# A functioning model of human time perception

Warrick Roseboom[*‡], Zafeirios Fountas[§], Kyriacos Nikiforou[§], David Bhowmik[§], Murray Shanahan[¶§], & Anil K. Seth[*†]

**Despite being a fundamental dimension of experience, how the human brain generates the perception of time remains unknown. Predominant models of human time perception propose the existence of oscillatory neural processes that continually track physical time - so called pacemakers - similar to the system clock of a computer[1,2,3]. However, clear neural evidence for pacemakers at psychologically relevant timescales is lacking, raising the question of whether internal pacemakers are necessary for time perception. Here we show that clock-like pacemaker processes are not required for human time perception. We built an artificial neural system based on a feed-forward image classification network[4], functionally similar to human visual processing[5,6]. In this system, input videos of natural scenes drive changes in activation within an image classification network and accumulation of salient changes in activations are used to estimate time. Estimates produced by this system match human reports made about the same videos, replicating key qualitative aspects such as report variability proportional to duration (scalar variability/Weber's law) and response regression to the mean (Vierordt's law)[2]. System-generated estimates also differentiate by scene type, such as walking around a busy city or sitting in a cafe, producing the same pattern of differences as human reports. Our results show how time perception can be derived from the operation of non-temporal perceptual classification processes, without any neural pacemaker, opening new opportunities for investigating the neural foundations of this central aspect of human experience.**

We recorded video of natural scenes such as walking through a city or the countryside, or sitting in an office or cafe (see Supplementary Video 1; Fig. 5). These videos were split into trials of one of thirteen durations between 1 and 64 seconds. Human participants watched these videos and made estimates of the presented video duration using a visual analogue scale (Fig. 1) while we recorded their gaze position using eye-tracking.

These same trial videos were used as the basis for input to a pre-trained feed-forward image classification network[4]. To estimate time, the system measured whether the Euclidean distance between successive activation patterns within a given layer, driven by the video input, exceeded a dynamic threshold (Fig. 2). For a given layer, when the activation difference exceeded the threshold a salient change was determined to have occurred, and a unit of time was accumulated. We implemented a dynamic threshold for each layer following a decaying exponential corrupted by Gaussian noise and resetting whenever a measured difference exceeded it, thus approximating the role of normalisation processes known to occur in biological sensory systems[10,11]. In order to transform the accumulated, abstract temporal units extracted by the system into a measure of time in standard units (seconds) for comparison with human reports, we trained a Support Vector Machine (SVM) to estimate the duration of the videos based on the accumulated salient changes (see Methods for full details of system design and training).

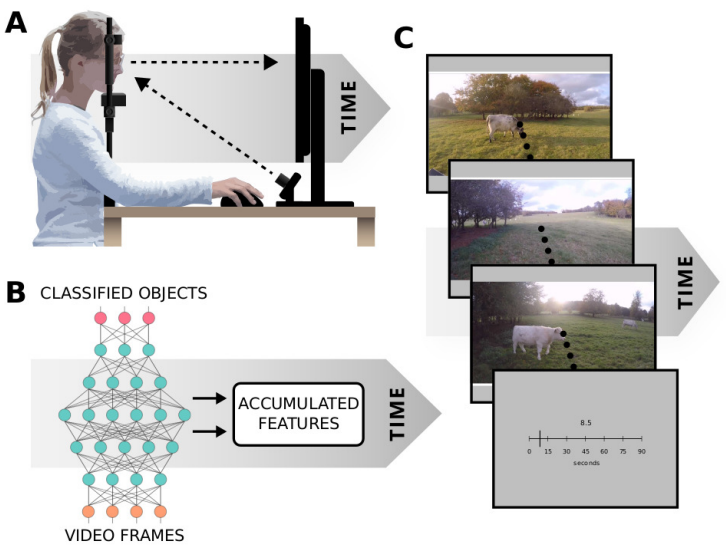


Figure 1: Experimental apparatus and procedure. (A) Human participants observed videos of natural scenes and reported the apparent duration while we tracked their gaze direction. (B) Depiction of the high-level architecture of the system used for simulations (Fig. 2). (C) Frames from a video used as a stimulus for human participants and input for simulated experiments. Human participants provided reports of the duration of a video in seconds using a visual analogue scale.

We initially had the system produce estimates under two input scenarios. In one scenario, the whole video frame was used as input. In the other, input was spatially constrained by biologically relevant filtering - the approximation of human visual spatial attention by a 'spotlight' centered on real human gaze fixation. The extent of this spotlight approximated an area equivalent to human parafoveal vision and was centered on the participants' fixation measured for each precise time-point in the video. Only the pixels inside this spotlight were

[*]Department of Informatics, University of Sussex, United Kingdom
[†]Sackler Centre for Consciousness Science, University of Sussex, United Kingdom
[‡]Corresponding author: wjroseboom@gmail.com
[§]Department of Computing, Imperial College London, United Kingdom
[¶]DeepMind, London, United Kingdom

