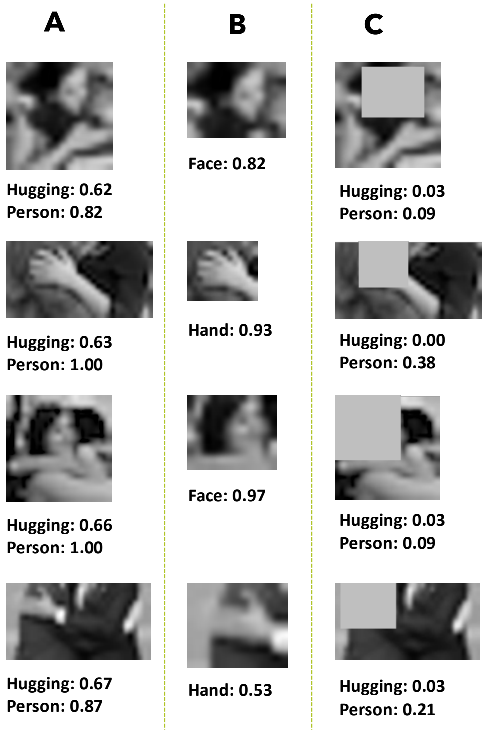

Figure 11. psychophysical support for the expansion of interpretation. **(A).** Hug-interaction minimal configurations, where humans recognize the agents and their type of interaction. **(B).** Sub-images containing an 'absolute' minimal configuration: the hug is no longer recognized, but the body part is recognized with accuracy above threshold. **(C).** The original image without the absolute minimal configuration. Both the interaction type as well as the remaining body parts become unrecognizable. Recognition rates are shown under each image.

local image region, which includes all the fine semantic components that humans can perceive in the region, along with a set of relations between them. We next suggested that the recognition of interactions between two agents, or an agent and an object, often depend in part on a detailed analysis of the regions where they interact. Consequently, the model for the full interpretation of local image regions can contribute not only to object recognition, but also to the recognition and interpretation of interactions between agents (as well as interactions between agents and objects).

Our studies focused on minimal recognizable images for three reasons. First, minimal images are useful for identifying the visual features and relations that play a role in image interpretation, using in particular comparisons between minimal images and their similar, but unrecognized sub-images. Second, humans can reliably recognize and interpret minimal images, and therefore a model of human image understanding should be able to account for these capacities. Third, as discussed further below, we suggest that during the recognition of natural images (of a large size and high resolution), local, recognizable image regions, similar to minimal images, provide useful building blocks for the image interpretation process. In this section, we turn therefore to discuss the possible role of minimal images in the recognition and interpretation of full scale real-world scenes. We suggest that the level of minimal images provides an effective starting point for the process of real-world scene understanding. Minimal images are by definition the smallest image regions that do not require any additional context to be recognized. They can therefore be recognized first, and then provide context for the subsequent recognition of additional image regions, which cannot be recognized on their own, but can be disambiguated and recognized based on the context provided by preceding recognition stages.

The interpretation model described above was developed for the task of interpreting a single minimal image in isolation. A full size natural image will usually contain multiple minimal images, at multiple locations and a range of resolutions. The availability of multiple minimal images raises the issue of integrating interpretation results across spatial locations and scales, but it also makes the process easier in certain respects as well as more robust. We turn next to consider aspects of the recognition process that go beyond a single minimal image, to recognize and interpret larger parts of a natural scene.

In the interpretation model discussed in Sec. 2.3, 2.4, the accurate recognition and full interpretation of a single minimal image is obtained by an initial fast feed-forward recognition stage, followed by an interpretation stage. The second stage provides a full local interpretation, and it also increases the accuracy of the initial recognition stage. A full object image provides not just a single minimal image, but multiple minimal configurations, at multiple locations and scales. Combining their results will increase the accuracy of the initial feed-forward recognition stage, and consequently for a full object image, the initial feed-forward stage will be able to produce accurate recognition, but still without providing the full interpretation of fine details. For example, a full horse image will contain minimal configurations such as the horse's overall shape, as well as different parts such as the head, torso, legs, or tail. A feed-forward activation of a subset of these configurations will indicate with high likelihood the presence of a horse. For some tasks such fast recognition without details may be sufficient, but others will depend on fine details of structure and interactions. Depending on the task, the recognition process can next proceed to provide a detailed interpretation of selected image parts. For example, for recognizing agents' interactions discussed above, a detailed interpretation of the interaction regions will often be required. The model suggests that the interpretation stage is applied in a top-down manner to selected locations, rather than being applied uniformly across the image. In our model, the interpretation starts from a subset of minimal images recognized in the first stage, and then expands to nearby regions.

