Cognition 201 (2020) 104263

Contents lists available at ScienceDirect

Cognition

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

Minimal videos: Trade-off between spatial and temporal information in human and machine vision

Guy Ben-Yosef[a,d,*], Gabriel Kreiman[b,d], Shimon Ullman[c,d]

[a] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[b] Children's Hospital, Harvard Medical School, Boston, MA 021155, USA
[c] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 7610001, Israel
[d] Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Keywords:
Minimal videos
Minimal images
Comparing deep neural networks and humans
Integration of spatial and temporal visual information
Visual dynamic recognition

ABSTRACT

Objects and their parts can be visually recognized from purely spatial or purely temporal information but the mechanisms integrating space and time are poorly understood. Here we show that visual recognition of objects and actions can be achieved by efficiently combining spatial and motion cues in configurations where each source on its own is insufficient for recognition. This analysis is obtained by identifying minimal videos: these are short and tiny video clips in which objects, parts, and actions can be reliably recognized, but any reduction in either space or time makes them unrecognizable. Human recognition in minimal videos is invariably accompanied by full interpretation of the internal components of the video. State-of-the-art deep convolutional networks for dynamic recognition cannot replicate human behavior in these configurations. The gap between human and machine vision demonstrated here is due to critical mechanisms for full spatiotemporal interpretation that are lacking in current computational models.

1. Introduction

Previous behavioral work has shown that visual recognition can be achieved on the basis of spatial information alone (Potter & Levy, 1969; Ullman, Assif, Fetaya, & Harari, 2016), and on the basis of motion information alone, as in the case of identifying human activities from biological motion (Johansson, 1973). At the neurophysiological level, neurons have been identified that respond selectively to objects and events based on purely spatial information, or motion information alone (Oram & Perrett, 1996; Perrett et al., 1985; Sáry, Vogels, & Orban, 1993; Vaina, Solomon, Chowdhury, Sinha, & Belliveau, 2001). However, several behavioral studies have also provided strong support suggesting that a combination of spatial and temporal information can aid recognition. A series of elegant experiments showing moving object image through a slit (Morgan, Findlay, & Watt, 1982; Parks, 1965; Rock, 1981; Zollner, 1862) suggest that both shape and motion cues may cooperate to help recognition, but whether and how space and time may be integrated remain unclear. Studies on perceptual organization from visual dynamics (e.g., dynamic grouping and segmentation from motion (Anstis, 1970); spatiotemporal continuation and completion (Kellman & Cohen, 1984)) also combine motion and shape information (e.g., spatial proximity or spatial orientation with common motion direction), but the role of motion is typically limited in this case to figure-ground segmentation. A recent study has shown limitations on the integration of spatial and temporal information in recognition by demonstrating how presenting different parts of an object asynchronously leads to a severe disruption in recognition (Singer & Kreiman, 2014) and that visually selective neurophysiological signals are sensitive to this temporal information (Singer, Madsen, Anderson, & Kreiman, 2015).

One of the domains in which temporal information is particularly relevant is action recognition. Several computational models have been developed to recognize actions from videos, combining spatial with temporal information. For example, in recent computer vision challenges, the goal is to classify a video clip (e.g., a 10 sec length video) into one of several possible types of human activities (e.g., Playing Guitar, Riding a Horse, etc.; UCF101 dataset by Soomro, Zamir, & Shah, 2012; Kinetics dataset by Kay et al., 2017). Earlier approaches for modeling action recognition (e.g., Giese & Poggio, 2003) suggest a dual path approach in which form (spatial) and motion (temporal) information are initially processed separately, and are then combined for the final action label prediction. Such models attempted to explain biological motion (e.g., by representing actions as collections of features from a "vocabulary" of dynamic templates or optical flow

* Corresponding author at: Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
E-mail address: gby@csail.mit.edu (G. Ben-Yosef).

https://doi.org/10.1016/j.cognition.2020.104263
Received 25 September 2019; Received in revised form 3 March 2020; Accepted 5 March 2020

patterns, e.g., Casile & Giese, 2005; Cavanagh, Labianca, & Thornton, 2001; Schindler & Van Gool, 2008; Thurman & Grossman, 2008) or to use mechanisms similar to structures in cortical visual motion areas (e.g., Jhuang, Serre, Wolf, & Poggio, 2007 via aggregation of spatiotemporal filters in a neural network model).

Modern models for action recognition from spatiotemporal input are based on deep network architectures, and can be partitioned into the following three groups depending on how they integrate spatial and temporal information: (i) Feed-forward networks with 3D convolutional filters, where the temporal features are processed together with the spatial ones via 3D convolutions in the space-time manifold (Hara, Kataoka, & Satoh, 2018; Karpathy et al., 2014; Tran et al., 2018; Tran, Bourdev, Fergus, Torresani, & Paluri, 2015), but it remains unclear if and how shape and motion cues are actually combined; (ii) two-stream networks based on late integration of two network 'modules' where one module is trained on spatial features (fine-tuned from a pre-trained static recognition network on ImageNet), and a second module that is trained on optical flow from consecutive frames (Feichtenhofer, Pinz, & Wildes, 2016; Feichtenhofer, Pinz, & Zisserman, 2016; Simonyan & Zisserman, 2014). Here, the integration of temporal and spatial features takes place at a subsequent, higher stage, whereas in human vision motion also has a low-level role as exemplified in figure-ground segmentation. (iii) Models combining deep convolutional networks with Long Short-Term Memory (Hochreiter & Schmidhuber, 1997) units based on recurrent connections (Donahue et al., 2017). The input is a sequence of frames, each of which is passed through a convolutional network followed by a layer of LSTM units with recurrent connections. Here too, the integration of temporal and spatial features takes place at late stages, and it is unclear how motion and spatial information are specifically integrated through the recurrent connections.

Despite progress in action classification, it remains unclear whether current models make an adequate and human-like use of spatiotemporal information. In order to evaluate the use of spatiotemporal integration by computational models, it is crucial to construct test stimuli that 'stress test' the combination of spatial and dynamic features. A difficulty with current efforts is that in many action recognition data sets (e.g. UCF101) high performance can be achieved by considering purely spatial information (Feichtenhofer, Pinz, & Wildes, 2016; Feichtenhofer, Pinz, & Zisserman, 2016), and therefore those stimuli are not ideally set up to rigorously test spatiotemporal integration. Furthermore, an important aspect of using spatiotemporal information in human vision is the ability to "fully interpret" an image, in contrast with current computational architectures, which merely assign action labels. Human recognition can not only label actions, but can also provide a full interpretation by identifying and localizing object parts, as well as inferring their spatiotemporal relations. We conjectured that when humans correctly recognize an action in a video, they can not only label the action, but they can also provide a detailed localization of the parts that are involved in the action, as well as the spatial and spatiotemporal properties and inter-relations between parts (Ben-Yosef, Assif, & Ullman, 2018). We refer to this detailed understanding of the video as 'spatiotemporal interpretation'. Existing schemes for spatiotemporal interpretation use direct extensions of static semantic segmentation techniques (Cheron, Laptev, & Schmid, 2015; Hur & Roth, 2016; Kundu, Vineet, & Koltun, 2016), which do not provide the full human-like spatiotemporal interpretation.