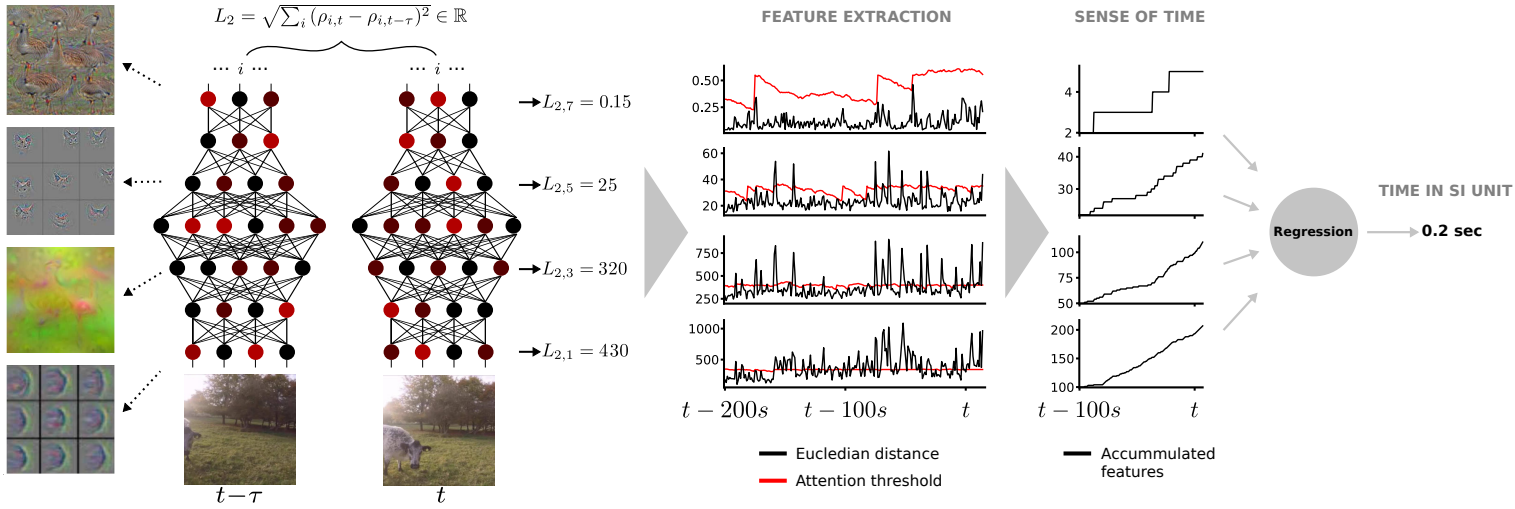


Figure 2: Depiction of the time estimation system. Salient differences in network activation driven by video input are accumulated and transformed into standard units for comparison with human reports. The left side shows visualisations of archetypal features to which layers in the classification network are responsive (adapted from[7,8,9]). The bottom left shows two consecutive frames of video input. The connected coloured nodes depict network structure and activation patterns in each layer in the classification network for the inputs. $L_2$ gives the Euclidean distance between network activations to successive inputs for a given network layer (layers conv2, pool5, fc7, output). In the Feature Extraction stage, the value of $L_2$ for a given network layer is compared to a dynamic threshold (red line). When $L_2$ exceeds the threshold level, a unit of subjective time is determined to have passed and is accumulated to form the base estimate of time. A regression method (support vector machine) is applied to convert this abstract time estimate into standard units (seconds).

used as input to the system (see Supplementary Video 2).

As time estimates generated by the system were made on the same videos as the reports made by humans, we could directly compare human and system estimates. Fig. 3 shows duration estimates produced by human participants and our system under the different input scenarios. Reports produced by our participants (Fig. 3A) demonstrated qualities typically found for human estimates of time: overestimation of short durations and underestimation of long durations (regression of responses to the mean/Vierordt's law), and variance of reports proportional to the reported duration (scalar variability/Weber's law).

System estimates produced when the full video frame was input (Fig. 3B; Full-frame model) revealed qualitative properties similar to human reports, though the degree of over and underestimation was exaggerated. This result demonstrates that the basic method of our system, accumulation of salient differences in network activation, can produce estimates of time - the slope of estimation is non-zero and short durations are discriminated from long durations by our system. However, while clearly able to produce temporal estimates and replicate qualitative aspects of human reports, the overall performance of the system in this case doesn't closely follow that of our human participants (Fig. 3E, F).

When the video input to the system was constrained to approximate human visual spatial attention by taking into account gaze position ("Gaze" model), system-produced estimates closely approximated reports made by human participants (Fig. 3B, E, F). This result was not simply due to the spatial reduction of input caused by the gaze-contingent spatial filtering, nor the movement of the input frame itself. When the gaze-contingent filtering was applied to videos other than the one from which gaze was recorded (i.e. gaze recorded while viewing one video then applied to a different video; "Shuffled" model), system estimates were poorer (Fig. 3D). These results indicate that the contents of where humans look in a scene play a key role in time perception.

To further test the idea that the contents of viewing play a fundamental role in time perception, and that our system reproduces this quality, we examined how system estimates differed by scene type. In our test videos, three different scenes could be broadly identified: scenes filmed moving around a city, moving around a leafy university campus and surrounding countryside, or from relatively stationary viewpoints inside a cafe or office (Fig. 5). Based on the idea that a more varied input should lead to more varied activation within the network layers, and therefore greater accumulation of salient changes, we looked at how many salient changes were accumulated by the system when given each scene type as input. As shown in (Fig. 4A), when the system was shown city scenes, which we would expect to be more visually dynamic and contain more changes in input over time, the system accumulated the greatest number of salient changes. This was true for each layer of the network that we examined. When the system was shown videos of scenes from moving around the campus or countryside, fewer salient changes were accumulated than the city, but more than the relatively stationary scenes of a cafe or office.

When we examined human estimates of duration for the different scene types, we found the same pattern of differences as for the system; human participants reported city scenes as longer in duration than campus/countryside and cafe/office scenes (Fig. 4B) (overall, human participants underestimated the duration of all scenes, but underestimated the duration of city scenes the least, indicating that city scenes were perceived as longer in duration). This scene dependency of time shows that the basic information from which our system estimates time is qualitatively similar to human participants - even without the final step of transforming accumulated salient changes into standard units of physical time. Both the location in the scene where information related to time can be found (gaze-contingency of time perception), and the broad temporal properties of different scenes appear to be key features for humans and our system when estimating time. These findings demonstrate that human-like time perception can be accomplished in the absence of processes that explicitly track physical time, as required by pacemaker-based proposals.

