CENTER FOR Brains Minds+ Machines

CBMM Memo No. 94 **November 21, 2018**

Spatiotemporal interpretation features in the recognition of dynamic images

Guy Ben-Yosef , Gabriel Kreiman , Shimon Ullman

Abstract

Objects and their parts can be visually recognized and localized from purely spatial information in static images and also from purely temporal information as in the perception of biological motion. Cortical regions have been identified, which appear to specialize in visual recognition based on either static or dynamic cues, but the mechanisms by which spatial and temporal information is integrated is only 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 the identification of minimal spatiotemporal configurations: these are short videos in which objects and their parts, along with an action being performed, can be reliably recognized, but any reduction in either space or time makes them unrecognizable. State-of-the-art computational models for recognition from dynamic images based on deep 2D and 3D convolutional networks cannot replicate human recognition in these configurations. Action recognition in minimal spatiotemporal configurations is invariably accompanied by full human interpretation of the internal components of the image and their inter-relations. We hypothesize that this gap is due to mechanisms for full spatiotemporal interpretation process, which in human vision is an integral part of recognizing dynamic event, but is not sufficiently represented in current DNNs.

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

Spatiotemporal interpretation features in the recognition of dynamic images

Guy Ben-Yosef[1,4] , Gabriel Kreiman[2,4] , Shimon Ullman[3,4]

1. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
2. Children's Hospital, Harvard Medical School, Boston ,MA 021155 ,USA
3. Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 7610001, Israel
4. Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

Text statistics: Number of figures: 4, Number of supplementary figures: 6, Number of supplementary. tables: 2, Number of words in abstract: 218, Number of words in main text: 6073

Abstract:

Objects and their parts can be visually recognized and localized from purely spatial information in static images and also from purely temporal information as in the perception of biological motion. Cortical regions have been identified, which appear to specialize in visual recognition based on either static or dynamic cues, but the mechanisms by which spatial and temporal information is integrated is only 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 the identification of minimal spatiotemporal configurations: these are short videos in which objects and their parts, along with an action being performed, can be reliably recognized, but any reduction in either space or time makes them unrecognizable. State-of-the-art computational models for recognition from dynamic images based on deep 2D and 3D convolutional networks cannot replicate human recognition in these configurations. Action recognition in minimal

spatiotemporal configurations is invariably accompanied by full human interpretation of the internal components of the image and their inter-relations. We hypothesize that this gap is due to mechanisms for full spatiotemporal interpretation process, which in human vision is an integral part of recognizing dynamic event, but is not sufficiently represented in current DNNs.

Introduction

Previous behavioral work has shown that visual recognition can be achieved on the basis of spatial information alone[1,2], and on the basis of motion information alone, as in biological motion[3]. At the neurophysiological level, neurons have been identified that respond selectively to objects and event based on purely spatial information, or motion information alone[4-7]. But 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[8-11] suggest that both shape and motion cues may cooperate to help recognition, but whether or how they may be integrated remain unclear. Studies on perceptual organization from visual dynamics (e.g., dynamic grouping and segmentation from motion[12]; spatiotemporal continuation and completion[13]) 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[14] and that visually selective neurophysiological signals are sensitive to this temporal information[15].

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 et al[16]; Kinetics dataset by Kay et al[17]). Modern models for action recognition from spatiotemporal input are based on deep network features, and in terms of combining spatial and temporal information they are partitioned into the following three groups: (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[18-21], 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 pre-trained static recognition network on ImageNet), and a second module is trained on optical flow from consecutive frames[22-24]. Here, the integration of temporal and spatial features takes place at a subsequent, higher stage, whereas in human vision motion has also a low-level role such as in figure-ground segmentation. (iii) Models combining deep convolutional networks with Long Short-Term Memory[25] units based on recurrent connections[26]. 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 spatio-temporal information. In order to evaluate the use of spatio-temporal 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[23,24], and therefore those stimuli are not ideally set up to rigorously test spatiotemporal integration. As elaborated below, an important aspect of using spatio-temporal 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. Existing schemes for spatiotemporal interpretation use direct extensions of static semantic segmentation techniques[27-29], which do not provide the full human-like spatiotemporal interpretation.

Here we sought to develop 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*, 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) would turn 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 MIRCs[2,30,31], 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[2], 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[30]. 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[30,31]. 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 configurations with both its 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 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 spatiotemporal configurations. Finally, the results suggest how computational models may be extended to better capture human performance.

Results

We first describe psychophysical experiments to find minimal spatiotemporal configurations in short video clips taken from computer vision datasets, and report how human behavior changes when varying critical dynamic parameters such as the frame rate in these configurations. We then describe human spatiotemporal interpretation of minimal configurations, including the identified components within the minimal configurations. Finally, we test existing computational models for recognition from spatiotemporal input on our set of minimal configurations, and we compare the models' results with human recognition.

A search for minimal spatiotemporal configurations

The search for each minimal spatiotemporal configuration started from a short video clip, taken from the UCF101 dataset[16], 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 S1 in the supplementary file for a full list). The original video snippets were reduced to a manually selected 50x50 pixel square region, cropped from 2 to 5 sequential non-consecutive frames, and taken at the same positions on each frame (see below for frame region selection). These regions served as the starting configurations in the search for minimal spatio-temporal configurations described below. In the default condition, frames were presented dynamically in a loop at a fixed frame rate of 2Hz (Methods). An example of a starting configuration and a minimal spatiotemporal configuration is shown in Figure 1 and the path to create it is illustrated in Figure 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 presented at a temporal interval of Δt (mean $\Delta t = 200msec \pm 100msec$, which encompasses the range of time interval to complete a natural body movement in the video clips that we considered, e.g., to lift a hand, etc.). An illustration of the starting configuration 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 pps 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[2]). At each step, we created reduced versions of the current configuration, namely five spatially reduced versions decreasing size and resolution, as well as temporally reduced versions where a single frame was removed from the spatiotemporal configurations (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 who tested on a particular spatiotemporal configuration were not tested on additional configurations from the same action type (thus we needed approximately 4000 different MTurk users to complete 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 spatiotemporal configuration as recognizable if more than 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 spatiotemporal configuration that was recognizable, but all of its reduced versions (in either space or time) were unrecognizable, and we refer to such a configuration as a 'minimal spatiotemporal configuration'. An example of a minimal spatiotemporal configuration is shown in Fig. 1B, and the reduced sub-minimal versions are shown in Fig. 1C-I. Most of the subjects (69%) were able to recognize the action ('mopping') in the spatiotemporal configuration in Fig. 1B, consisting of two frames shown every 500 ms (2 Hz, the default frame rate used for all minimal configurations). 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. 1C-D (and Fig. S1), in the cases tested the spatial content of the minimal and (temporal) sub-minimal configurations is very similar (namely only minor spatial content is added to frame#1 by frame#2). Yet a large difference in human recognition is recorded due to the motion signal. Image crop 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 image (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 (namely, the pixels that are cropped out do not cut off significant image motion). This implies that the motion signal alone is not a sufficient condition for human recognition of minimal spatio-temporal configurations.