Here we developed a set of stimuli that can directly test the synergistic interactions of dynamic and spatial information, to identify spatiotemporal features that are critical for visual recognition and to evaluate current computational architectures on these novel stimuli. We tested *minimal spatiotemporal configurations* (also referred to below as *minimal videos*), which are composed of a set of sequential frames (i.e., a video clip), in which humans can recognize an object and an action, but where further small reductions in either the spatial dimension (i.e., reduction by cropping or down sampling of one or more frames) or in the time dimension (i.e., removal of one or more frames from the video) turns the configuration unrecognizable, and therefore also uninterpretable, for humans. This work follows recent studies on minimal configurations in static images (termed minimal recognizable configurations, or 'minimal images' (Ben-Yosef et al., 2018; Ben-Yosef & Ullman, 2018; Ullman et al., 2016)), extending the concept of minimal configurations to the spatiotemporal domain. In static images, it was shown that at the level of minimal configurations, small image changes can cause a sharp drop in human recognition (Ullman et al., 2016), and that recognizable minimal object images are also interpretable, i.e., humans can identify not only the object category but also the internal object parts and their inter-relations (Ben-Yosef et al., 2018). These properties provided a mechanism to study computational models for human interpretation, and also to study the link between object recognition and object interpretation in the human visual system (Ben-Yosef et al., 2018; Ben-Yosef & Ullman, 2018). In particular, the sharp drop in recognition between minimal images, and their similar, but unrecognizable sub-minimal images (i.e., the slightly reduced images) was used to identify critical recognition features, which appear in the minimal, but not the corresponding sub-minimal images. The goal in this study is then to similarly investigate critical spatiotemporal features for recognition and interpretation, as well as integration of spatial and motion cues, comparing minimal videos with their spatial and temporal sub-minimal versions.

We show that recognition can be achieved by efficiently combining spatial and motion cues, in configurations where each source on its own is insufficient for recognition. Recognition and spatiotemporal interpretation go together in these minimal video configurations: once humans can recognize the object or action, they can also provide a detailed spatiotemporal interpretation for them. These results pose a new challenge for current spatiotemporal recognition models, since our tests show that existing models cannot replicate human behavior on minimal videos. The results further add to the growing recent discussion about the differences between deep neural networks and human vision (Schofield, Gilchrist, Bloj, Leonardis, & Bellotto, 2018), here in the important domain of dynamic visual recognition Finally, the results suggest directions for future extensions of computational models, to better capture human performance in the interpretation of dynamic patterns.

2. Results

We first describe psychophysical experiments to find minimal videos in short video clips taken from computer vision datasets. We then describe human spatiotemporal interpretation of minimal videos, including the identified components within the minimal videos. Finally, we test existing computational models for recognition from spatiotemporal input on our set of minimal videos, and we compare the models' results with human recognition.

2.1. Search for minimal videos

The search for each minimal video started from a short video clip, taken from the UCF101 dataset (Soomro et al., 2012), in which humans could recognize a human-object interaction. We used examples from the UCF101 dataset because they contain a single agent, performing a single action, and it is a common benchmark for evaluating video classification algorithms in the computer vision literature. The search included 18 different video snippets, from various human-object interaction categories (e.g., 'a person rowing', 'a person playing violin', 'a person mopping', etc., see Table 1 for a full list).

Similarly to the work with static minimal images (Ullman et al., 2016), we think of the size of image patches in terms of samples required to represent the image without redundancy. The original video snippets were reduced to a manually selected 50 × 50 pixel square region, cropped from 2 to 5 sequential non-consecutive frames, and taken at the same positions on each frame (details below). These

Table 1
UCF101 categories used for search of minimal videos.

Biking
Rowing
Playing violin
Playing flute
Playing tennis
Playing piano
Mopping
Cutting
Typing

regions served as the starting configurations in the search for minimal video configurations described below. In the default condition, frames were presented dynamically in a loop at a fixed frame rate of 2 Hz (Materials and methods). An example of a starting configuration and a minimal video is shown in Fig. 1 and the path to create it is illustrated in Fig. 2.

Frames and frame regions for the starting configurations were selected such that the agent, the object, and the agent-object interaction were recognizable from each frame. The selected frames were taken at a temporal interval Δt ($\Delta t = 500 \pm 100$ ms), which encompasses the range of time intervals to complete natural body movements in the video clips that we considered (e.g., to lift a hand, etc.). An illustration of the starting configuration for one of these examples is shown in Fig. 1A. Because of the dynamic nature of the stimuli used in this study, it is difficult to appreciate the effects from static renderings. Therefore, we accompany the static figures with supplementary slide show files (e.g. Supplementary slide show 1 for Fig. 1A). The starting configuration was then gradually reduced in small steps of 20% in size and resolution (same procedure as in a previous study (Ullman et al., 2016)). At each step, we created reduced versions of the current configuration, namely five spatially reduced versions decreasing in size and resolution, as well as temporally reduced versions where a single frame was removed from the video configurations (Materials and methods). Each reduced version was then sent to Amazon's Mechanical Turk (MTurk), where 30 human subjects were asked to freely describe the object and action. MTurk workers tested on a particular video configuration were not tested on additional configurations from the same action type. Because of this restriction, approximately 4000 different MTurk users participated in all the behavioral tasks in this study. The success rates in recognizing the object and the action were recorded for each example. We defined a video configuration as recognizable if > 50% of the subjects described both the object and the action correctly.

The search continued recursively for the recognizable reduced versions, until it reached a video configuration that was recognizable, but all of its reduced versions (in either space or time) were unrecognizable. We refer to such a configuration as a 'minimal video'. An example of a minimal video is shown in Fig. 1B, and the reduced subminimal versions are shown in Fig. 1C–I. Most of the subjects (69%)