A potential criticism of our work would be that we haven't eliminated the need for a pacemaker, but simply found a proxy for its

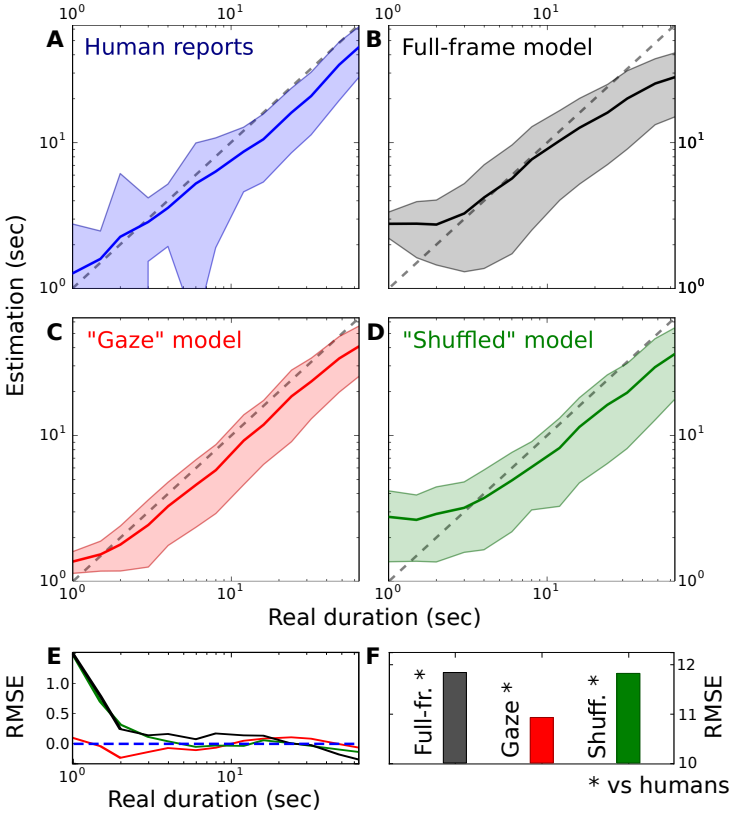


Figure 3: The mean duration estimates for 4290 trials for both human (A) and system (B,C,D) for the range of presented durations (1-64s). Shaded areas show $\pm 1$ standard deviation of the mean. Human reports (A) show typical qualities of human temporal estimation with overestimation of short and underestimation of long durations. (B) System estimates when input the full video frame replicate similar qualitative properties, but temporal estimation is poorer than humans. (C) System estimates when the input was constrained to approximate human attention based on human gaze data very closely approximated human reports made on the same videos, though when we "Shuffle" the gaze contingency such that the spotlight is applied to a different video than it was obtained on (D), performance decreases. (E) Comparison of mean absolute error between different estimations across presented durations. (F) Comparison of the root mean squared error of the system estimates compared to the human data. The "Gaze" model is most closely matched.

operation. However, such criticisms miss the point of our approach. Pacemaker-based models necessitate that time is estimated from an internal operation that attempts to match physical time - analogous to a computer system clock. Our approach moves away from the requirement that subjective time be tightly related to physical time in this way. For our system, accumulated temporal units don't represent the passage of physical time as in pacemaker-based systems, they *are* subjective time. Consequently, our model can easily account for many context-based distortions in subjective time, such as the scene-wise differences in time estimation seen in Fig. 4. By contrast, pacemaker-based approaches require a change in internal pacemaker operation, such as spontaneous changes in pacemaker rates (e.g.[12]) to produce such cases.

One might still worry that we do retain an underlying physical pacemaker because calculation of salient network activation changes occurs at some frequency. In the reported model, the video was input and activation difference calculated at 30 Hz. However, it is easy to demonstrate that the update rate is not the predominant feature in determining time estimates. If it were, duration estimates for the "Gaze" versus "Shuffled" models would be highly similar, as they contain the same input rate (30 Hz) and temporal features induced by movement of the gaze spotlight. However, this is clearly not the case (Fig. 3C) and (Fig. 3D). To thoroughly reject the idea that system update rate was the main determinant of time estimation in our system, we compared the salient changes accumulated by the system when inputing the 'normal' videos at 30 Hz, with accumulated changes under three conditions: videos in which the frame rate was halved (skipped every second frame), videos in which some frames were skipped psuedo-randomly with a frequency of 20%, or videos input at 30Hz, but with the video frames presented in a shuffled order. We found that the manipulations of frame rate (skipping every second frame or 20% of frames) produced only small differences in accumulated changes over time compared to the normal input videos (Fig. 6). However, when the input rate was kept at 30 Hz, but the presentation order of the frames was shuffled, thereby disrupting the flow of content in the video, the number of accumulated changes was very different (around 40 times *more* different from standard than either the halved or randomly skipped frame cases; see Fig. 6). These results underline that our system was producing temporal estimates based predominantly on the content of the scene, not the update rate of the system.