From the set of original video snippets, we searched for 20 minimal spatiotemporal configurations similar to the one shown in Fig. 1. Four additional examples of minimal spatiotemporal configurations and their sub-minimal versions are shown in Fig. S1. A prominent characteristic of minimal spatiotemporal configurations was a clear and consistent gap in human recognition of the minimal configurations, compared to their sub-minimal versions. The mean recognition rate was 0.71±0.11 (mean±SD) for the 20 minimal spatiotemporal configurations (such as the one in Fig. 1B), 0.29±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 were statistically highly significant: $P < 3.08 \times 10^{-12}$ and $P < 5.16 \times 10^{-08}$, n=20, one-tailed paired *t test*, for the spatial and temporal sub-minimal configurations, respectively. The minimal spatiotemporal configurations included 2 frames of *n* x *n* pixels, where $n = 20 \pm 7.1$ on average. Although highly reduced in size, the recognition rate for the minimal spatiotemporal configurations was high, and not far from the recognition rate of the original UCF101 video clips (mean recognition was 0.94±6.7 for the original UCF101 video clips, an average of 175 frames, each with 320x240 colored RGB pixels versus the 2 grayscale frames of average size 20x20 pixels). Recognition rates for the minimal spatiotemporal configurations was also close to the recognition rates for the level above it in the search tree (the ‘super minimal configuration’: mean recognition was 0.81±0.74).

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 0.21±0.14) was not statistically significant ($P > 0.05$, n=20, one-tailed paired *t test*).

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 (Fig. 2). 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). This trade-off between spatial and temporal information was consistent for all the tested minimal configuration examples. In the example in Figure 2, 204 pixels were added (20x20 pixels versus 14x14 pixels), which was the maximum amount of pixels that needed to be added to make temporal sub-minimal images recognizable across all the examples. Spatial sub-minimal images required one additional frame to pass the recognition threshold. (within this range, maximal recognition of the sub-minimal with additional pixels, i.e., the case where improvement was highest, was 0.66±0.09, and of sub-minimal images with additional frame 0.59±0.10. These are significant improvements of the average recognition of the spatial sub-minimal, and temporal sub-minimal, as reported above. $P < 3.04 \times 10^{-3}$, and $P < 8.38 \times 10^{-4}$, n=6, one-tailed paired *t test*, respectively).

The frame rate impacts recognition of minimal spatiotemporal configurations

Linking two or more frames for recognition, requires temporal integration of dynamic information. We conjectured that the degree of temporal integration would be dependent on temporal spacing between the frames. The results presented thus far were based on a fixed frame rate (2 Hz) and a fixed frame duration (500 milliseconds), based on pilot experiments. Next, we investigated the dependence of recognition on the presentation rate. The dependence of recognition on frame rate could be used to infer the role of motion frequency as a component of natural dynamic recognition. We conducted further psychophysics experiments by creating modified versions of the minimal spatiotemporal configurations in which we varied the frame rate from 0.5 Hz to 8 Hz (Figure S4). Examples for such modified configurations are shown in Fig. S4B (dynamic version shown in Supplementary Slide Show S4). There was a significant difference in human recognition of the modified configurations for different frame rates ($P \leq 0.003$, n=5, one-way *ANOVA*). Recognition rates dropped when the frame rate was reduced from the default of 2 Hz and there was a lesser drop for higher frame rates (Figure S4A). We interpret these results to imply that too slow a presentation impairs temporal integration and essentially recapitulates the temporally sub-minimal condition where the two frames are presented separately or side-by-side.

There was a slight but noticeable dependence of the optimum frame rate on the specific action type of image tested. Some spatiotemporal configurations were highly recognizable for one of the tested frame rates but recognition dropped drastically as frame rate changed towards either higher or lower rates (e.g., Fig. S4B). In some cases, there was a phenomenon of 'dramatic pairs' showing large recognition drop between two spatiotemporal configurations with identical frames but different frame rates. As examples, for 'playing a flute' recognition rate was 0.65 when shown in frame rate of 4Hz but only 0.37 when shown in 8Hz. For 'Biking' recognition rate was 0.71 when shown in frame rate of 2Hz, but only 0.37 when shown in 1Hz. Still, we note that further investigation is required to quantify the dependence on the action in frame-rate require, which is left for further research.

Action recognition in minimal images is accompanied by full image interpretation

We conjectured that when humans correctly recognize the action in the minimal spatiotemporal configuration, 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 in the image sequence (a similar case of identifying parts and relations was shown in static minimal images[30]). We refer to this detailed understanding of the image as 'spatiotemporal interpretation'. To test this conjecture, we ran a new series of experiments where subjects were instructed to describe internal components of the images. MTurk subjects were presented with the minimal spatiotemporal configurations, along with a probe pointing to one of its internal 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 image interpretation in 5 minimal spatiotemporal configurations and tested with MTurk users. We defined a component as 'recognized' if it was correctly labeled by more than 50% of the subjects. Average recognition for the 31 components that we evaluated was 0.77±0.17 (see examples in Fig. 3). To assess whether the dynamic spatiotemporal configurations were necessary for interpretation, we repeated the experiment using the sub-minimal spatial and temporal versions, using the same procedure of inserting a probe in the images. 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±0.22, $P \leq 6.8 \times 10^{-9}$, 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±0.20, $P \leq 5.2 \times 10^{-9}$, n=31, one-tailed paired *t test*). An example of image interpretation for the "mopping" action is shown in Fig. 3A (upper panel). Subjects could identify the action (mopping), the presence of a person, and also the internal parts of the person figure, such as the legs, the internal parts of the object of action, namely the mop stick and the mop head. In contrast, none of these internal parts could be reliably identified in the reduced temporal and spatial sub-minimal versions, when one frame was removed (Fig. 3A, lower panel), or when the frames were slightly cropped.

Interpretation of image 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) 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 all components reported above: the mean recognition rate for 10 object parts was 0.61±0.08 for the minimal spatiotemporal configuration, 0.21±0.11 for the spatial sub-minimal configuration ($P \leq 5.5 \times 10^{-5}$, n=10, one-tailed paired *t test*), and 0.11±0.06 for the temporal sub-minimal configuration ($P \leq 6.3 \times 10^{-8}$, n=10, one-tailed paired *t test*).

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 spatiotemporal configurations, and compared their recognition performance to human recognition. Our working hypothesis was that minimal dynamic configurations require integrating spatial and dynamic features, which are not used by current models. The tested models included the C3D model by Tran et al[19,20], the two-stream network model by Simonyan & Zisserman[22], and the RNN-based model by Donahue et al[26], 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 spatiotemporal configurations, to explore their capability to generalize from real-world video clips to minimal configurations. Our test set included 20 minimal spatiotemporal configurations, 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[21], and 1/20 for a C3D deep convolutional network based on ResNet-101[21] (see Methods 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 Methods). These recognition rates are significantly lower than the classification accuracy achieved by these models for the original full video clips, from which we cropped the minimal configurations ($P \leq 3.8 \times 10^{-5}$, n=4, one-tailed paired *t test*)). An example comparing humans and the C3D model for a minimal spatiotemporal configuration is shown in Fig. S5. The correct answer is not among the top 10 in this case.

The models considered thus far had no training with the minimal configurations (the same holds for the human subject). Next, we evaluated whether training the models with minimal spatiotemporal configurations (fine-tuning) could help improve their performance. We used a binary classifier based on the convolutional 3D network model (C3D[19,20]), which was pre-trained on the SportM dataset: the network was originally trained on 1M video clips from 427 different sport actions[18]. The network was then fine-tuned on a training set including 25 positive examples similar to a minimal spatiotemporal configuration from a single category and type (the ‘rowing’ minimal configuration, see examples in Fig. 4A. All positive examples were validated as recognizable to humans), as well as 10000 negative examples (e.g., Fig. 4B. See methods). The binary classifier was then tested on a novel set of 10 positive examples and 5000 negative examples, similar to the ones in 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) was 0.941. 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 spatiotemporal configurations’ for further tests. The hard negative configurations included 30 examples of spatiotemporal 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; see Fig. S6-C), while the fine-tuned network scored the hard negatives higher than most positive examples (AP=0.18; See Fig. S6-F).