The expansion process from minimal images to surrounding regions is illustrated schematically in Fig. 10, for the case of agents' interactions. The figure shows a graph composed of internal object parts, together with relations between parts. The graph components come from the image of two interacting people, with blue nodes coming from one person, and yellow nodes from the other. The first stage in our model detects minimal configurations of the interacting agents, such as a human hand, or a human face (Fig. 10A, dashed purple contours). The recognition process continues to produce an internal interpretation of the hand or face regions, and then extends the interpretation to nearby regions. The process eventually gets to an extended region in which the interpretation is sufficient to provide information about the agents' interaction (Fig. 10A, purple ellipses). These extended regions are the minimal interaction configurations, discussed in Sec. 3.2.1. On this view, minimal interaction configurations contain smaller minimal 'absolute' recognizable configurations, from which recognition and interpretation start.

In recent psychophysical studies, we obtained support for the structure of minimal interaction configurations, as composed of a minimal absolute configuration, combined with additional features, which are not recognizable on their own. Examples are shown in Fig. 11. Column A shows example of Hug-interaction minimal configurations. Column B shows sub-images containing an 'absolute' minimal configuration: the hug is no longer recognized, but the body part is recognized with accuracy above threshold. Column C shows the original image without the absolute minimal configuration. Here, not only the hug becomes unrecognizable, but also all the body parts in the remaining image become unrecognized. In the model, the absolute configuration is recognized first, and the interaction is subsequently recognized by an expansion process. It will be of interest to examine this predicted dynamics of the recognition process in further psychophysical experiments.

Acknowledgments

We thank Alon Yachin for help with data collection, and Deborah Ben-Yosef Gelernter for help with creating the figures.

Funding Statement

This work was supported by grant 2016731 from the United States-Israel Binational Science Foundation (BSF) and US National Science Foundation (NSF), The German Research Foundation DFG grant ZO 349/1-1, the MIT-IBM Brain-Inspired Multimedia Comprehension project, and the Center for Brains, Minds and Machines, funded by NSF Science and Technology Centers Award CCF-1231216.

Appendix A: A search for minimal interaction images - Experimental details

This section describes in more details the MTurk procedure for testing human recognition of social interaction. A Turk subject was presented with an interaction image, and was asked to describe in free text what is the object, object part, action or interaction that he/she saw in the image. In the presented set of images to the subject, there was always one control object image (e.g., a swan), and a control image of interaction (e.g., shaking hands), from which the object or interaction is easy to recognize (verified in a preliminary pilot experiment). A subject's survey was rejected if the answers to these two 'catch' images were incorrect.

A decision if the subject's answer is correct or not was based on a list of keywords that were acceptable as true interaction descriptions, which we selected and predefined in a pilot experiment before the main experiment was conducted. In this pilot experiment we showed subjects fully viewed high-resolution interaction images, and collected the union of the words used for describing them. Table 2 shows the list of keywords that we accepted as 'correct' description of 'hugging', and the list of keywords that we considered as incorrect. To verify repeatability of the psychophysics results, we re-sent 10% of the tested images in our experiment again to the MTurk, and matched human recognition results from the first and second surveys. The match showed high correlation.

Interaction	Considered 'correct'	Considered 'incorrect'
Hugging	▪ **People Hugging, embracing** ▪ **A parent holding a child** ▪ **A couple holding each other** ▪ **People kissing** ▪ **People dancing**	▪ **A man is holding another man** ▪ **A man is touching another man** ▪ **People fighting, arguing, yelling, scolding** ▪ **People smiling, talking, playing**

Table 2. Shows examples for the answers that we considered as correct and incorrect for recognizing 'hugging'.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2425-2433).

2. A. Arnab, S. Jayasumana, S. Zheng, and P. H. Torr. Higher order conditional random fields in deep neural networks. In European Conference on Computer Vision (pp. 524-540), 2016.

3. Ben-Yosef, G., Assif, L., Harari, D., & Ullman, S. (2015). A model for full local image interpretation. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society.*

4. Ben-Yosef, G., Assif, L., & Ullman, S. (2018). Full interpretation of minimal images. *Cognition,* 171, 65-84.

5. Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. Psychological review, 94(2), 115-147.

6. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

7. Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017, July). Realtime multi-person 2d pose estimation using part affinity fields. In CVPR (Vol. 1, No. 2, p. 7).

8. Chao, Y. W., Wang, Z., He, Y., Wang, J., & Deng, J. (2015). Hico: A benchmark for recognizing human-object interactions in images. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1017-1025).

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

10. 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), 1-1. doi: 10.1109/TPAMI.2017.2699184 4.