While the update rate is not critical in our system, how a biological system such as the human brain produces a comparison of successive activation states in sensory networks is an interesting question. Previous work has implicated sub-cortical areas such as the basal ganglia and the striatum in time perception[1,2]. Responses of medium-spiny projection neurons (MSN), the most prevalent type of neuron in the striatum, have been shown to encode time intervals with a bell-shaped distribution of errors that expands over time proportionally, like human reports[1]. MSN are known to have a many-to-one input configuration[13], are highly connected with cortical sensory areas[14,13], have uncommonly low membrane excitability and dynamic firing thresholds[1,13]. In previous pacemaker-based models, such as the striatal-beat-frequency model (SBF[1,2]) striatal neurons, due to their connectivity, are thought to act as integrators and fire when neural pacemaker oscillations occur synchronously. Alternatively, we propose that striatal neurons may be firing in response to the difference between successive states of sensory networks, acting as the difference calculator depicted in (Fig. 2). Striatal lateral inhibition[15] and short-term intrinsic depression of striatal afferents[13] have the potential to ensure that a new sensory state would be accumulated only once. In addition, short-term potentiation of striatal connections to other basal ganglia nuclei (globus pallidus[16] and substantia nigra[17]) could encode the accumulated features (i.e. subjective units of time). Outputs from striatal MSNs would communicate the accumulation of time to the remainder of basal ganglia circuitry. This view of the role of the basal ganglia in time perception is intriguing as it facilitates integration of the extant literature of time perception-related neural regions with our finding that no pacemaker-like processes are required for a sensory system to estimate time.

The core feature that allows our model to produce human-like time perception is the identification and accumulation of salient changes in network activation (salient perceptual events). In the current model, accumulation is accomplished by memory dedicated only to tracking the occurrence of these events. However, our ongoing work seeks to link identification of salient events with content-based memory for these events (i.e., episodic-like memory). In this way, our approach may accommodate data consistent with both prospective (online time estimation, e.g. waiting for an event to happen)[18,19] and retrospective time estimation (based on memory of past events)[18,19], as the basic

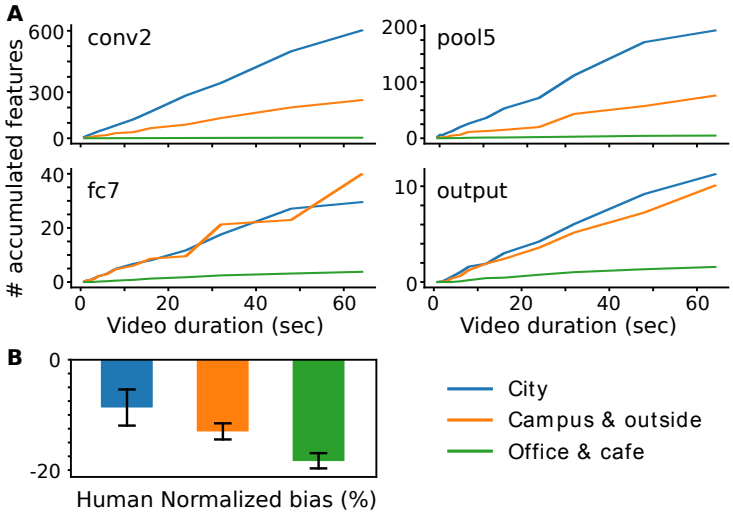


Figure 4: System salient change accumulation and human reports by scene type. (A) The number of accumulated salient changes over time in the different network layers (lowest to highest: conv2, pool5, fc7, output), depending on input scene type. (B) The normalised bias in human participants' duration reports relative to physical duration. The less the underestimation (closer to zero), the longer the reported duration for that scene.

process of identifying changes in sensory network activation would underlie both.

Our results demonstrate that internal pacemaker operations are not necessary for modeling the quantitative properties of human time perception. System-produced time estimates replicated well-known features of human reports of time that differed based on biologically relevant cues, such as where in a scene sensory input is coming from, as well as the general content of a scene. That our system produces human-like time estimates based on only natural video inputs is a major achievement in building artificial systems with human-like temporal cognition, and presents a fresh opportunity to understand human perception and experience of time.

# Acknowledgments

This work was supported by the European Union Future and Emerging Technologies grant (GA:641100) TIMESTORM – Mind and Time: Investigation of the Temporal Traits of Human-Machine Convergence and the Dr. Mortimer and Theresa Sackler Foundation, supporting the Sackler Centre for Consciousness Science. Thanks to Michaela Klimova, Francesca Simonelli and Virginia Mahieu for assistance with the human experiment. Thanks to Tom Wallis for comments on the initial version of the manuscript.

# References

[1] Matthew S Matell and Warren H Meck. Cortico-striatal circuits and interval timing: coincidence detection of oscillatory processes. *Cognitive brain research*, 21(2):139–170, 2004.

[2] Hedderik Van Rijn, Bon-Mi Gu, and Warren H Meck. Dedicated clock/timing-circuit theories of time perception and timed performance. In *Neurobiology of interval timing*, pages 75–99. Springer, 2014.

[3] Bon-Mi Gu, Hedderik van Rijn, and Warren H Meck. Oscillatory multiplexing of neural population codes for interval timing and working memory. *Neuroscience & Biobehavioral Reviews*, 48:160–185, 2015.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[5] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1(10):417–446, 2015.

[6] Tomoyasu Horikawa and Yukiyasu Kamitani. Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features. *Frontiers in computational neuroscience*, 11, 2017.

[7] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *arXiv preprint arXiv:1311.2901*, 2013.

[8] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. *arXiv preprint arXiv:1412.0035*, 2014.

[9] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[10] Matteo Carandini and David Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13:51–62, 2013.

[11] Samuel G. Solomon and Adam Kohn. Moving sensory adaptation beyond suppressive effects in single neurons. *Current Biology*, 24:R1012–R1022, 2014.

[12] Sylvie Droit-Volet and John Wearden. Speeding up an internal clock in children? effects of visual flicker on subjective duration. *The Quarterly Journal of Experimental Psychology Section B*, 55(3):193–211, 2002.