A distinctive property of recognition at the minimal level is the sharp gap between minimal and sub-minimal images. We therefore further compare recognition by the binary CNN classifier and human recognition, we tested 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 dynamic configurations 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 configurations, 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 over this test set. While there was a clear gap in human recognition between minimal and spatial sub-minimal spatiotemporal configurations (average gap in human recognition rate 0.63; see Fig. S6-A), and between minimal and temporal sub-minimal spatiotemporal configurations (average gap in human recognition rate 0.68; see Fig. S6-B), the differences in recognition scores given by the network model for the minimal and sub-minimal examples were small (see Methods). In sum, none of the tested models, even when fine-tuned with minimal dynamic configurations described here, were able to account for human recognition of minimal spatiotemporal configurations.

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

The psychophysics data in Sec. 2 and Sec. 3 shows that processing of minimal spatiotemporal configurations in the human visual system requires the combining of motion and spatial information. We next compared the use of motion information by the human system and current CNN models (such as C3D) in the recognition of minimal spatiotemporal configurations. For this purpose, we compared the recognition of minimal and sub-minimal spatiotemporal configurations by two network models: (i) A purely spatial VGG19 network model, pre-trained on ImageNet and fine-tuned on frames of minimal configurations (see Methods), 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 configurations. Our goal was to quantify the match between the two models and human recognition on minimal configurations, in order to understand the contribution of temporal processing in the C3D model compared with static VGG19 architectures and to human behavior.

For the static VGG19 model, the recall gap between 'rowing' minimal configurations and spatial sub-minimal configurations was 0.34 (see Fig. S6-G), a large difference from the corresponding human gap (0.63, as mentioned above). For the dynamic C3D model, the recall gap between the temporal sub-minimal and the minimal configurations was 0.37 (see Fig. S6-H), 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 we repeated the test for hard negatives for C3D, and collected a set of 30 hard negative examples for the fine-tuned VGG19 model). Comparing human and VGG19 recognition for the set of hard negatives showed a difference in recognition accuracy (AP=0.64 for VGG19, See Fig. S6-I. Humans were not confused by any of the hard negatives: AP=1), but also a gap in recognition accuracy between the VGG19 and C3D models (0.64 vs. 0.18). This shows that the Average Precision for VGG19 is higher, closer to humans, than the AP for C3D model, indicating that the VGG19 was better at rejecting hard negative examples.

To conclude, the test results show that VGG19 is better than C3D in replicating human behavior for spatial sub-minimal configurations (recall gap: 0.34 for VGG19, 0.02 for C3D, 0.63 for humans) 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 was 0.37 for VGG19 vs. 0.78 for C3D, 0.68 for humans). We suspect that the reason for the latter is that the C3D is sensitive to basic dynamic features, which are not contained in our temporal sub-configurations, and which the spatial VGG19 cannot capture. The more surprising point is that for the spatial sub-configurations and the hard negative examples, the motion information that is added in the C3D is contributing very little, if any, to replicating human behavior. The different conditions and results above as summarized in Table S2.

Since minimal dynamic configurations 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 integration of the available time and space in formation by humans and the tested network models.

Discussion

We presented here minimal spatiotemporal configurations in which, by construction, all spatial and temporal visual information is required for human recognition (Figure 1). A slight change of the minimal configurations either in the spatial or temporal dimensions, led to a drastic drop in recognition of the action and objects in the scene. There was a trade-off between spatial and temporal information: adding more spatial information could enhance recognition when temporal information was insufficient and adding temporal information could enhance recognition when spatial information was insufficient (Figure 2). Action recognition in the minimal configurations was accompanied by interpretation of the different image parts and their interactions (Figure 3). State-of-the-art computational models of action recognition were unable to replicate the human behavior findings.

The minimal spatiotemporal configurations contained a mixture of both static features (e.g., the legs and torso of the person playing the violin do not change in time) and moving features (e.g., the hand and bow are moving); both are crucial for human recognition and interpretation, as revealed by the sharp transition to unrecognizable spatial and temporal sub-minimal configurations. Previous works have shown how moving features alone (e.g., all features are moving in biological motion studies[32] and in the slit experiments[11]) can be sufficient for action recognition. Many previous studies have also shown that static features can be sufficient for action recognition[31,33]. In contrast to the distinction between dynamic and static features suggested by those previous studies, we show that the 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 spatiotemporal configurations.

A known role of dynamics 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 for the action ‘moving a box’ to be recognized. Much of the computational vision literature has focused on this aspect of dynamics – the motion trajectories associated with objects that can be identified statically[27,34]. Minimal spatiotemporal configurations 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 of minimal spatiotemporal configurations, even when the models are fine-tuned for the task, and are trained with similar minimal configurations. This limitation motivates a future study of spatiotemporal features and computational recognition models that can better predict human behavior. The minimal spatiotemporal configurations provide a tool to study critical spatiotemporal features, as well as space-time dependency, by exploring the differences between the recognizable minimal configurations and their slightly reduced but unrecognizable sub-minimal versions. Future studies could extend recent modeling of full interpretation of spatial minimal images[30,31], to the modeling of full spatiotemporal interpretation, leading to a better understanding and more accurate modeling of spatio-temporal integration and human recognition.

Methods

Setting initial spatiotemporal configuration: The normalized frame size, the frame rate, and presentation as animated GIF. The initial spatiotemporal 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 50x50 image samples (pixels) and to graylevel colors. We then built a spatiotemporal configuration in which the selected normalized frames repeat in a loop at a fixed frame rate of 2 frames/second (2Hz). The spatiotemporal configuration was presented as animated GIF format. The choice of 2Hz frame rate was made since it provided the best recognition accuracy by the MTurk users.

Testing pre-trained network models on minimal spatiotemporal configurations: For 3D convolutional networks, we used the implementations by Hara et al[21], 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, then fine-tuned for the UCF101 benchmark. For two-stream network we used the implementation by Feichtenhofer et al., 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., 2015. Frames are input to layer of CNNs (based on AlexNet), then input to layer of LSTMs, scored by averaging across all video frames.

Negative examples for fine-tuning DNNs with minimal spatiotemporal configuration: 10000 negative examples were collected containing spatiotemporal configurations of a similar frame size and frame length as the positive set (minimal spatiotemporal configurations 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 because 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[35] when using standard data balancing techniques.

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 spatiotemporal configuration 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 Ullman et al., 2016[2]). For the C3D model, the recall gap between ‘rowing’ minimal configurations and spatial sub-minimal configurations was 0.02 (see Fig. S6-D), 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 (see methods). The reason 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. S6-E. All temporal sub-minimal configurations received a very low recognition score by the C3D model), which was close to the human gap.

Constructing spatial VGG19 model for recognizing minimal spatiotemporal configuration: 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 dynamic examples in the train set for the C3D mentioned above. When a novel dynamic 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.

Reporting Summary: Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability: The data that support the finding of this study are available from the corresponding author upon request.

Code availability: The computer codes are available from the corresponding author upon request.

Acknowledgements

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.

Author contributions

The experiments and ideas were jointly developed by GBY, GK and SU. GBY conducted all the experiments, computational simulations and analyzed all the data. The paper was written by GBY, GK and SU.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper (file attached).