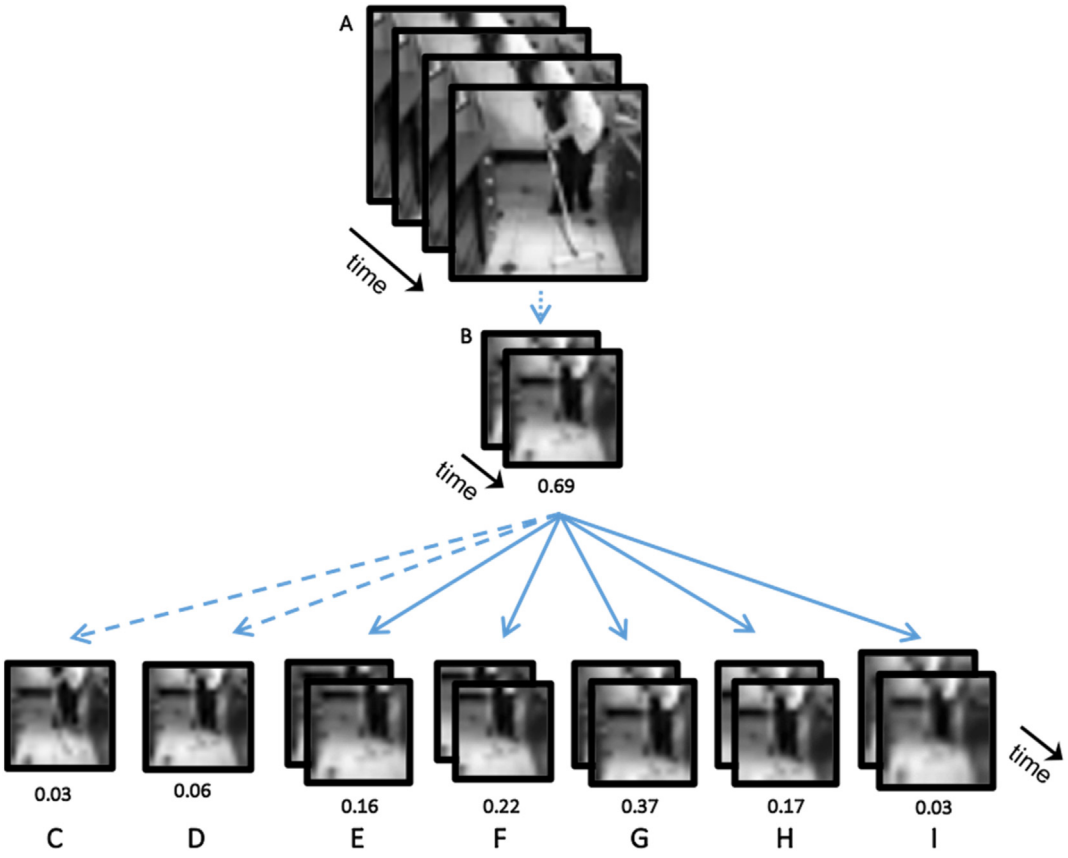

Fig. 1. Example of a minimal video.
A short initial video clip showing 'mopping' activity (A) was gradually reduced in both space and time to a minimal recognizable configuration (B) (Materials and methods). The numbers on the bottom of each image show the fraction of subjects who correctly recognized the action (each subject saw only one of these images). The spatial and temporal trimming was repeated until none of the spatially reduced versions (E–I, solid connections) or temporally reduced versions (C, D, dashed connections) reached the recognition criterion of 50% correct answers. Spatial reduced versions: In E each frame was cropped in the top-right corner, leaving 80% of the original pixel size in B. F, G, H are similar versions where the crop is on the top-left, bottom-right, and bottom-left corners, respectively, I is a version where the resolution of each frame was reduced to 80% of the frame in B. Temporal reduced versions: A single frame was removed, resulting in static frame#1 in C, and static frame#2 in D. See Supplementary file 'fig 1.mp4' or https://www.dropbox.com/s/nil8uyzxarkiadz/fig1.mp4?dl=0 for animated version of the dynamic configurations.

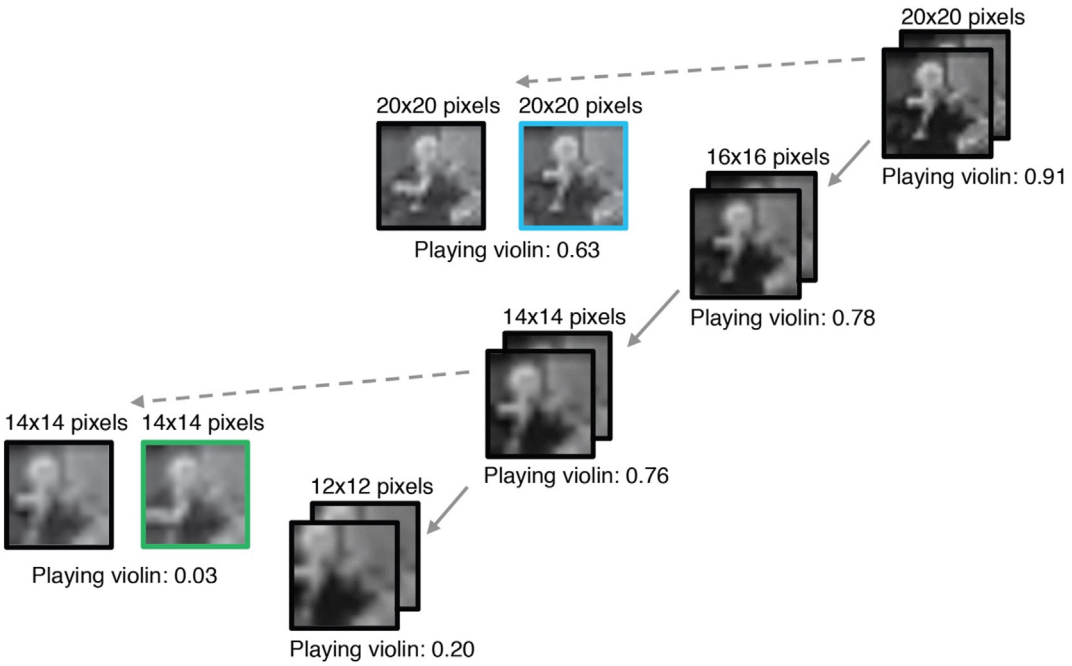

Fig. 2. Trade-off between spatial and temporal information.
Solid connectors represent spatially reduced versions, dashed connectors represent temporal reduced versions. The numbers below each configuration represent the fraction of subjects that correctly identified the action "playing violin". The temporally sub-minimal single-frame green configuration is not recognizable, but it becomes recognizable when more spatial information (i.e., more pixels) is added in the single-frame configuration in blue. The converse also holds: adding temporal information to a spatial sub-minimal configuration can recover performance (Fig. S2). See Supplementary file 'fig 2.mp4' or https://www.dropbox.com/s/ei5yaz6c6kaab3e/fig2.mp4?dl=0 for for animated version of the dynamic configurations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

were able to recognize the action ('mopping') in the video in Fig. 1B, consisting of two frames shown every 500 ms (2 Hz, the default frame rate used for all minimal videos). Showing each frame separately led to recognition rates of 3% and 6%, respectively (Fig. 1C–D, we refer to these as temporal sub-minimal configurations). As shown in Fig. 1B–D, the spatial content of the minimal and temporal sub-minimal configurations was very similar: there were only minor spatial content added from frame#1 to frame#2. Despite the similarity of the two frames, there was a large difference in human recognition due to the removal of the motion signal. Spatially cropping the video also led to a large drop in recognition (16–37%, Fig. 1E–H, we refer to these as spatial sub-minimal configurations). Keeping the number of pixels but blurring the video (reducing sampling distance by 20%) also led to a large drop in recognition (to 3%, Fig. 1I). As shown in Fig. 1E–H, in the tested cases the motion content of the minimal and (spatial) sub-minimal is very similar, that is, the pixels that are cropped out do not remove a large amount of image motion. This implies that the motion signal alone is not a sufficient condition for human recognition of minimal videos.

From the set of 18 original video snippets, we obtained 20 minimal videos similar to the one shown in Fig. 1. Four additional examples of minimal videos and their sub-minimal versions are shown in Fig. S1. A prominent characteristic of minimal video configurations was a clear and consistent gap in recognition between the minimal configurations and their sub-minimal versions. The mean recognition rate was 0.71 ± 0.11 (mean ± SD) for the 20 minimal video configurations (such as the one in Fig. 1B), 0.28 ± 0.15 for the spatial sub-minimal configurations (such as the ones in Fig. 1E–I), and 0.16 ± 0.14 for the temporal sub-minimal configurations (such as the ones in Fig. 1C–D). The difference in recognition rates between the minimal and sub-minimal configurations was statistically highly significant: $P < 4 \times 10^{-12}$ and $P < 6 \times 10^{-8}$, n = 20, one-tailed paired *t*-test, for the spatial and temporal sub-minimal configurations, respectively. The minimal videos included 2 frames of $n \times n$ pixels, where n = 20 ± 7.1 on average. Although highly reduced in size, the recognition rate for the minimal videos was high, and was not too far from the recognition rate of the original UCF101 video clips (mean recognition = 0.94 ± 0.07), even though the original clips had an average of 175 frames (versus 2 frames), each with 320 × 240 colored RGB pixels (versus the 2 grayscale frames of average size 20 × 20 pixels). Recognition rates for the minimal videos were also close to the recognition rates for the level above it in the search tree (the 'super minimal video configuration') (mean recognition = 0.81 ± 0.07).