[13] Heinz Steiner and Kuei Y Tseng. *Handbook of basal ganglia structure and function*, volume 24. Academic Press, 2016.

[14] Garrett E Alexander, Mahlon R DeLong, and Peter L Strick. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual review of neuroscience*, 9(1):357–381, 1986.

[15] Mark D Humphries, Ric Wood, and Kevin Gurney. Reconstructing the three-dimensional gabaergic microcircuit of the striatum. *PLoS computational biology*, 6(11):e1001011, 2010.

[16] Robert E Sims, Gavin L Woodhall, Claire L Wilson, and Ian M Stanford. Functional characterization of gabaergic pallidopallidal and striatopallidal synapses in the rat globus pallidus in vitro. *European journal of neuroscience*, 28(12):2401–2408, 2008.

[17] William M Connelly, Jan M Schulz, George Lees, and John NJ Reynolds. Differential short-term plasticity at convergent inhibitory synapses to the substantia nigra pars reticulata. *Journal of Neuroscience*, 30(44):14854–14861, 2010.

[18] Simon Grondin. Timing and time perception: A review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception, and Psychophysics*, 72(3):561–582, 2010.

[19] Christopher J. MacDonald. Prospective and retrospective duration memory in the hippocampus: is time in the foreground or background? *Phil. Trans. R. Soc. B*, 369(1637), 2014.

[20] David Brainard. The psychophysics toolbox. *Spatial Vision*, 10, 1997.

[21] Denis Pelli. Videotoolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, 10, 1997.

[22] Mario Kleiner, David Brainard, and Denis Pelli. What's new in psychtoolbox-3? *Perception*, 36, 2007.

[23] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[25] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

# Methods

**Participants** Participants were 55 adults (21.2 years, 40 female) recruited from the University of Sussex, participating for course credit or £5 per hour. Participants typically completed 80 trials in the 1 hour experimental session, though due to time or other constraints some participants only completed as few as 20 trials (see Supplemental Data for specific trial completion details). This experiment was approved by the University of Sussex ethics committee.

**Apparatus** Experiments were programmed using Psychtoolbox 3 [20,21,22] in MATLAB 2012b (MathWorks Inc., Natick, US-MA) and the Eyelink Toolbox (Cornelissen et al., 2002), and displayed on a LaCie Electron 22 BLUE II 22" with screen resolution of 1280 x 1024 pixels and refresh rate of 60 Hz. Eye tracking was performed with Eyelink 1000 Plus (SR Research, Mississauga, Ontario, Canada) at a sampling rate of 1000 Hz, using a desktop camera mount. Head position was stabilized at 57 cm from the screen with a chin and forehead rest.

**Stimuli** Experimental stimuli were based on videos collected throughout the City of Brighton in the UK, the University of Sussex campus, and the local surrounding area. They were recorded using a GoPro Hero 4 at 60 Hz and 1920 x 1080 pixels, from face height. These videos were processed into candidate stimulus videos 165 minutes in total duration, at 30 Hz and 1280 x 720 pixels. To create individual trial videos, a psuedo-random list of 4290 trials was generated - 330 repetitions of each of 13 durations (1, 1.5, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64s). The duration of each trial was psuedo-randomly assigned to the equivalent number of frames in the 165 minutes of video. There was no attempt to restrict overlap of frames between different trials. The complete trial list and associated videos are available in the Supplemental Data.

For computational experiments when we refer to the 'full frame' we used the center 720 x 720 pixel patch from the video (56 percent of pixels; approximately equivalent to 18 degrees of visual angle (dva) for human observers). When computational experiments used human gaze data, a 400 x 400 pixel patch was centered on the gaze position measured from human participants on that specific trial (about 17 percent of the image; approximately 10 dva for human observers).

**Computational model architecture** The computational model is made up of four parts: 1) A classification deep neural network, 2) an threshold mechanism, 3) a set of accumulators and 4) a regression scheme. We used the convolutional deep neural network AlexNet [4] available through the python library caffe [23]. AlexNet has been pretrained to classify high-resolution images in the LSVRC-2010 ImageNet training set [24] into 1000 different classes, with state-of-the-art performance. It consists of five convolutional layers, some of which are followed by normalisation and max-pooling layers, and two fully connected layers before the final 1000 class probability output. It has been argued that convolutional networks' connectivity and functionality resemble the connectivity and processing taking place in human visual processing [5] and thus we use this network as the main visual processing system for our computational model. At each time-step (30 Hz), a video frame is fed into the input layer of the network and the subsequent higher layers are activated. For each frame, we extract the activations of all neurons from layers conv2, pool5, fc7 and the output probabilities. For each layer, we calculate the Euclidean distance between successive states. If the activations are similar, the Euclidean distance will be low, while the distance between neural activations corresponding to frames which include different objects will be high.

A 'temporal attention' mechanism is implemented to dynamically calibrate the detection of changes between neural activations (threshold) resulting from successive frames. Each of the four layers has an initial threshold value for the distance in neural space. This threshold decays with some stochasticity (Eq. 1) over time to replicate normalisation of neural responses to stimulation over time. A new salient feature for each layer is registered once the Euclidean distance between activations for two successive frames exceeds this threshold, the counter in each of the layers' accumulators is incremented by one and the threshold of that layer is reset to its maximum value. The purpose of a decaying function is to accommodate time perception across various environments with too few or too many features. Implementation details for each layer can be found in the table below, and the threshold was calculated as:

$$T^k_{t+1} = T^k_t - \left(\frac{T^k_{max} - T^k_{min}}{\tau^k}\right) e^{-\left(\frac{D}{\tau^k}\right)} + \mathcal{N}\left(0, \frac{T^k_{max} - T^k_{min}}{\alpha}\right) \quad (1)$$

where $T^k_t$ is the threshold value of $k^{th}$ layer at timestep $t$ and $D$ indicates the number of timesteps since the last time the threshold value was reset. $T^k_{max}$, $T^k_{min}$ and $\tau^k$ are the maximum threshold value, minimum threshold value and decay timeconstant for $k^{th}$ layer respectively, values for which are provided in Table 1. Stochastic noise drawn from a Gaussian is added to the threshold and $\alpha$ a dividing constant to adjust the variance of the noise.