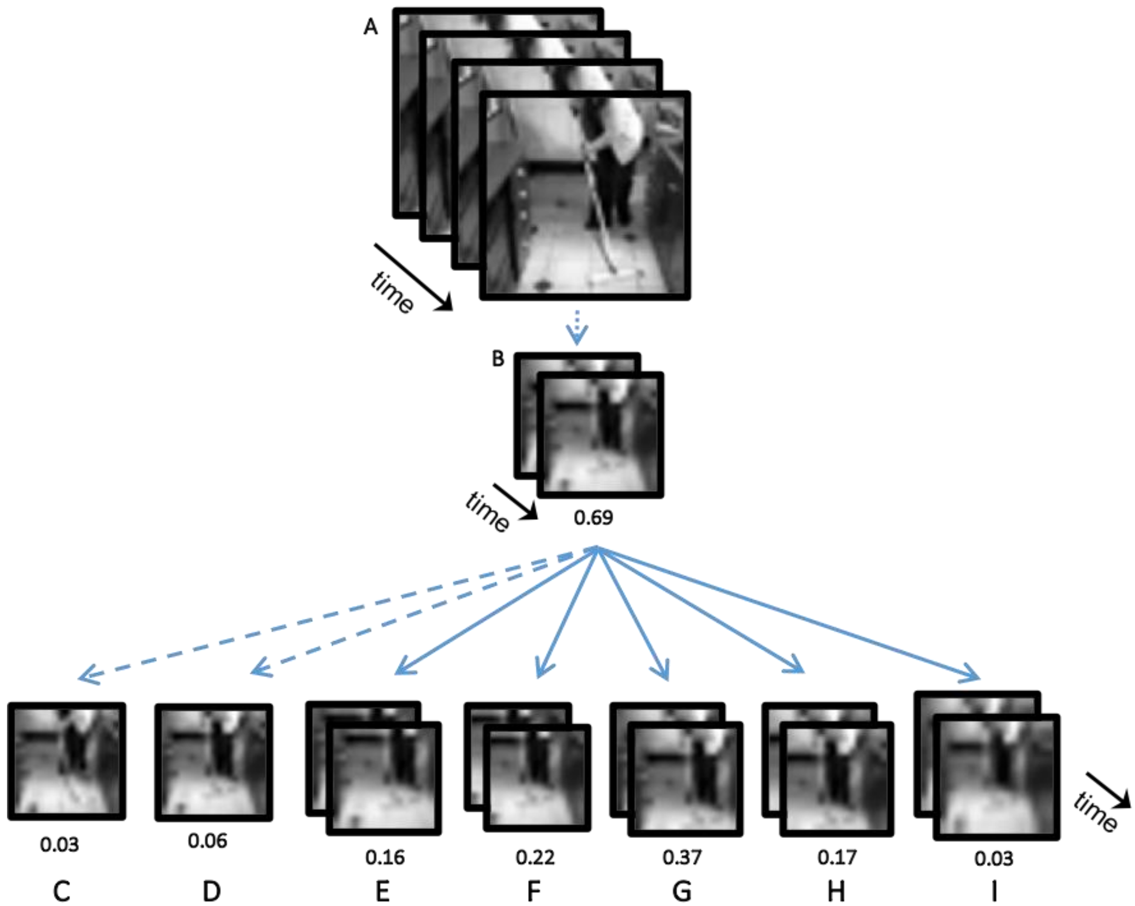


Figure 1. ***Example of a minimal spatiotemporal configuration***. *A short initial video clip showing 'mopping' activity (**A**) was gradually reduced in both space and time to a minimal recognizable configuration (**B**) (Methods). The numbers on the bottom of each image show the fraction of subjects who correctly recognized the action (subjects see 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**, in 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 'fig1.ppsx' for animated version of the dynamic configurations.*

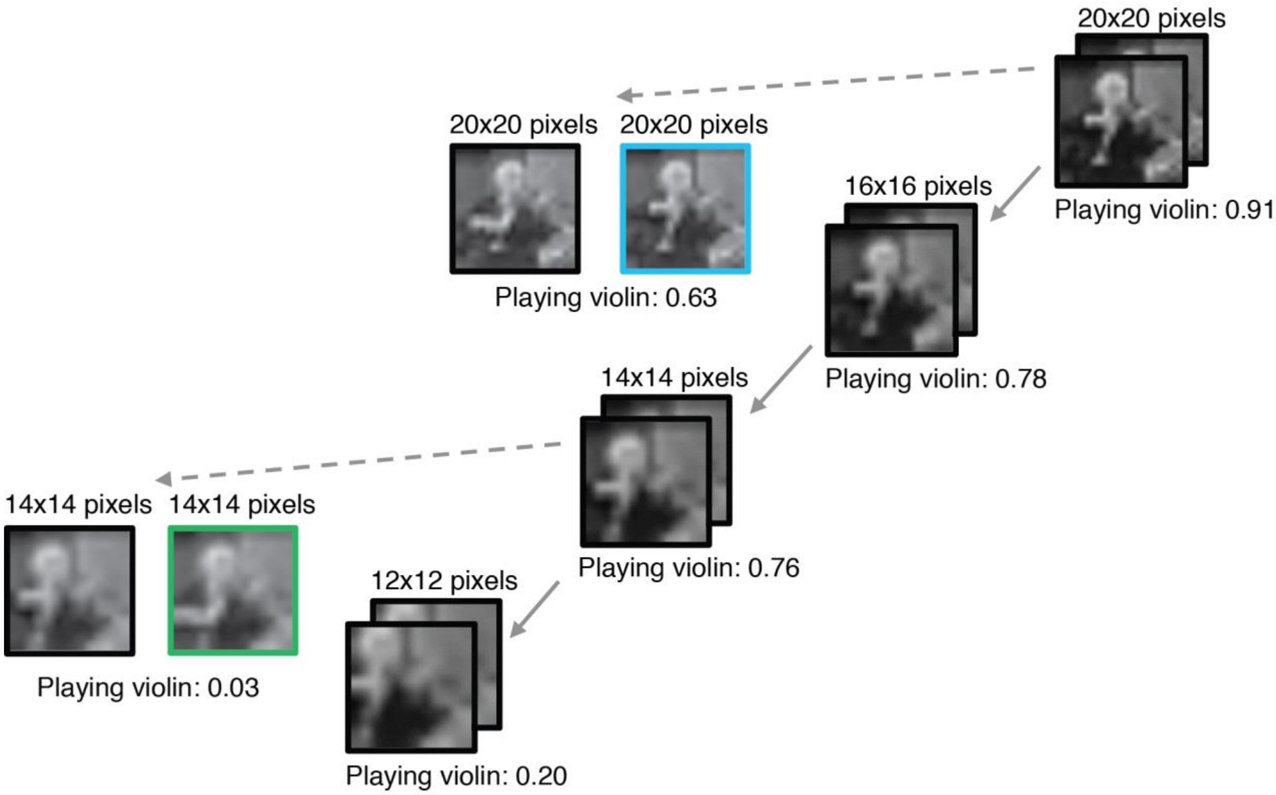


Figure 2. Trade-off between spatial and temporal information. *Solid connectors represent spatially reduced versions, dashed connectors represent temporal reduced versions. The numbers below each configuration represent the fraction of subjects that correctly identified the action "playing violin". The 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 (Figure S2). See Supplementary file 'fig2.ppsx' for animated version of the dynamic configurations.*