In the temporally reduced single frames shown in Fig. 1C–D, there is an entire frame of spatial information missing. We asked whether the drop in recognition could be ascribed to the missing spatial information, without the need to combine information temporally. To evaluate this possibility, we introduced a condition where the two frames were presented side-by-side. The side-by-side simultaneous presentation of the two frames from the minimal configuration without the dynamics was not sufficient to improve recognition (mean performance 0.27 ± 0.17), and the gap between the side-by-side recognition rate and the maximal single frame recognition rate (mean recognition = 0.21 ± 0.14) was not statistically significant ($P > 0.05$, n = 20, one-tailed paired *t*-test).

The way selected here to reduce spatiotemporal configurations keeps the reduced configuration in the natural videos domain (albeit having gray-level and low-resolution frames). This is different than e.g., applying a sort of spatiotemporal noise such as spatiotempoal Fourier transform, which would return synthetic videos. As we argue below in the Discussion section, this aspect is important to explore the critical natural image features that exist in the minimal videos but not in the sub-minimal ones.

2.2. Spatial information can compensate for lack of temporal information and vice versa

Given that removing either spatial information or temporal information led to a large drop in recognition performance, we asked whether it is possible to compensate for lack of spatial information by adding more temporal information or, conversely, to compensate for the lack of temporal information by adding more spatial information. A temporal sub-minimal configuration (e.g., a single frame) became recognizable when more spatial information (i.e., more pixels) was added (e.g., Fig. 2). In the example in Fig. 2, 204 pixels were added (20×20 pixels versus 14×14 pixels), which was the maximum amount of pixels that needed to be added to make temporal sub-minimal videos recognizable across all the examples. Results were consistent across the $n = 6$ examples tested ($P < 9 \times 10^{-4}$, one-tailed *t*-test); the maximum improvement in recognition obtained upon adding spatial information to the temporally sub-minimal configurations was 0.66 ± 0.09. Similarly, a spatial sub-minimal configuration (e.g., two dynamic frames of smaller size) became recognizable when more temporal information (i.e., more frames) was added (Fig. S2). Spatial sub-minimal videos required only one additional frame to pass the recognition threshold. Results were consistent across the $n = 6$ examples tested ($P < 4 \times 10^{-3}$, one-tailed t-test); the maximum improvement in recognition obtained upon adding temporal information to the spatial sub-minimal configurations was 0.59 ± 0.10. Thus, there is a trade-off between spatial and temporal information and both dimensions can compensate for each other to aid recognition.

2.3. Action recognition in minimal videos is accompanied by full spatiotemporal interpretation

To test the conjecture that action recognition is accompanied by detailed interpretation of the image parts as well as spatiotemporal relations between parts, we ran a new series of experiments where subjects were instructed to describe internal components of the videos. MTurk subjects were presented with the minimal videos, along with a probe pointing to one of its internal spatial components. The probe could be either an arrow pointing to a frame region, or a contour separating two regions of the frame (Fig. S3).

We evaluated spatiotemporal interpretation in 5 minimal videos. We defined a component as 'recognized' if it was correctly labeled by > 50% of the subjects. For example, Fig. 3A (top) shows the results of the spatiotemporal interpretation experiment for the "mopping" minimal video: most of the subjects were able to correctly label the arm, legs, stick and vacuum. Average recognition for the 31 components that we evaluated was 0.77 ± 0.17 (see examples in Fig. 3). To assess whether the dynamic video configurations were necessary for interpretation, we repeated the experiment using the spatial sub-minimal and temporal sub-minimal versions, using the same procedure of inserting a probe in the frames. In contrast to the reports obtained from the minimal videos, subjects consistently struggled to recognize the parts in the sub-minimal videos. For example, the percentage of subjects that correctly identified the arm dropped from 75% to 48% or 38% and recognition of the stick dropped from 52% to 0% or 19% (Fig. 3A). We computed the gap in recognition rate for each component when it appeared in the minimal configuration versus when it appeared in its sub-minimal version. There was a significant decrease in component recognition for the spatial sub-minimal versions (difference in component recognition rates $= 0.41 \pm 0.22$, $P < 7 \times 10^{-9}$, $n = 31$, one-tailed paired *t*-test), as well as a significant decrease in component recognition for the temporal sub-minimal versions (difference in component recognition rates $= 0.29 \pm 0.20$, $P < 6 \times 10^{-9}$, $n = 31$, one-tailed paired *t*-test).

Interpretation of video components was not necessarily all-or-none. In some cases of partial interpretation, subjects could recognize the human body, or body parts, but could not recognize the action object and hence the activity type. In the example of 'Playing a Violin' in Fig. 3B, humans could recognize few body parts (e.g., the arm and the head) but not the action, from the sub-minimal configurations (lower panel), while in the minimal configuration (upper panel) they could identify a richer set of body parts, as well as the objects of action (i.e., the violin, the bow). The gap in recognition for object components was higher than that obtained for body components reported above: the mean recognition rate for 10 object parts was 0.61 ± 0.08 for the minimal videos, 0.21 ± 0.11 for the spatial sub-minimal videos ($P < 6 \times 10^{-5}$, $n = 10$, one-tailed paired *t*-test), and 0.11 ± 0.06 for the temporal sub-minimal videos ($P < 7 \times 10^{-8}$, $n = 10$, one-tailed paired *t*-test). In sum, action recognition in the minimal videos is accompanied by a rich description of the image components and how they interact to produce the action.

2.4. Existing computational architectures for action recognition fail to explain human behavior

To further understand the mechanisms of spatiotemporal integration in recognition, we tested current models of spatiotemporal recognition on our set of minimal videos, and compared their recognition performance to human recognition. Our working hypothesis was that minimal videos require integrating spatial and dynamic features, which are not used by current models. The tested models included the C3D model by Tran et al. (2015, 2018), the two-stream network model by Simonyan and Zisserman (2014), and the RNN-based model by Donahue et al. (2017), which have recently achieved a winning record on popular benchmarks for action classification in videos (e.g., the UCF-101 challenge), and which come from three different approaches to spatiotemporal recognition (namely, the 3D Convolutional Networks, the Two-Stream Networks, and RNN networks, respectively, as mentioned in the Introduction).

Our computational experiments included three types of tests with increasing amount of specific training, to compare human visual spatiotemporal recognition with existing models. In the first tests, models were pre-trained on the UCF-101 dataset for video classification. We tested such pre-trained models on our set of minimal videos, to explore their capability to generalize from real- world video clips to minimal configurations (see Materials and methods). Classification accuracy by the C3D model for minimal videos was significantly lower than the classification accuracy achieved for the original full video clips, from which we cropped the minimal videos (e.g., in testing four variants of the C3D model: $P < 4 \times 10^{-5}$, $n = 4$, one-tailed paired *t*-test). For the full videos, both humans and the model were able to correctly recognize the action (Fig. S4A–B), but for minimal videos there was a large gap between humans and the model performance: While 75% of humans correctly identified the action in the minimal video shown in Fig. S4C, the correct answer was not even among the top 10 for the model (Fig. S4D).