Table 1: Threshold mechanism parameters

| **Parameters for implementing salient event threshold** | | | | |
|---|---|---|---|---|
| Layer | No. neurons | $T_{max}$ | $T_{min}$ | $\tau$ |
| conv2 | 290400 | 340 | 100 | 100 |
| pool5 | 9216 | 400 | 100 | 100 |
| fc7 | 4096 | 35 | 5 | 100 |
| output | 1000 | 0.55 | 0.15 | 100 |

The parameters of the model, $T^k_{max}$, $T^k_{min}$ and $\tau^k$, were chosen so that the Euclidean distances of each layer exceed the threshold only when a large increase occurs. The choice of particular values is not very important as the model performance is robust across a broad range of these values. When we scaled the values of $T^k_{max}$, $T^k_{min}$ by a factor allowing us to vary the level of the threshold mechanism ('Attention Level'), our model can still estimate time with relatively good accuracy across a broad range of these values (Fig. 7A) and, most importantly, still differentiate between short and long durations (slope is greater than zero for most levels). To further examine the effect of $T^k_{max}$ and $T^k_{min}$, we scaled each parameter by an independent scaling factor to show that the model estimations (compared to the real physical duration) are robust over a wide range of values for these two parameters (Fig. 7B).

The number of accumulated features in the accumulators represent the elapsed duration between two points in time. In order to convert estimates of subjective time into units of time in seconds, a simple regression method was used based on epsilon-Support Vector Regression (SVR) from sklearn python toolkit [25]. The kernel used was the radial basis function with a kernel coefficient of $10^{-4}$ and a penalty parameter for the error term of $10^{-3}$. We used 10-fold cross-validation. To produce the presented data, we used 9 out of 10 groups for training

and one (i.e. 10% of data) for testing. This process was repeated 10 times so that each group was used for validation only once. In order to verify that our system performance was not simply due to overfitting of the regression method for the set of durations we included, rather than the ability of the system to estimate time, we tested the model estimation performance when excluding some durations from the training set, but keeping them in the testing set. The mean normalised error for durations included and excluded in each experiment is shown in (Fig. 8). As can be seen, only when excluding a large number of training levels (e.g. 10 out of 13 possible levels) does the estimation error get notably larger.