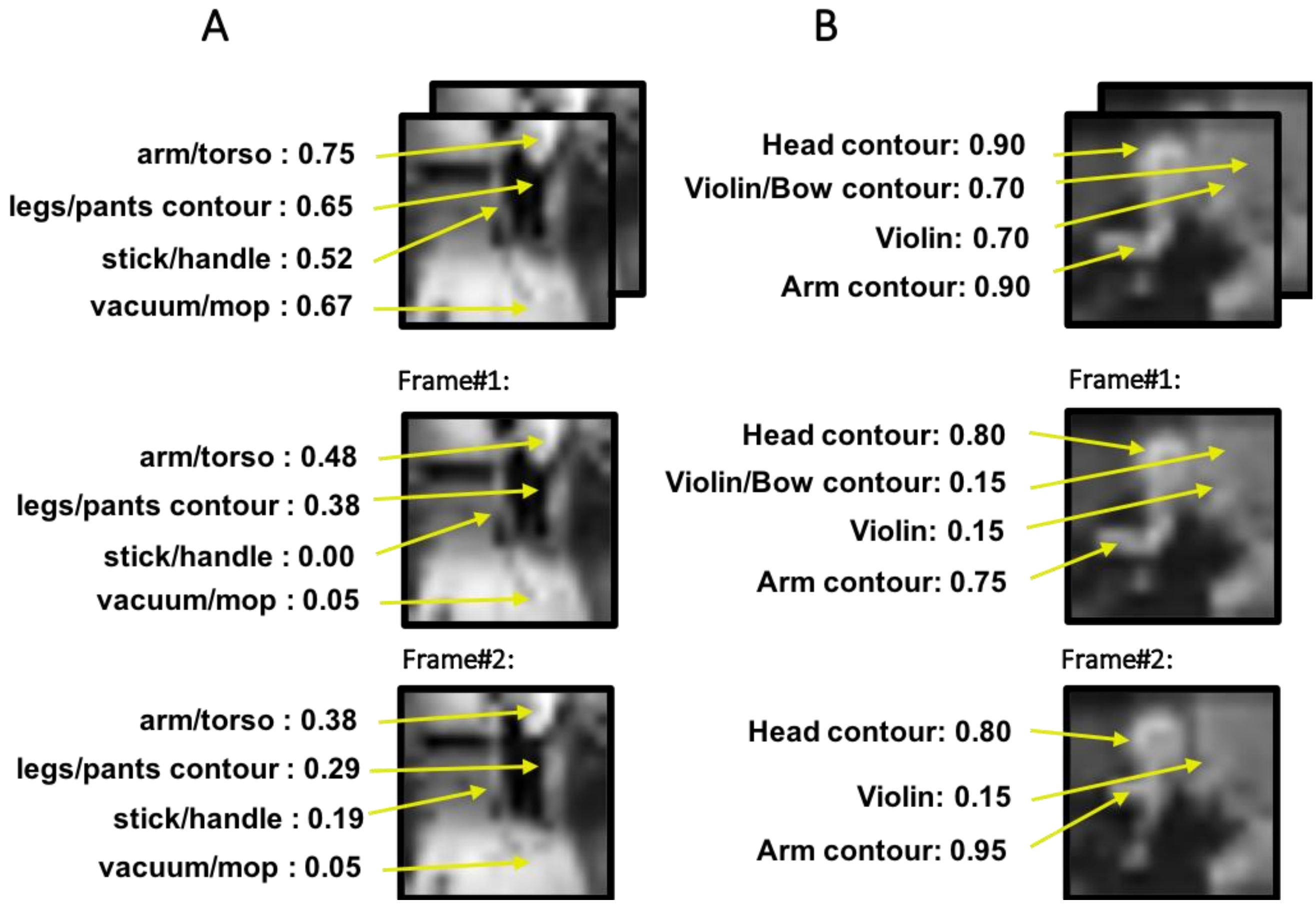


Figure 3. ***Spatiotemporal interpretation.*** *When humans could recognize the object and action, they could also identify a set of internal components of the agent and the object of action (top). In contrast, humans could not recognize these internal components (or could partially recognize them) in the sub-minimal versions (bottom four panels). Here are some of the recognized semantic components of minimal spatiotemporal 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.*

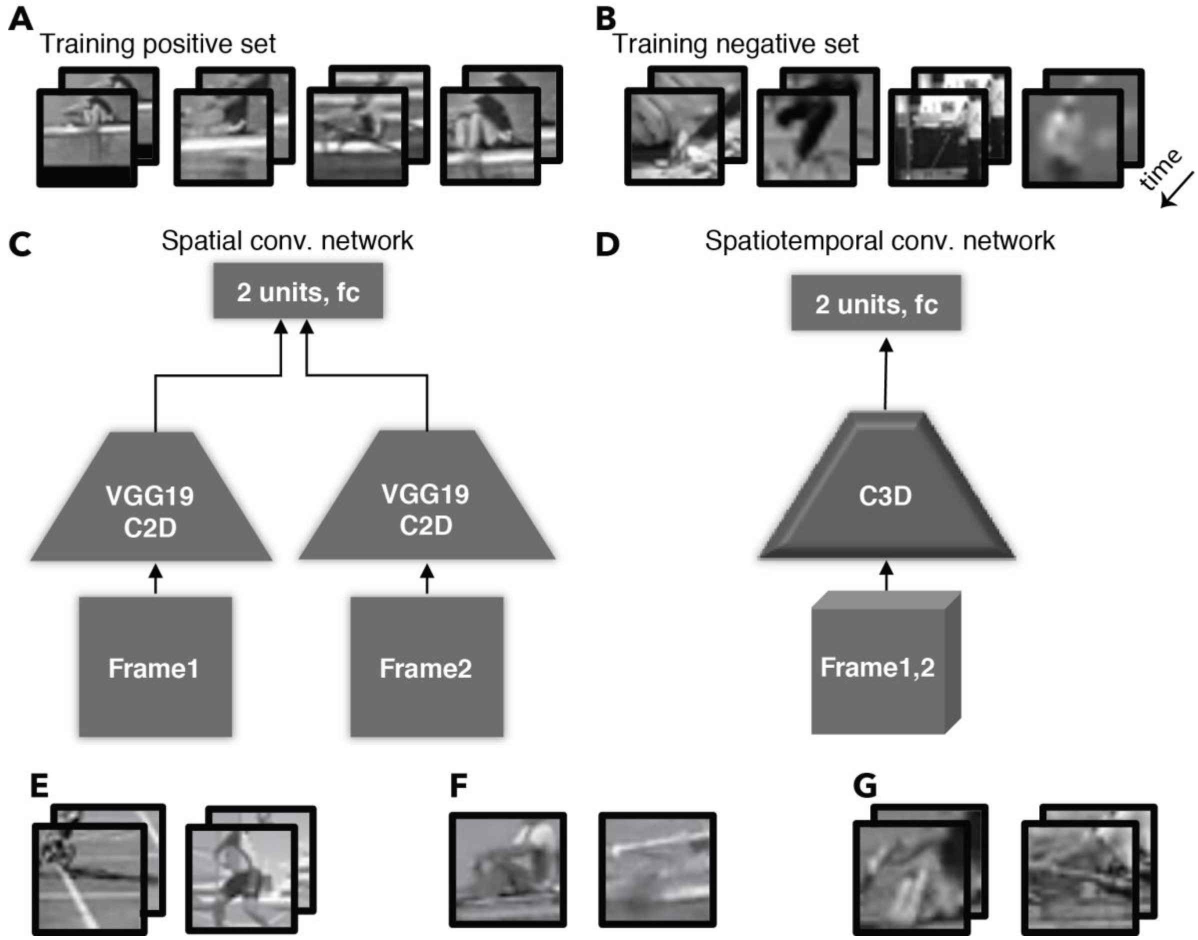


Figure 4. ***Testing minimal configurations with existing models for spatiotemporal recognition. (A-B)*** *A binary classifier is trained to separate a positive set of similar minimal images ("rowing"), showing the same action at the same body region and viewing position (**A**) from a negative set ("not rowing") including non-class images of the same size and style as the minimal configurations (**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 (Duran 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 3D and 2D 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 'fig4.ppsx' for animated version of the dynamic configurations.*

References

1 Potter, M. C. & Levy, E. I. Recognition memory for a rapid sequence of pictures. *J Exp Psychol* **81**, 10-15 (1969).

2 Ullman, S., Assif, L., Fetaya, E. & Harari, D. Atoms of recognition in human and computer vision. *Proc Natl Acad Sci U S A* **113**, 2744-2749, doi:10.1073/pnas.1513198113 (2016).