The models considered thus far had no training with the minimal videos (the same holds for the human subjects). Next, we evaluated whether training the models with minimal videos (fine-tuning) could help improve their performance. We used a binary classifier based on the convolutional 3D network model (C3D (Tran et al., 2015; Tran et al., 2018)), which was pre-trained on the SportM dataset: the network was originally trained on 1M video clips from 427 different sport actions (Karpathy et al., 2014). The C3D model does not have explicit mechanisms for predicting action trajectories into the future. We selected the C3D model for this task despite being less biologically plausible because it is a standard and common model in the computer vision literature. The network was then fine-tuned on a training set including 25 positive examples similar to a minimal video from a single category and type (the 'rowing' minimal video, see examples in Fig. 4A, all positive examples were validated as recognizable to humans), as well as 10,000 negative examples (e.g., Fig. 4B). Data balancing techniques were used to ensure that the results would not be biased by the

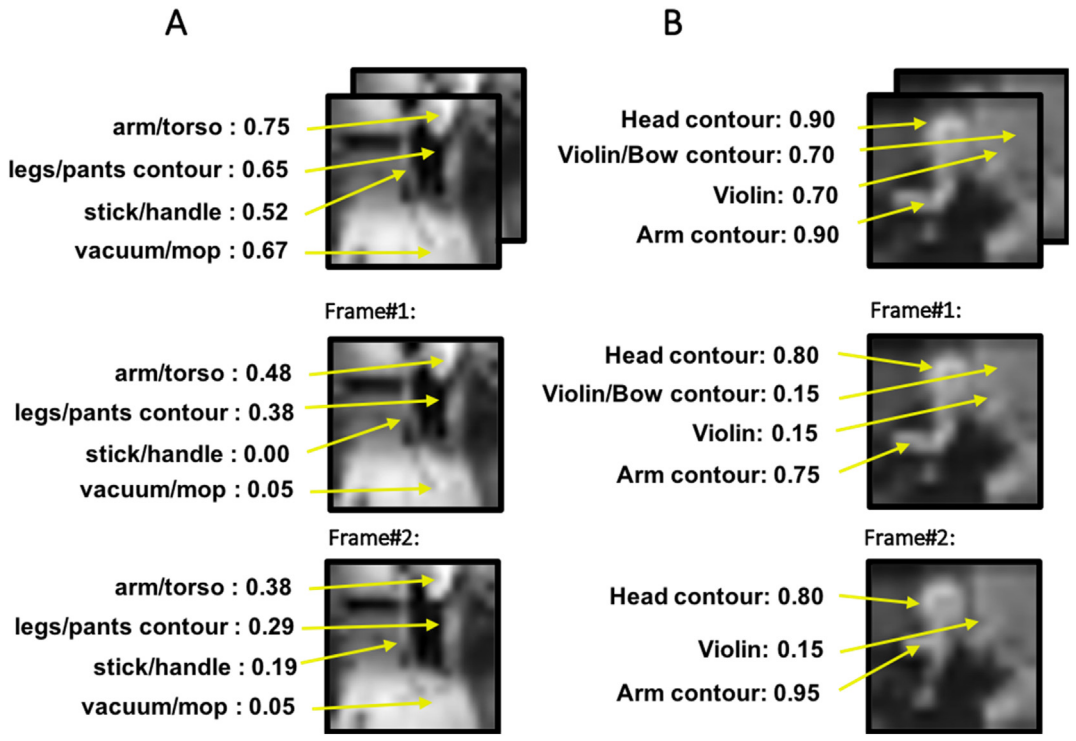

Fig. 3. Spatiotemporal interpretation.
When humans could recognize the object and action, they could also identify a set of internal components of the agent and the object of action (top). In contrast, humans could not recognize these internal components (or could partially recognize them) in the sub-minimal versions (bottom four panels). Here are some of the recognized semantic components of minimal video configurations for 'mopping' (in A) and 'Playing a violin' (in B). The numbers indicate the rate of correct identification of part, when human subjects were presented with the minimal configuration along with a probe pointing to the part location. Bolded entries indicate large differences between the minimal and sub-minimal configurations.

imbalance between positive and negative examples (see Appendix B). The binary classifier was then tested on a novel set of 10 positive examples and 5000 negative examples, similar to the ones used during training. Since our set of positive examples was constrained to specific body parts and specific viewing positions in 'rowing' video clips, the fine-tuned classifier was able to correctly classify most of the random negative examples; the average precision (AP, see Appendix D for term definition) was 0.94. Still, a non-negligible set of negative examples was given high positive score by the fine-tuned model, from which we composed a new set that we refer to as 'hard negative video configurations' for further analysis. The hard negative configurations included 30 examples of video configurations that were erroneously labeled by the fine-tuned network model (see examples in Fig. 4E). Comparing accuracy of human and network recognition for the set of hard negative configurations further revealed a significant gap: humans were not confused by any of the hard negative examples (AP = 1; Fig. S5C), while the fine-tuned network scored the hard negatives higher than most positive examples (AP = 0.18; Fig. S5F).

A distinctive property of recognition at the minimal level is the sharp gap between minimal and sub-minimal videos. We therefore further compared recognition by the binary CNN classifier and human recognition by testing whether the network model was able to reproduce the gap in human recognition between the minimal configurations and their spatial and temporal sub-minimal ones. For this purpose, we collected a set of minimal and sub-minimal videos showing a large gap in human recognition, which did not overlap with the training set for the network model. We tested the fine-tuned network model on a set containing 10 minimal videos, 20 temporal sub-minimal configurations (e.g., Fig. 4F), and 20 spatial sub-minimal configurations (as in Fig. 4D), all from the same category of 'rowing' in a similar viewing position and size. The network model was not able to replicate human recognition performance over this test set. While there was a clear gap in human recognition between minimal and spatial sub-minimal videos (average gap in human recognition rate 0.63; Fig. S5A), and between minimal and temporal sub-minimal videos (average gap in human recognition rate 0.68; Fig. S5B), the differences in recognition scores given by the network model for the minimal and sub-minimal examples were small (see Materials and methods). This discrepancy between human behavior and the models persisted even after using standard data augmentation and fine-tuning techniques. In sum, none of the tested models, even when fine-tuned with minimal videos, were able to account for human recognition of minimal videos.

2.5. Existing computational architectures do not integrate time and space cues the way humans do

The psychophysics results show that processing of minimal videos in the human visual system requires the combining of motion and spatial information. We next compared the combination of motion and spatial information by the human system and current CNN models (such as C3D) in the recognition of minimal videos. For this purpose, we compared the recognition of minimal and sub-minimal videos by two network models: (i) A purely spatial VGG19 network model, pre-trained on ImageNet and fine-tuned on frames of minimal videos (see Appendix C), and (ii) the C3D model, which is a spatiotemporal adaptation of the spatial VGG19 via 3D convolutional operations, pre-trained on ImageNet and UCF101 and fine-tuned on minimal videos. Our goal was to quantify the similarities and differences between the two models and human recognition on minimal videos, in order to understand the contribution of temporal processing in the C3D model compared with static VGG19 architectures as well as the contributions of temporal information to human behavior.