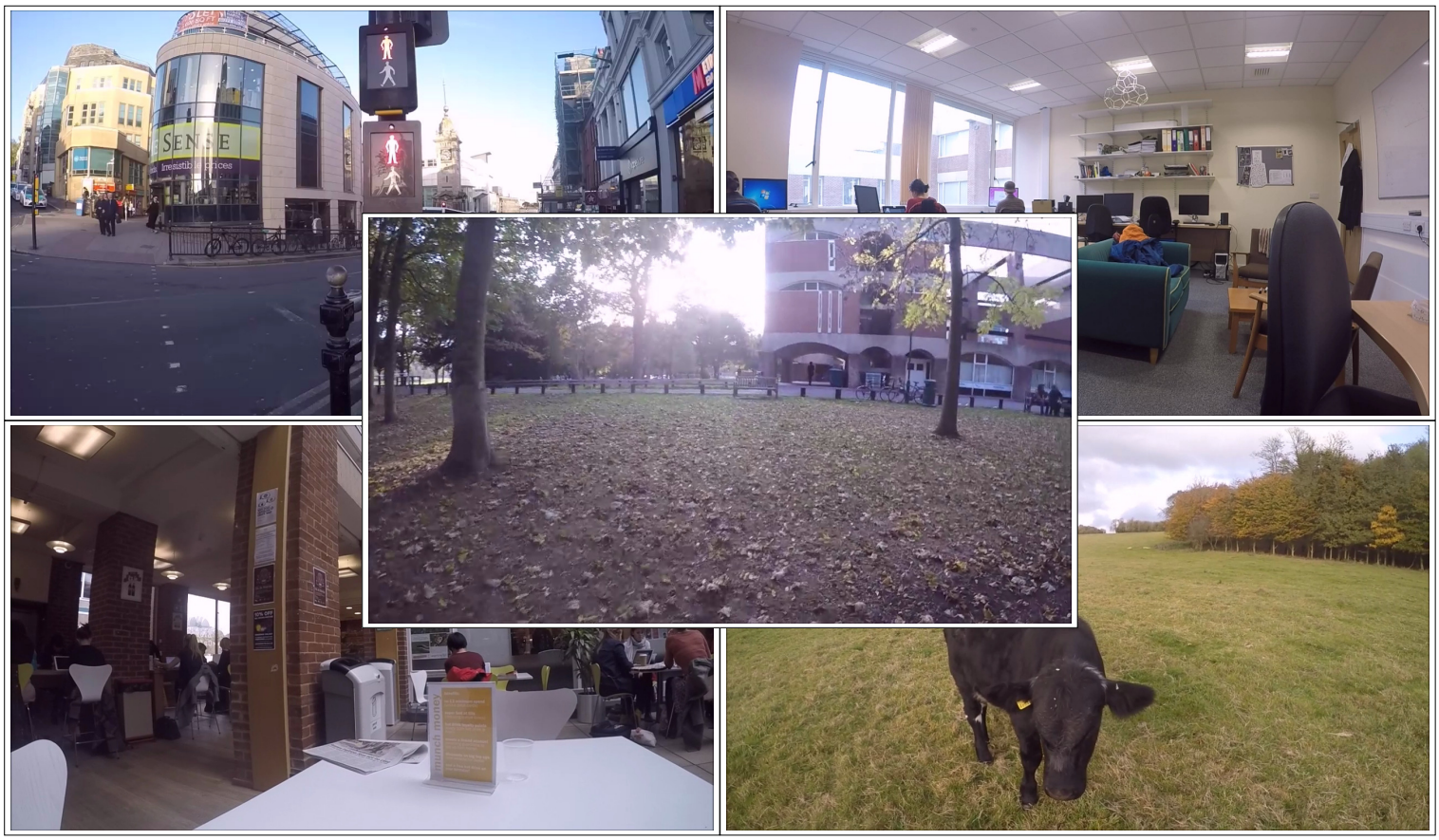


Figure 5: **(Extended figure)** Videos used as stimuli for the human experiment and input for the system experiments included scenes recorded walking around a city (top left), in an office (top right), in a cafe (bottom left), walking in the countryside (bottom right) and walking around a leafy campus (center).

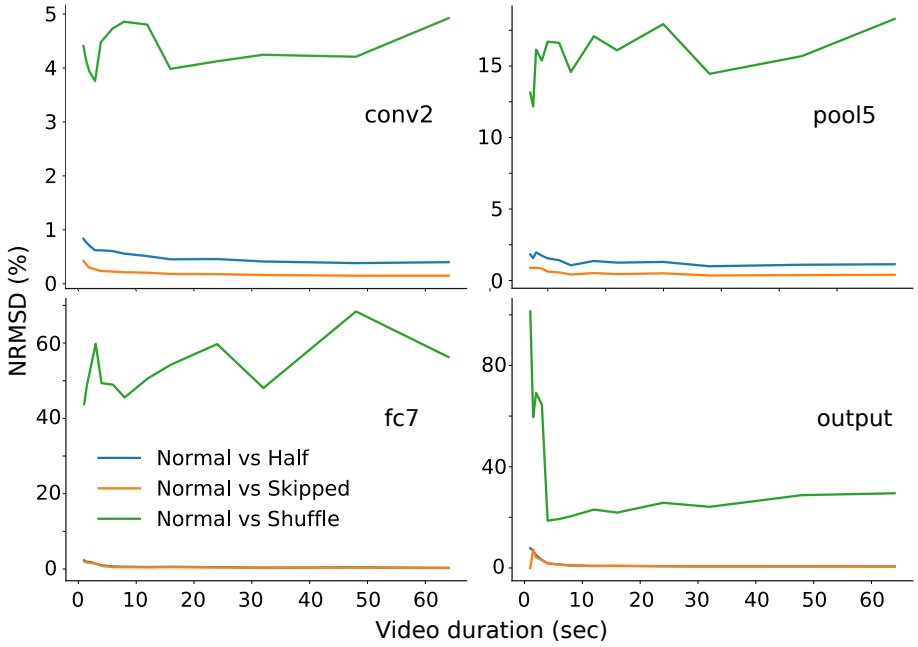


Figure 6: **(Extended figure)** Comparison of system accumulation of salient changes depending on input frame rate and composition of the input video. Each panel shows the normalised root-mean squared difference between the accumulated salient changes in the system when given the normal input video at 30 Hz, compared to input videos at half the frame rate, inputs videos with 20% of frames pseudo-randomly skipped, and input videos presented at 30 Hz (same as the normal input videos), but with the order of presentation of the video frames shuffled. The manipulations of frame rate (halving or skipping 20%) had little effect on the accumulated changes (blue and orange lines), while shuffling the order of presentation of the frames altered the accumulation of salient changes dramatically (green lines).