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

4 Sary, G., Vogels, R. & Orban, G. A. Cue-invariant shape selectivity of macaque inferior temporal neurons. *Science* **260**, 995-997 (1993).

5 Vaina, L., Solomon, J., Chowdhury, S., Sinha, P. & Belliveau, J. Functional neuroanatomy of biological motion perception in humans. *Proc Natl Acad Sci USA* **98**, 11656-11661 (2001).

6 Perrett, D. *et al.* Visual analysis of body movements by neurones in the temporal cortex of the macaque monkey: A preliminary report. *Behavioral Brain Research* **16**, 153-170 (1985).

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

8 Zollner, F. Über eine neue Art anorthoskopischer Zerrbilder. *Annalen der Physik* (1862).

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

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

11 Morgan, M. J., Findlay, J. M. & Watt, R. J. Aperture viewing: a review and a synthesis. *Q J Exp Psychol A* **34**, 211-233 (1982).

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

13 Kellman, P. J. & Cohen, M. H. Kinetic subjective contours. *Percept Psychophys* **35**, 237-244 (1984).

14 Singer, J. M. & Kreiman, G. Short temporal asynchrony disrupts visual object recognition. *J Vis* **14**, 7, doi:10.1167/14.5.7 (2014).

15 Singer, J. M., Madsen, J. R., Anderson, W. S. & Kreiman, G. Sensitivity to timing and order in human visual cortex. *J Neurophysiol* **113**, 1656-1669, doi:10.1152/jn.00556.2014 (2015).

16 Soomro, K., Zamir, A. R. & Shah, M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv preprint arXiv:1212.0402* (2012).

17 Kay, W. *et al.* The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

18 Karpathy, A. *et al.* Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* 1725-1732 (2014).

19 Tran, D. *et al.* A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6450-6459 (2018).

20 Tran, D., Bourdev, L., Fergus, R., Torresani, L. & Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4489-4497 (2015).

21 Hara, K., Kataoka, H. & Satoh, Y. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 18-22 (2018).

22 Simonyan, K. & Zisserman, A. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*. 568-576 (2014).

23 Feichtenhofer, C., Pinz, A. & Zisserman, A. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1933-1941 (2016).

24 Feichtenhofer, C., Pinz, A. & Wildes, R. Spatiotemporal residual networks for video action recognition. In *Advances in neural information processing systems*. 3468-3476 (2016).
25 Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput* **9**, 1735-1780 (1997).
26 Donahue, J. *et al.* Long-term recurrent convolutional networks for visual recogniton and description. *IEEE transactions on Pattern Analysis and Machine Intelligence* **39**, 677-691 (2017).
27 Cheron, G., Laptev, I. & Schmid, C. P-cnn: Pose based cnn features for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 3218-3226 (2015).
28 Kundu, A., Vineet, V. & Koltun, V. Feature space optimization for semantic video segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725-1732 (2016).
29 Hur, J. & Roth, S. Joint optical flow and temporally consistent semantic segmentation. In *European Conference on Computer Vision*. 163-177 (2016).
30 Ben-Yosef, G., Assif, L. & Ullman, S. Full interpretation of minimal images. *Cognition* **171**, 65-84, doi:10.1016/j.cognition.2017.10.006 (2018).
31 Ben-Yosef, G. & Ullman, S. Image interpretation above and below the object level. *Interface Focus* **8**, 20180020, doi:10.1098/rsfs.2018.0020 (2018).
32 Blake, R. & Shiffrar, M. Perception of human motion. *Annu Rev Psychol* **58**, 47-73, doi:10.1146/annurev.psych.57.102904.190152 (2007).
33 Yao, B. *et al.* Human action recognition by learning bases of action attributes and parts. In *Proceedings of the IEEE International Conference on Computer Vision*. 1331-1338 (2011).
34 Blank, M., Gorelick, L., Shechtman, E., Irani, M. & Basri, R. Actions as space-time shapes. In *Proceedings of the IEEE International Conference on Computer Vision*. 1395-1402 (2005).
35 Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning*. (2016).

Supplementary

Tables:

UCF101 Categories use to for search of minimal spatiotemporal configuration
Biking,
Rowing,
Playing violin,
Playing flute,
Playing Tennis,
Playing Piano,
Mopping,
Cutting,
Typing.

Table S1

Tests comparing humans and computational models:	**Humans**	**C3D model** (fine-tuned on minimal configurations)	**VGG19 model** (fine-tuned on minimal configurations)
Classifying minimal configurations vs. 'hard' non-class examples	Ave. Precision =1	Ave. Precision =0.18	Ave. 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

Table S2

Figures:

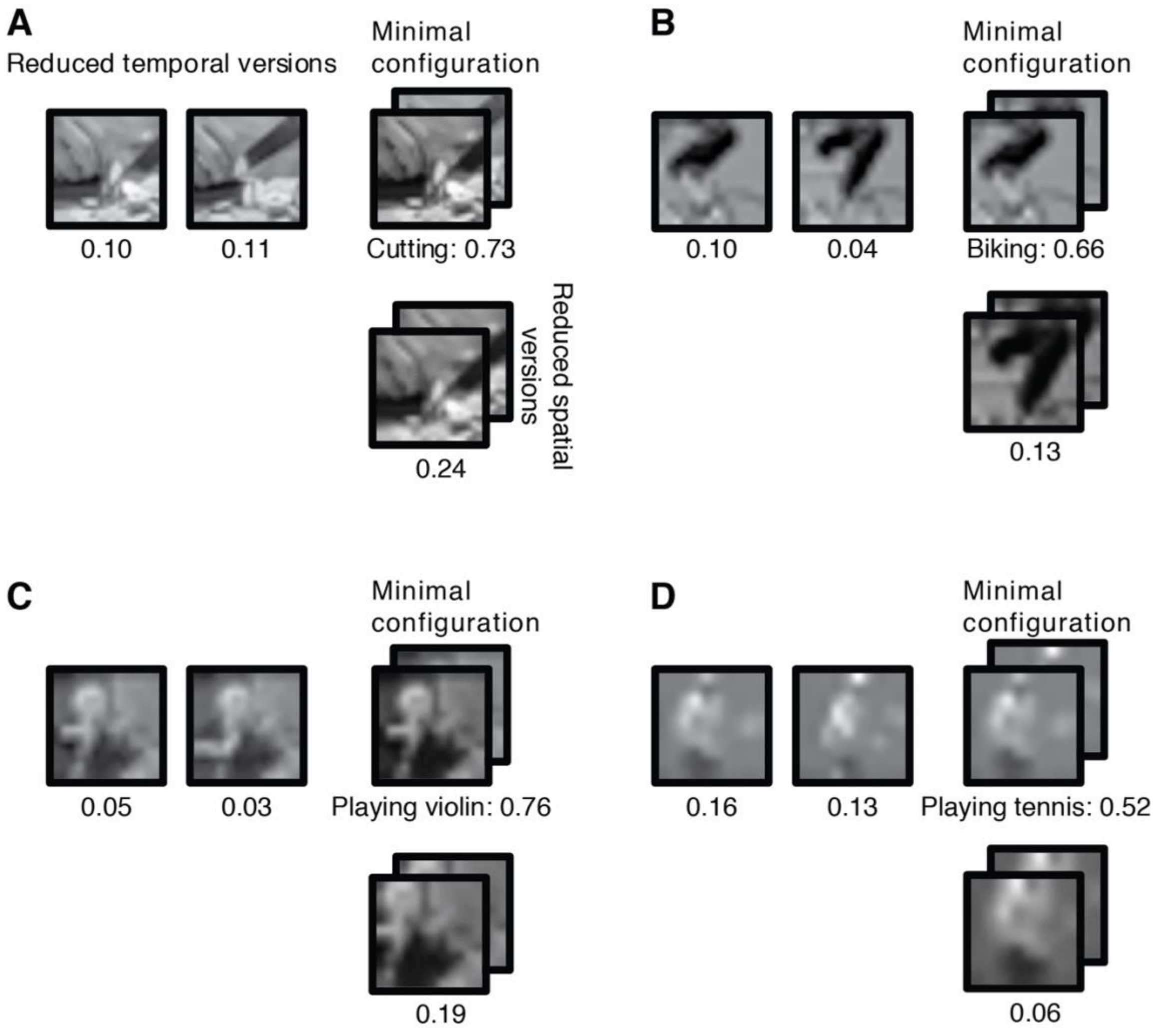
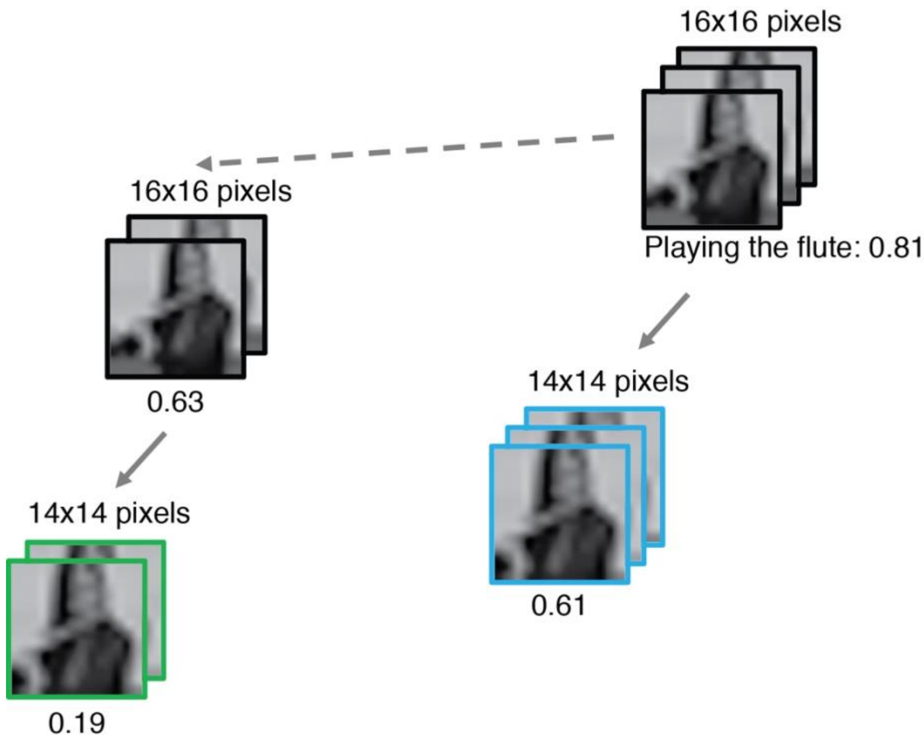
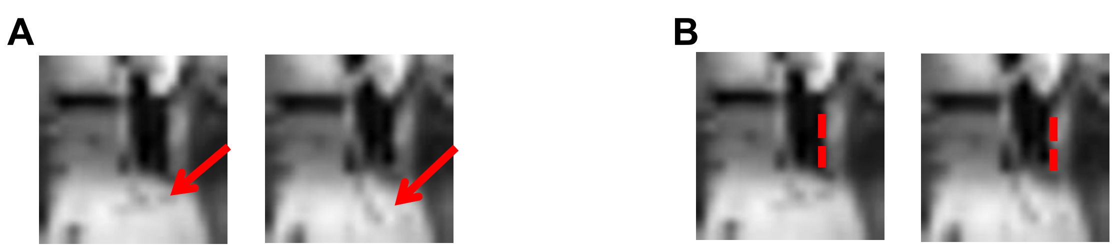


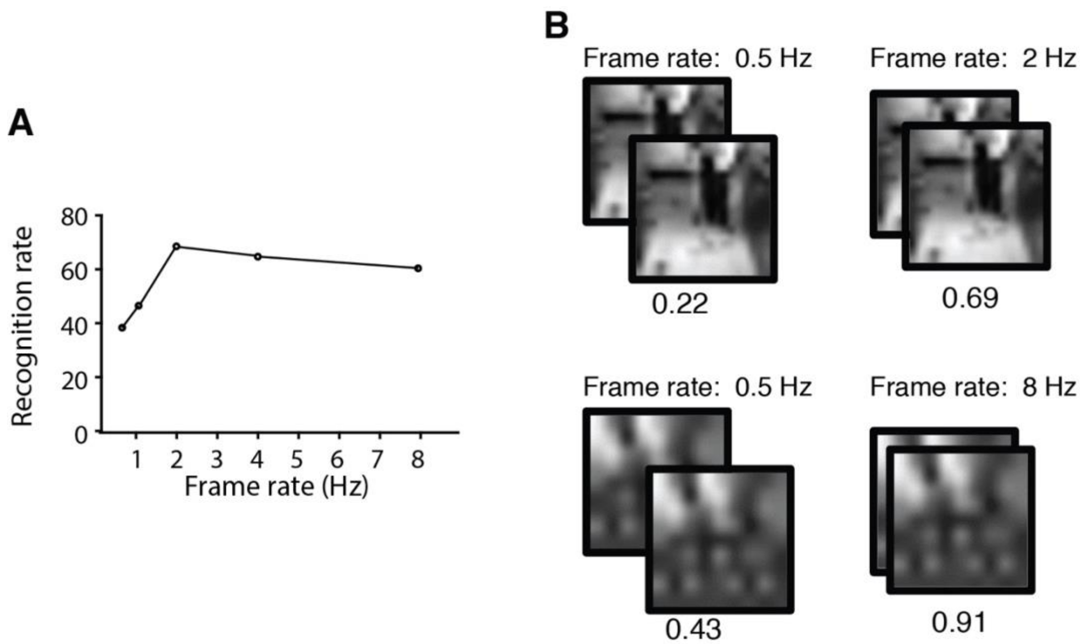
Figure S1. ***Examples of minimal and sub-minimal spatiotemporal configurations***. *Each minimal spatiotemporal configuration is shown next to its temporal sub-minimal versions (left), and its spatial sub-minimal version (below). The number represents the percentage of correct recognition responses for the action denoted below each minimal configuration (recall that MTurk users who tested on a minimal configuration were not tested on its sub-minimal configurations). Tags for similar actions were considered correct as well (e.g., Playing Baseball was considered similar to Playing Tennis). In the presented minimal images both the human object and the action category were recognized. In the presented sub-minimal image the actions were not recognized. The person object was partially recognized in C and D (see Fig. 3), and was not recognized in either A or B. See Supplementary file 'figS1.ppsx' for an animated version.*