For the static VGG19 model, the recognition gap between 'rowing' minimal videos and spatial sub-minimal videos was 0.34 (Fig. S5G, see

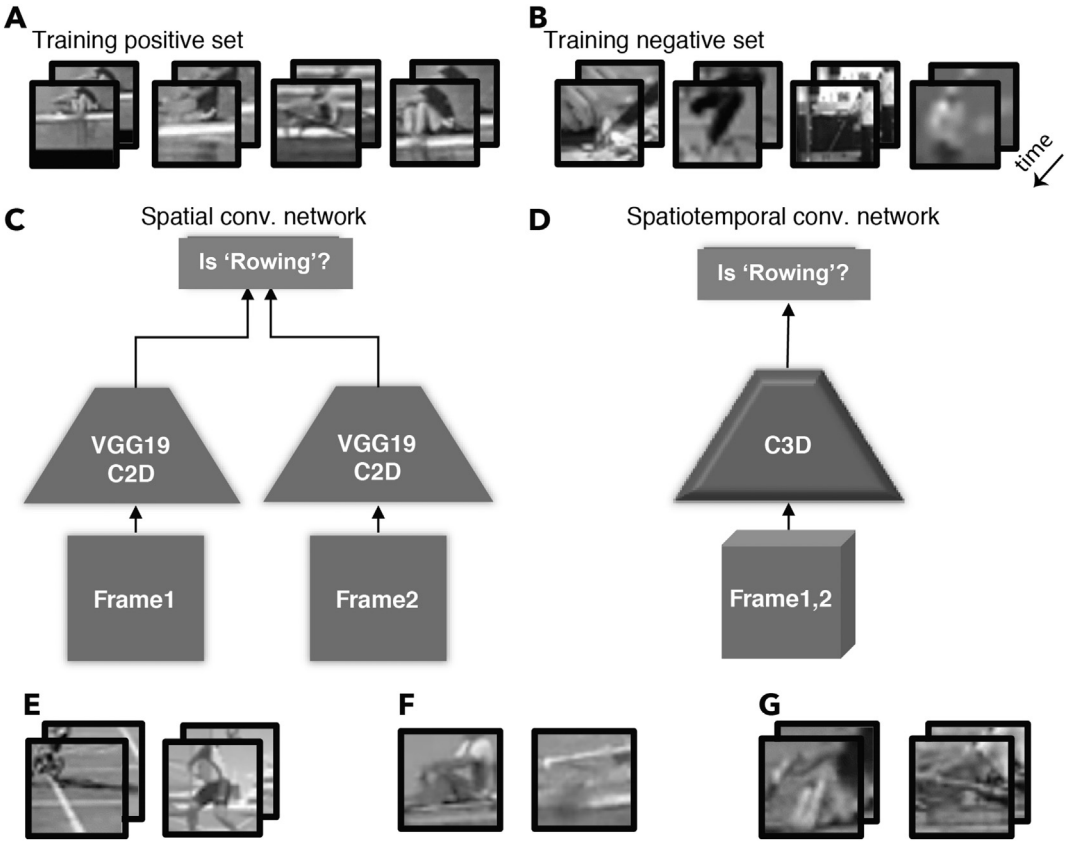

Fig. 4. Testing minimal videos with existing models for spatiotemporal recognition.
A–B. A binary classifier is trained to separate a positive set of similar minimal videos ("rowing"), showing the same action with the same body region and viewing position (A) from a negative set ("not rowing") including non-class videos of the same size and style as the minimal (B).
C. One type of binary classifier was based on CNNs with 2D convolutional filters, followed by taking the maximum detection score from each frame.
D. Another type of binary classifier was based on CNNs with 3D convolutional filters (Tran et al., 2015, 2018), which was fine-tuned with the positive and negative sets in A and B.
E–G. The binary classifiers could not replicate human recognition, and performance by 2D and 3D CNNs was similar. Six example configurations that were misclassified including two of the same size (E), two temporally sub-minimal (F) and two spatially sub-minimal (G). See Supplementary file 'fig 4.mp4' or https://www.dropbox.com/s/ei5yaz6c6kaab3e/fig2.mp4?dl=0 for animated version of the dynamic configurations.

Materials and methods for how we compare minimal vs. sub-minimal recognition gap between humans and models), which was smaller than the corresponding gap in human behavior (0.63, as mentioned above). For the dynamic C3D model, the gap between the temporal sub-minimal and the minimal videos was 0.37 (Fig. S5H), which was also very different from the corresponding human gap (0.68, as mentioned above). We also tested the VGG19 and C3D models on a set of hard-negative examples. For this analysis, we repeated the test described in the previous paragraph to generate hard negative examples for C3D, and collected a new set of 30 hard negative examples for the fine-tuned VGG19 model. Comparing human and VGG19 recognition for the set of hard negative examples showed a difference in recognition accuracy (AP = 0.64 for VGG19 whereas humans were not confused by any of the hard negatives, AP = 1, Fig. S5I). Besides the gap between humans and models, there was also a gap in recognition accuracy between the two models, namely VGG19 and C3D models (0.64 AP vs. 0.18 AP). This shows that VGG19 was better at rejecting hard negative examples, and in this aspect closer to humans than the C3D model.

To conclude, the test results show that VGG19 is better than C3D in replicating human behavior for spatial sub-minimal videos (recall gap: 0.34 for VGG19, 0.02 for C3D, and 0.63 for humans. See Appendix D for term definition) and for hard negative examples (AP = 0.64 for VGG19, 0.18 for C3D, 1 for humans), but the C3D is better than VGG19 in replicating human behavior for temporal sub-minimal examples (recall gap: 0.37 for VGG19, 0.78 for C3D, and 0.68 for humans). We suspect that the reason for the latter is that the C3D model is sensitive to basic dynamic features, which are not contained in our temporal sub-minimal configurations, and which the strictly spatial VGG19 model cannot capture. The more surprising point is that for the spatial sub-minimal configurations and the hard negative examples, the motion information that is added in the C3D contributes very little, if anything, to replicating human behavior. The different conditions and results above are summarized in Table 2. Since minimal videos are limited in their amount of visual information, and require efficient use of the existing spatial and dynamic cues, comparing their recognition by humans and existing models uncovers differences in the use of the available information. By using these configurations, the experimental results above point to fundamentally different mechanisms of integration of the available time and space information by humans and state-of-the-art network models.

3. Discussion

We generated *minimal videos* where human observers can readily identify both objects and actions but which become unrecognizable upon any reduction in the amount of either spatial or temporal information. Object and action recognition is accompanied by full interpretation of the different components of the video as well as the interaction between components. These minimal videos demonstrate a large discrepancy between humans and state-of-the-art computational

Table 2
A summary of test results comparing humans and computational models (C3D [Trun et al., 2018] and VGG19 [***]) on recognition of minimal videos. See test details and anlysis of results in the main text. *** Karen Simonyan, Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations (ICLR) 2015.

Tests comparing humans and computational models	Humans	C3D model (fine-tuned on minimal configurations)	VGG19 model (fine-tuned on minimal configurations)
Classifying minimal videos vs. 'hard' non-class examples	Average Precision = 1	Average Precision = 0.18	Average Precision = 0.64
Recognizing minimal vs. spatial sub-minimal configurations	Recall gap = 0.68	Recall gap = 0.78	Recall gap = 0.37
Recognizing minimal vs. temporal sub-minimal configurations	Recall gap = 0.63	Recall gap = 0.02	Recall gap = 0.34

models, emphasizing that current models lack the fundamental interactions between spatial and temporal features that characterize human perception.