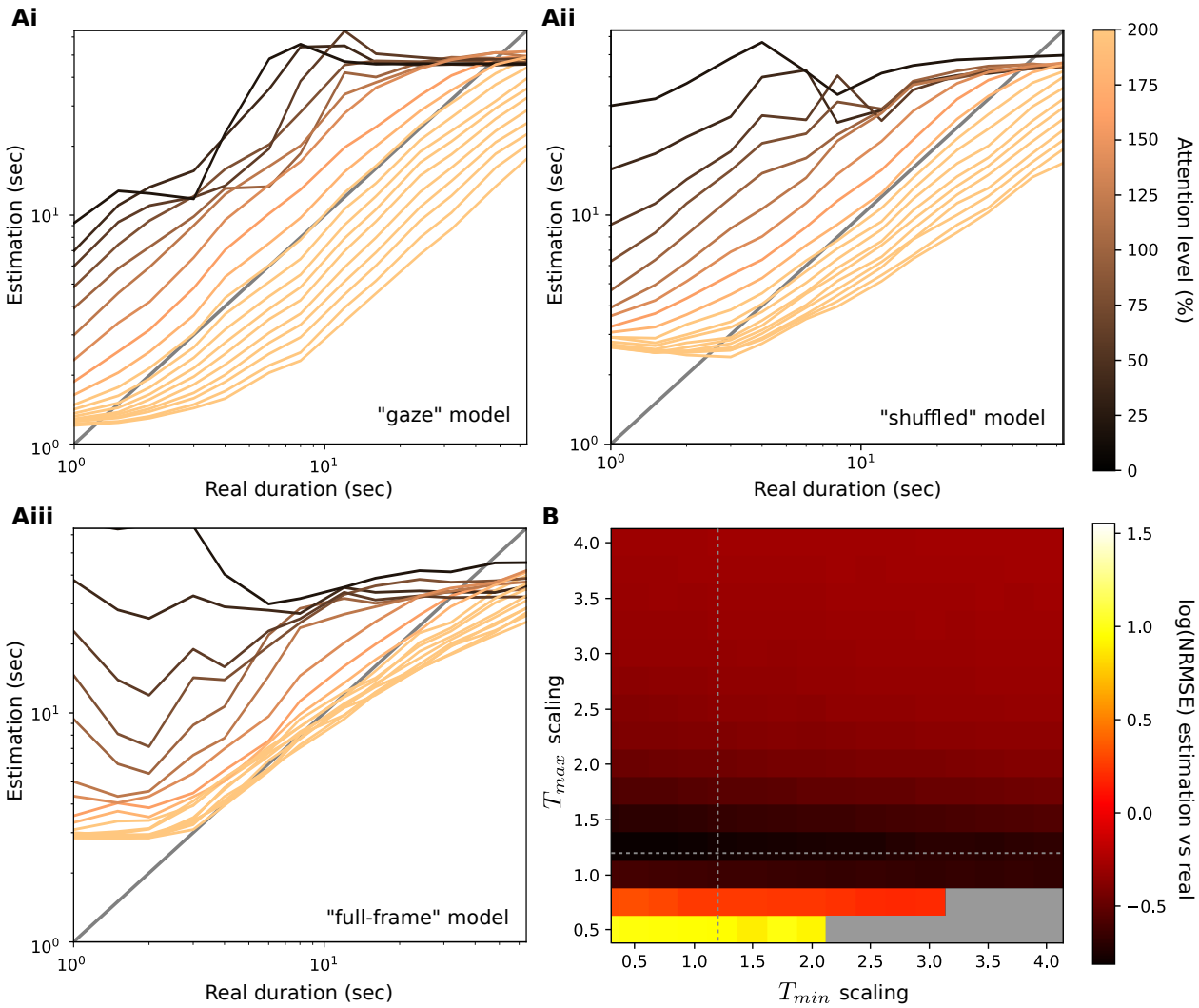


Figure 7: **(Extended figure)** Robustness of the temporal attention mechanism. **A:** Comparison of system duration estimation at different attention levels. Attention level refers to a scaling factor applied to the parameters $T_{max}$ and $T_{min}$, specified in Table 1. and equation 1 Each panel shows the performance for a different variant of the model ("Gaze", "Shuffled" and "Full-frame"). While changing the Attention level did affect duration estimates, often resulting in a bias in estimation (e.g. many levels of the "Full-frame" exhibit a bias towards over-estimation), across a broad range of Attention levels the models (particularly in the "Gaze" model) still differentiate longer from shorter durations, as indicated by the positive slopes with increasing real duration. For the models in Fig. 3, the following scalings were used: ("Gaze": 1.20, "Shuffled": 1.10 and "Full-frame": 1.06) as they were found to produce estimations most closely matching human reports. **B:** Normalised root mean squared error (NRMSE) of duration estimations of the "gaze" model versus real physical durations, for different combinations of values for the parameters $T_{max}$ and $T_{min}$ in equation (1). The gray areas in the heatmap represent combinations of values that cannot be defined. Dotted lines represent the chosen attention threshold scaling used for the "Gaze" model in Fig. 3.

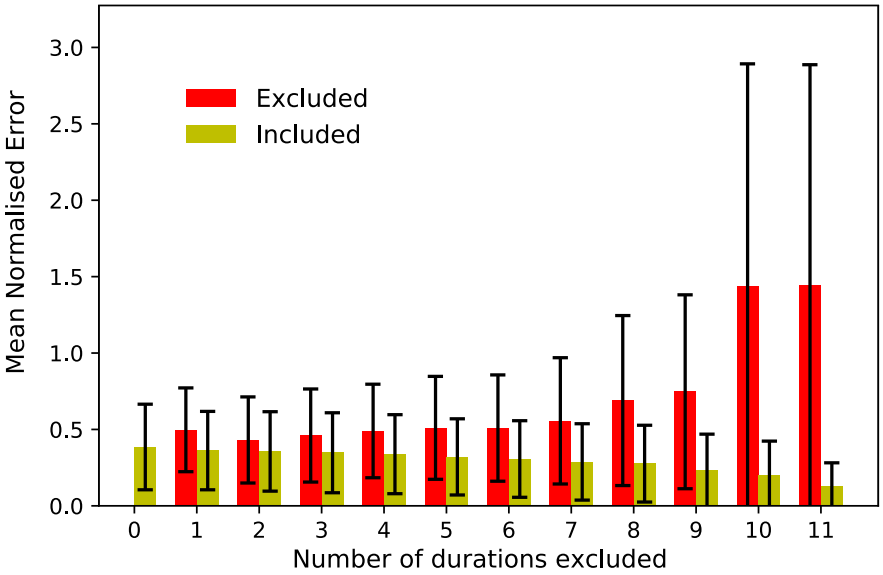


Figure 8: **(Extended figure)** Comparison of system performance by means of normalized duration estimation error, when a subset of testing durations were not used in the training process. For each pair of bars, 10 trials of N randomly chosen durations (out of 13 possible durations) have been excluded (x-axis). The SVR was trained on the remainder of the durations and tested on all durations. The errors for excluded and included trials are reported for each N. Only when excluding a large number of training levels (e.g. 10 out of 13 possible levels) does the estimation error get notably larger.