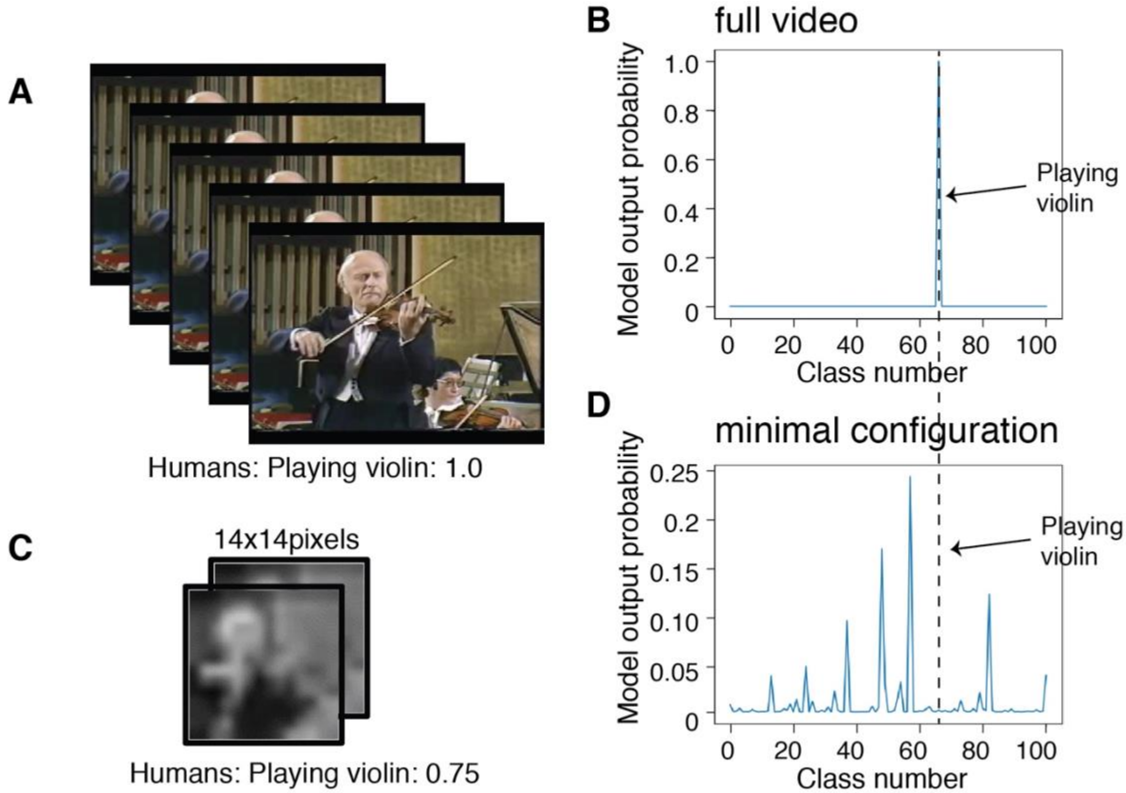
FigureS2. Trade-off between spatial and temporal information*.* *Solid connectors represent spatially reduced versions, while dashed connectors represent temporal reduced versions. The spatial sub-minimal 2-frame green configuration is not recognizable, but it becomes recognizable when more temporal information (i.e., more frames) is added, as shown in the 3-frame configuration in blue. The converse also holds: adding spatial information can recover performance for a temporal sub-minimal configuration (Figure 2). See supplementary file 'figS2.ppsx' for animated version.*



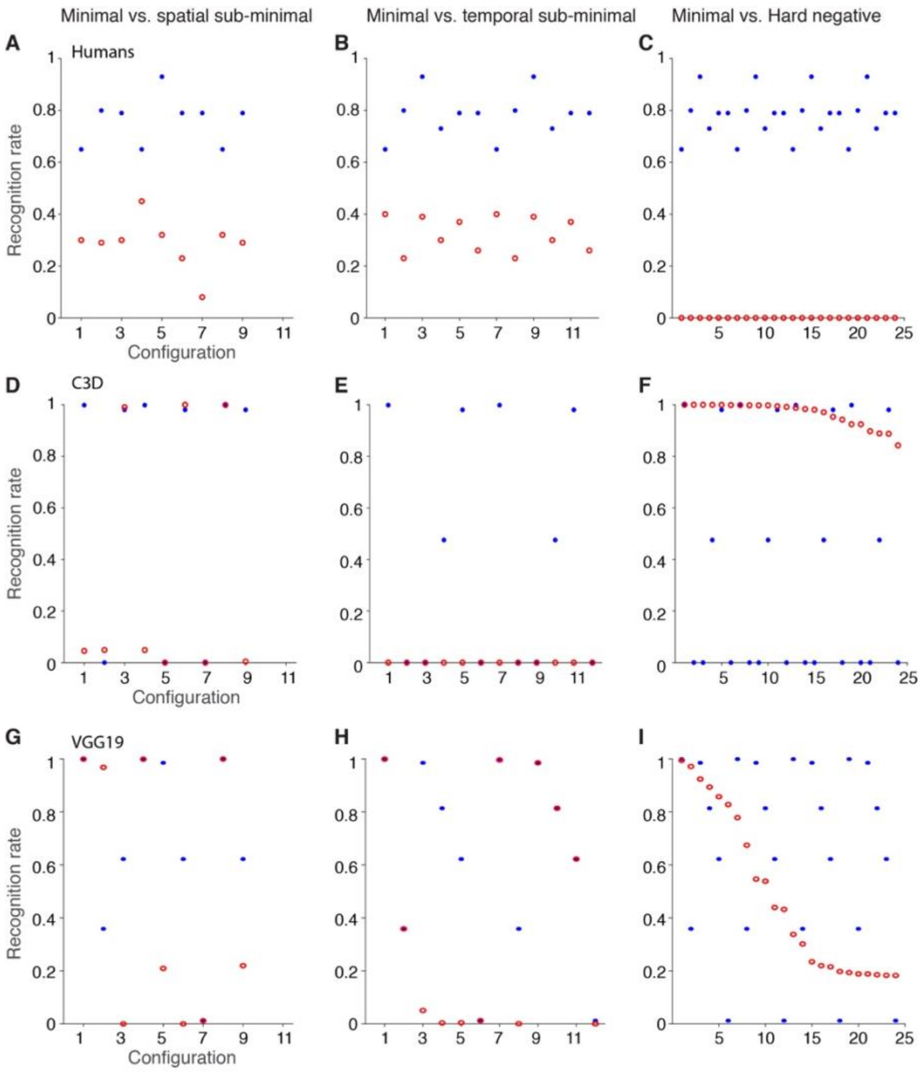
FigureS3. *Interpretation experiment via MTurk***.**
(A). *Arrow probes (red) were inserted to each frame in a minimal configuration (here: 'mopping' action) pointing to a specific part (here pointing to the mop/vaccum). The modified frames were then shown repeatedly one after another as a spatiotemporal configuration with 2Hz frame rate. Human subjects were then asked to tag the object part pointed by the arrow.*
*(**B**). A contour (red) was plotted along the border of a given object part (here along the border of a 'legs', or 'pants') for each frame of the minimal spatiotemporal configuration. Subjects were then asked to tag the parts shown on both sides of the contour.*



FigureS4.
*(**A**) Human recognition rate as a function of frames rate for the minimal configurations. Recognition decreases below frame rate of 2 Hz.*
***(B)** Two examples of the effect of changing frame rates on recognition of minimal spatiotemporal configurations. The same frames were shown to different MTurk users at different frame rates. The numbers show recognition success rate. See Supplementary file 'figS4.ppsx' for animated version of the dynamic configurations.*



FigureS5. *Pre-trained CNNs for spatiotemporal input were tested over full-viewed video clips (**A-B**), similar to the ones on their training process, and over minimal spatiotemporal configurations (**C-D**). Here is an example of typical behavior of the tested network models (here shown for the C3D model). The model could correctly classify the original video clip shown in A yielding a probability of 1 for the correct class number and 0 otherwise (B). However, the model failed to recognize the minimal configuration shown in C, yielding a probability of almost 0 for the correct class (D). This behavior stands in stark contrast with human recognition performance (percentage correct shown below the spatiotemporal configurations in A and C).*



FigureS6. *Comparison between humans (A-C), the fine-tuned C3D computational model (D-F) and the fine-tuned VGG19 computational model (G-I) for the 'rowing' example. The plot compares minimal (blue) versus spatial sub-minimal (red) configurations (A, D, G), minimal (blue) versus temporal sub-minimal (red) configurations (B, E, H) and minimal (blue) versus hard negative (red) configurations (C, F, I).*