The minimal videos contained a mixture of both static spatial features (e.g., the legs and torso of the person playing the violin do not change in time, Fig. 2) and dynamic temporal features (e.g., the hand and bow are moving). Both spatial and temporal features are crucial for human recognition and interpretation, as revealed by the sharp transition to unrecognizable spatial and temporal sub-minimal videos. Previous works have shown how moving features alone can be sufficient for action recognition; for example, all features are moving in biological motion studies (Blake & Shiffrar, 2007) and in the slit experiments (Morgan et al., 1982). In biological motion the spatial information is very limited (when viewed without motion), namely simple unconnected and meaningless dots. Many previous studies have also shown that static features can be sufficient for action recognition (Ben-Yosef & Ullman, 2018; Yao et al., 2011). In contrast to the distinction between dynamic and static features suggested by those previous studies, here we show that recognition (and interpretation) is not divided into two separate channels, one for motion-based recognition, the other static: a particular mix of spatial and temporal features drives recognition and interpretation of minimal videos. Furthermore, adding spatial information can compensate for the lack of temporal information and vice versa, thus reconciling the current work with previous studies that provided only either spatial or temporal features with overcomplete information.

A known role of temporal features in scene understanding is to provide the dynamical aspects of objects in the scene. For example, a 'hand touching a box' can already be recognized in each individual frame in a sequence; however, a sequence of the hand and box objects in motion is required to recognize the action 'moving a box'. Much of the computational vision literature has focused on this aspect of dynamics – the motion trajectories associated with objects that can be identified statically (Blank, Gorelick, Shechtman, Irani, & Basri, 2005; Cheron et al., 2015). Minimal videos identify natural images that must have dynamics, as well as specific spatial cues, to allow recognition and interpretation by humans. These spatiotemporal configurations can thus be used to study the mechanisms subserving integration of spatial and temporal information, and the trade-off in human visual processing, between static and motion cues during visual recognition.

State-of-the-art deep learning models failed to capture human recognition and interpretation of minimal videos, even when those models were fine-tuned for the specific tasks evaluated here by training them with similar minimal configurations. A mere gap between existing DNN models and human cognition should not be surprising by itself, of course. However, studying and characterizing this limitation of existing models is important because it motivates further investigation of spatiotemporal features and computational recognition models (and in particular neural network models) that can better predict human behavior. The minimal videos provide a tool to study critical spatiotemporal features, as well as space-time dependency, by exploring the differences between the recognizable minimal video configurations and their slightly reduced but unrecognizable sub-minimal versions. These sharp differences hint to the type of critical features in recognition of minimal videos, which could be more cognitive, high-level features, rather than low-level visual features that disrupt basic perception. Fig. 2 demonstrates this point: frames that include the violinist's right elbow are recognizable, while frames with partial view of the elbow are not. Adding motion cues, even where the elbow is partially visible, improves recognition. Other body parts, such as the head or bow, are still fully visible but not recognizable in sub-minimal videos. Interestingly, the specific mix of features in the sub-minimal videos is not sufficient for recognition.

Future studies could extend recent modeling of full interpretation of spatial minimal images (Ben-Yosef et al., 2018; Ben-Yosef & Ullman, 2018), to the modeling of full spatiotemporal interpretation. More specifically, full spatiotemporal interpretation can be achieved by a structural learning approached in which configurations of parts assignments and their associated spatiotemporal properties and relations are explored and matched to stored configurations of action categories. The spatiotemporal features and relations can be complex: e.g., spatiotemporal relations of 'parts containment' or 'contours parallelism' (Ben-Yosef et al., 2018; Ben-Yosef & Ullman, 2018) that characterizes the differences in interpretation between the minimal and sub-minimal videos. Triggering these more complex features will therefore be done in a selective, top-down manner, and complementary to a first-stage forward aggregation of spatiotemporal filters (Jhuang et al., 2007; Tran et al., 2015; Tran et al., 2018). Focusing on a new type of models that cannot only label a particular action, but can also perform spatiotemporal interpretation will lead to a better understanding and more accurate modeling of spatiotemporal integration and human recognition.

4. Materials and methods

4.1. Setting initial video configuration

The normalized frame size, the frame rate, and presentation as animated GIF. The initial video configuration was created as follows: we selected 2 to 5 frames from the original video clip, from which the action and object were recognizable to the MTurk users, according to our criterion, and normalized their frame size to 50 × 50 image samples (pixels) and to gray level colors. We then built a video configuration in which the selected normalized frames repeat in a loop at a fixed frame rate of 2 frames/s (2 Hz). The configuration was presented as animated GIF format. The choice of 2 Hz frame rate was made since it provided the best recognition accuracy by the MTurk users.

4.2. Testing pre-trained networks on minimal videos

Our test set included 20 minimal videos, from 9 different human action categories: Biking, Rowing, Playing violin, Playing flute, Playing Tennis, Playing Piano, Mopping, Cutting, and Typing. The accuracy for all the models was low: top-1 average accuracy was 0/20 for a C3D

deep convolutional network based on ResNet-18 (Hara et al., 2018), and 1/20 for a C3D deep convolutional network based on ResNet-101 (Hara et al., 2018) (see Appendix A for implementation details). Although humans were only given one chance for labeling the video sequences, several studies in the computer vision literature report top-5 accuracy (a label is considered to be correct if any of the top 5 labels is correct). The average top-5 accuracy was 0.10 for C3D based on ResNet-18, and 0.20 for the C3D based on ResNet-101 (algorithms based on the two-stream network, and the RNN-based model did not provide better results, see Appendix A).

4.3. Comparing minimal vs. sub-minimal recognition gap between humans and models

To compare the model and human recognition gap, we set the acceptance rate of the binary classifier to match the average human recognition rate (e.g., 78% of the minimal videos for 'rowing'), and then compared the percentage of the minimal vs. spatial sub-minimal configurations that exceeded the network-based classifier's acceptance (hereinafter the network 'recall'; a similar method was used in previous work (Ullman et al., 2016)). For the C3D model, the recall gap between 'rowing' minimal configurations and spatial sub-minimal configurations was 0.02 (see Fig. S5D), which is far from the recognition gap observed in humans. To test temporal sub-minimal configurations, we composed spatiotemporal configurations containing one frame from the minimal configuration, and a noise frame. The reason for this construct is that configurations with zero dynamics are trivially rejected by the C3D model. Nevertheless, distinguishing between the 'rowing' temporal sub-minimal and the minimal configurations was less difficult for the C3D model, with a recall gap of 0.78 (see Fig. S5E). All temporal sub-minimal configurations received a very low recognition score by the C3D model, which was close to the human gap.

Funding

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, NSF grant 1745365, NIH grants R01EY026025, 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. SU was supported by EU Horizon 2020 Framework 785907 and Israel Science Foundation grant 320/16.

Author contributions

The ideas and experiments were jointly developed by GBY, GK and SU. All the experiments, analyses and computational models were implemented by GBY. The manuscript was written by GBY with the help and edits of GK and SU.

Declaration of competing interest

The authors declare that they do not have any conflict of interest.

Data and code availability

The data that support the finding of this study and the computer code are available at https://github.com/guybenyosef/introducing_minimal_videos.git.

Appendix A. Testing pre-trained network models on minimal videos (implementation details)

For 3D convolutional networks, we used the implementations by Hara et al. (2018), based on Resent-18 and Resnet-101, which are currently the leading architectures in the UCF101 challenge. The models were pre-trained on the very large Kinetics dataset by Kay et al. (2017), and then fine-tuned for the UCF101 benchmark. For two-stream network we used the implementation by Feichtenhofer, Pinz, and Wildes (2016), based on Resset-50. The model was pre-trained on ImageNet, and then fine-tuned on the UCF101 benchmark. For the RNN-based model we used the implementation by Donahue et al. (2017). Frames are input to layer of CNNs (based on AlexNet), then input to layer of LSTMs, scored by averaging across all video frames.