11. Centelles, L., Assaiante, C., Nazarian, B., Anton, J. L., & Schmitz, C. (2011). Recruitment of both the mirror and the mentalizing networks when observing social interactions depicted by point-lights: a neuroimaging study. PloS one, 6(1), e15749.

12. DeepDreaming with TensorFlow (2017). https://github.com/tensorflow/tensorflow/blob/master/tensorflow/examples/tutorials/deepdream/deepdream.ipynb

13. 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/.
14. 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.
15. 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.
16. 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.
17. Felleman, D. J., & Van, D. E. (1991). Distributed hierarchical processing in the primate cerebral cortex. Cerebral cortex (New York, NY: 1991), 1(1), 1-47.
18. Ferrari, V., Jurie, F., & Schmid, C. (2010). From images to shape models for object detection. International journal of computer vision, 87(3), 284-303.
19. Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., ... & Schmid, C. (2018). AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. In press.
20. Gu, C., Sun, C., Vijayanarasimhan, S., Pantofaru, C., Ross, D. A., Toderici, G., ... & Malik, J. (2018). AVA: A video dataset of spatio-temporally localized atomic visual actions. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. In press. (arXiv preprint arXiv:1705.08421.)
21. Girshick, R., Iandola, F., Darrell, T., & Malik, J. (2015). Deformable part models are convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 437-446).
22. Hall, E. T. (1966). The hidden dimension.
23. Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive development*, 26(1), 30-39.
24. Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in cognitive sciences, 11(10), 428-434.*
25. Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. Proceedings of the National Academy of Sciences, 201714471.
26. 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.
27. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., & Girshick, R. (2017, July). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *In Computer Vision and Pattern Recognition (CVPR)*, 2017 IEEE Conference on (pp. 1988-1997). IEEE.
28. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... & Suleyman, M. (2017). The kinetics human action video dataset. arXiv preprint arXiv:1705.06950.
29. Kong, Y., & Fu, Y. (2016). Close human interaction recognition using patch-aware models. IEEE Transactions on Image Processing, 25(1), 167-178.
30. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... & Bernstein, M. S. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision, 123(1), 32-73.
31. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 1097-1105.
32. Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
33. Leary, T. (1958). Interpersonal diagnosis of personality. *American Journal of Physical Medicine & Rehabilitation*, 37(6), 331.
34. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

35. Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 3431-3440.
36. Li, L. J., Socher, R., & Fei-Fei, L. (2009). Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 2036-2043). IEEE.
37. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. Proceedings of the European Conference on Computer Vision, 740-755.
38. Mascaro, O., & Csibra, G. (2012). Representation of stable social dominance relations by human infants. Proceedings of the National Academy of Sciences, 109(18), 6862-6867
39. Patron-Perez, A., Marszalek, M., Reid, I., & Zisserman, A. (2012). Structured learning of human interactions in TV shows. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(12), 2441-2453.
40. Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. Cognitive psychology, 8(3), 382-439.
41. Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nature neuroscience, 2(11), 1019-1025.
42. 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.
43. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations.
44. Sliwa, J., & Freiwald, W. A. (2017). A dedicated network for social interaction processing in the primate brain. Science, 356(6339), 745-749.
45. Tan, P. N., Steinbach, M., & Kumar, V. (2006). Introduction to data mining (Vol. 1). Boston: Pearson Addison Wesley.
46. Tanisik, G., Zalluhoglu, C., & Ikizler-Cinbis, N. (2016). Facial Descriptors for Human Interaction Recognition In Still Images. *Pattern Recognition Letters.*
47. Thomsen, L., Frankenhuis, W. E., Ingold-Smith, M., & Carey, S. (2011). Big and mighty: Preverbal infants mentally represent social dominance. science, 331(6016), 477-480.
48. Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. Journal of personality and social psychology, 37(3), 395.
49. 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.
50. 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*, 3622-3629.
51. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2017). Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence, 39(4), 652-663.
52. 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 Computer Vision (ICCV), 2011 IEEE International Conference on (pp. 1331-1338). IEEE.
53. Yang, Y., Baker, S., Kannan, A., & Ramanan, D. (2012). Recognizing proxemics in personal photos. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3522-3529. IEEE.
54. Yang, S., Luo, P., Loy, C. C., & Tang, X. (2015). From facial parts responses to face detection: A deep learning approach. In Proceedings of the IEEE International Conference on Computer Vision (pp. 3676-3684).

55. 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.
56. Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience, 19(3), 356.
57. Zhu, S. C., & Mumford, D. (2007). A stochastic grammar of images. Foundations and Trends® in Computer Graphics and Vision, 2(4), 259-362.
58. Zhu, Y., Groth, O., Bernstein, M., & Fei-Fei, L. (2016). Visual7w: Grounded question answering in images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4995-5004).