Appendix B. Negative examples for classification of minimal spatiotemporal configuration

10,000 negative examples were collected containing video configurations of a similar frame size and frame length as the positive set (minimal videos of the same class and type, e.g., 'rowing' as in Fig. 4A), but taken from different categories (i.e., non-'rowing') video clips (e.g., Fig. 4B). This asymmetry in size of positive and negative sets is due to the observation that negative examples were easier to find and to test psychophysically than the positive examples. Despite this asymmetry, a large set of negative examples can still contribute to the training process of deep CNNs (Goodfellow, Bengio, & Courville, 2016) when using standard data balancing techniques. In our case the technique that worked best was to duplicate the number of positive examples, such that the number of positive and negative examples is roughly even. Data augmentation techniques such as rotating and/or flipping the image examples were not used since such techniques often turn the minimal image un-recognizable to humans.

Appendix C. Constructing spatial VGG19 model for recognizing minimal videos

The spatial VGG19 model was constructed as a binary classifier (based on the pre-trained ImageNet version), which was fine-tuned on all frames from the positive and negative video examples in the train set for the C3D mentioned above. When a novel video configuration example was given to the VGG19, we applied the VGG19 network separately to each frame, and considered the maximal VGG score for the frames as the final returned recognition score. We tested the VGG19 on the three test sets mentioned above for the C3D, and then compared results for the VGG19 and C3D convolutional networks.

Appendix D. Using average precision and recall for evaluating model classification

To evaluate performance of classification models (namely the models used here for action classification from videos) we use average precision metric, which is popular in the evaluating image and video classification models in the machine learning literature. Precision is a term describing the fraction of true positive classifications out of all positive classifications made by the model. Recall is the fraction of true positive classifications out of all the positive examples in the dataset. Average precision (AP) is a measure that combines recall and precision for ranked classifications. It computes the average precision value for recall value over 0 to 1, namely after each example in the dataset is classified by the model. High AP value would then mean good classification of the dataset by the model, even if the dataset is not balanced (i.e., dataset does not have roughly equal number of positive and negative examples). Gap in recognition performance between two recognition systems can be measured by the gap between the AP (or only recall/precision) that is calculated for each model on the same dataset.

Appendix E. Supplementary data

Supplementary materials archived include 5 videos showing animated versions of the main and supplementary figures in the paper, and

raw data including the minimal and sub-minimal video files and statistics of MTurk human test results. Supplementary data to this article can be found online at https://github.com/guybenyosef/introducing_minimal_videos.git, https://doi.org/10.1016/j.cognition.2020.104263.

References

Anstis, S. M. (1970). Phi movement as a subtraction process. *Vision Research, 10*, 1411–1430.

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

Ben-Yosef, G., & Ullman, S. (2018). Image interpretation above and below the object level. *Journal of the Royal Society Interface Focus, 8*(4).

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annual Review of Psychology, 58*, 47–73.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005). *Actions as space-time shapes. Paper presented at the IEEE International Conference on Computer Vision*.

Casile, A., & Giese, M. A. (2005). Critical features for the recognition of biological motion. *Journal of Vision, 5*(4), 348–360.

Cavanagh, P., Labianca, A. T., & Thornton, I. M. (2001). Attention-based visual routines: Sprites. *Cognition, 80*(1–2), 47–60.

Cheron, G., Laptev, I., & Schmid, C. (2015). P-cnn: Pose based cnn features for action recognition. *IEEE International Conference on Computer Vision*.

Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., & Darrell, T. (2017). Long-term recurrent convolutional networks for visual recogniton and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Feichtenhofer, C., Pinz, A., & Wildes, R. (2016). Spatiotemporal residual networks for video action recognition. *Advances in neural information processing systems*.

Feichtenhofer, C., Pinz, A., & Zisserman, A. (2016). *Convolutional two-stream network fusion for video action recognition. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Giese, M. A., & Poggio, T. (2003). Cognitive neuroscience: Neural mechanisms for the recognition of biological movements. *Nature Reviews Neuroscience, 4*(3), 179–192.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.

Hara, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet. *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6546–6555). .

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Hur, J., & Roth, S. (2016). Joint optical flow and temporally consistent semantic segmentation. *European Conference on Computer Vision*.

Jhuang, H., Serre, T., Wolf, L., & Poggio, T. (2007). A biologically inspired system for action recognition. *IEEE International Conference on Computer Vision*.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14*, 201–211.

[dataset]Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. *IEEE Conference on Computer Vision and Pattern Recognition*.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., & Suleyman, M. (2017). The kinetics human action video dataset. arXiv, 1705.06950.

Kellman, P. J., & Cohen, M. H. (1984). Kinetic subjective contours. *Perception & Psychophysics, 35*, 237–244.

Kundu, A., Vineet, V., & Koltun, V. (2016). Feature space optimization for semantic video segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*.

Morgan, M. J., Findlay, J. M., & Watt, R. J. (1982). Aperture viewing: A review and a synthesis. *The Quarterly Journal of Experimental Psychology Section A, 34*, 211–233.

Oram, M., & Perrett, D. (1996). Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. *Journal of Neurophysiology, 76*(109–129).

Parks, T. E. (1965). Post-retinal visual storage. *The American Journal of Psychology, 78*, 145–147.

Perrett, D., Smith, P., Mistlin, A., Chitty, A., Head, A., Potter, D., ... Jeeves, M. (1985). Visual analysis of body movements by neurones in the temporal cortex of the macaque monkey: A preliminary report. *Behavioral Brain Research, 16*, 153–170.

Potter, M. C., & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology, 81*, 10–15.

Rock, I. (1981). *Anorthoscopic perception.* Scientific American.

Sáry, G., Vogels, R., & Orban, G. A. (1993). Cue-invariant shape selectivity of macaque inferior temporal neurons. *Science, 2605*, 995–997.

Schindler, K., & Van Gool, L. (2008). Action snippets: How many frames does human action recognition require? *IEEE Conference on Computer Vision and Pattern Recognition*.

Schofield, A., Gilchrist, I., Bloj, M., Leonardis, A., & Bellotto, N. (2018). Understanding images in biological and computer vision. *Interface focus, 8*, 20180027.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*.

Singer, J. M., & Kreiman, G. (2014). Short temporal asynchrony disrupts visual object recognition. *Journal of Vision, 14*(5).

Singer, J. M., Madsen, J. R., Anderson, W. S., & Kreiman, G. (2015). Sensitivity to timing and order in human visual cortex. *Journal of Neurophysiology, 113*(5), 1656–1669.

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv, 1212.0402.

Thurman, S. M., & Grossman, E. D. (2008). Temporal "bubbles" reveal key features for point-light biological motion perception. *J Vis, 8*(3), 28 (21-11).

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). *Learning spatiotemporal features with 3d convolutional networks. Paper presented at the IEEE International Conference on Computer Vision*.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6450–6459). .

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*, 2744–2749.

Vaina, L., Solomon, J., Chowdhury, S., Sinha, P., & Belliveau, J. (2001). Functional neuroanatomy of biological motion perception in humans. *Proceedings of the National Academy of Sciences of the United States of America, 98*, 11656–11661.

Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L., & Fei-Fei, L. (2011). Human action recognition by learning bases of action attributes and parts. *IEEE International Conference on Computer Vision* (pp. 1331–1338). .

Zollner, F. (1862). Über eine neue Art anorthoskopischer Zerrbilder. *Ann. Phys. 193*(11), 477–484.