Clockless biologically-plausible architecture for temporal perception using convolutional neural networks

Zafeirios Fountas[1], **Kyriacos Nikiforou, David Bhowmik and Murray Shanahan** ([1]**zfountas@imperial.ac.uk**)
Department of Computing, Imperial College London, United Kingdom

Warrick Roseboom and Anil Seth
Department of Informatics & Sackler Centre for Consciousness Science
University of Sussex, United Kingdom

Abstract

The focus in understanding how humans or other animal species perceive time has been on mechanisms relying on fine alignments of intrinsic dynamical features, such as heartbeat, or the oscillatory cortical input to the striatum. Although they are often biologically consistent, these theories fail to address a number of important characteristics of temporal perception, including age-related differences, domain-specific biases and how interval timing can be scaled up to long durations (e.g. hours or days). In this work, we present a novel neural architecture that is able to make accurate time estimations of a given episode without the need for internal, clock-like processes. Instead, it relies on the amount of information that flows through hierarchical sensory areas and a feature detection mechanism, also employed for episodic memory formation. Using the power of convolutional neural networks for image classification, we built an implementation of this architecture in the visual domain. In this system, egocentric visual streams resulted in accurate interval estimations across time scales from 1-64s, both during real-time exposure to novel scenes and for episodic memory recall. Our results demonstrate that sufficient information exists in sensory classification networks to estimate duration without the need for any internal clock-like process.

Keywords: Time estimation; Hierarchical Neural Networks; Episodic Memory; Cognitive Architectures; Computational Modeling

The architecture

Low-level sensory hierarchy At the low level, the system performs visual object classification via AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), a deep convolutional neural network trained on 1000 categories of animals and objects. AlexNet receives the raw pixels of a single video frame as input, to output an array of probabilities indicating the class to which the objects of the image belong.

Feature detection To detect the changes of salient features and determine the flow of visual information, the architecture uses an adaptive "attention mechanism". The basic principle of the attention mechanism is that a new feature is detected if the new visual input is different from that experienced in the near past. If visual content is similar for an extended period, the attention threshold decreases so as to detect finer changes in the scene. The implementation of the attention mechanism involves calculating the Euclidean distance between the moving average of recent neuronal states in each of four chosen layers of the network and the new state corresponding to the current observed scene. If the distance exceeds the adaptive threshold value (Fig. 1 A), a new feature is registered in the accumulator of the corresponding layer (Fig. 1 B) and the threshold is reset. If the distance persistently remains below the threshold, the threshold decreases until a smaller change between successive frames exceeds it.

Episodic memory system In order to augment the capabilities of the proposed time estimation model, a separate Episodic Memory module is added to the architecture. Previous work (Bhowmik, Nikiforou, Shanahan, Maniadakis, & Trahanias, 2016) has shown that a network of firing-rate neurons can be trained to store and recall different sequences of a video in an episodic memory-like manner. In this system, memories are encoded in the activity of a neural network that is trained to reproduce sensory information through reproducing its own activity. When the highest-level threshold is exceeded, indicating that the overall context of the scene has changed, the storing of a new episode in memory is triggered.

Time duration estimation The proposed system provides enough information through the total number of accumulated features in each layer's accumulator for estimating how much time has elapsed from the start of accumulation. The total number of accumulated features can be used as input to a regression method, pre-trained to map these to physical duration. The same principle works for episodes recalled from memory: as the experience of the episode is reconstructed, it passes through the architecture and the salient features of the memory are accumulated. From this accumulation a duration estimation can be obtained for the remembered episode.

High-level sensory hierarchy At a higher level, the output of the neural network is directly connected to a Gaussian mixture model (GMM), trained to classify categories of episodes based on the recognized objects. GMMs provide a massive dimensionality reduction, where the 1000 object categories can be mapped to a small number of episodes. Due to the high-performance and consistency of AlexNet's classification, even when objects not included in the set of labels are misclassified, they are misclassified robustly and consistently.

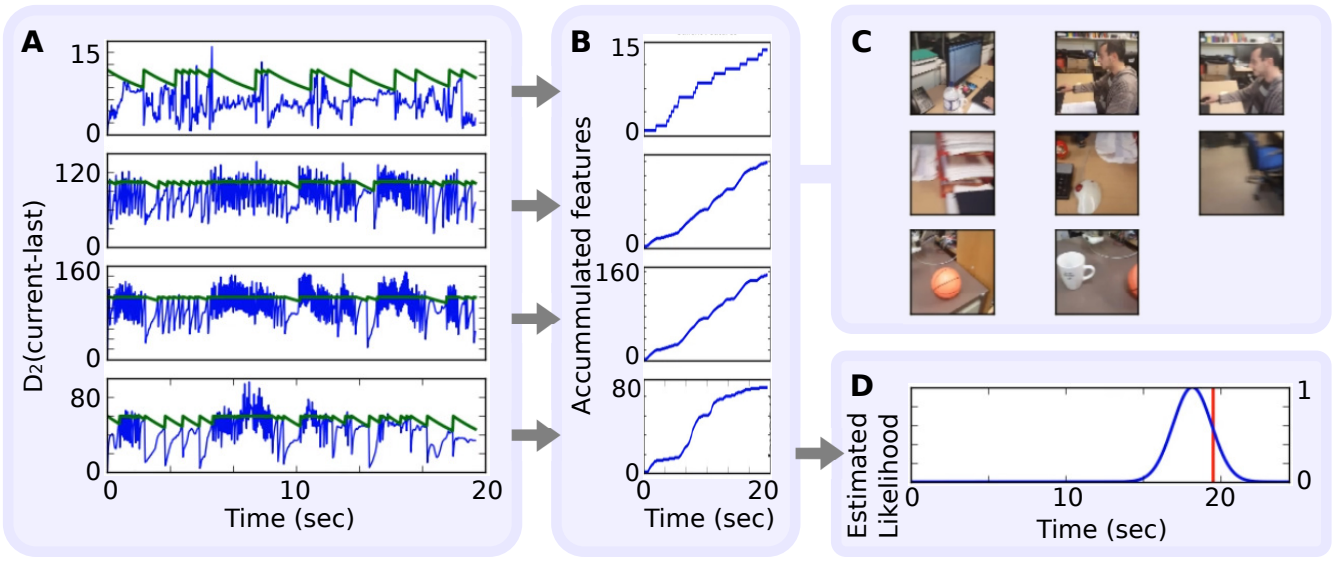

Figure 1: Instance of the time estimation process. **A:** Distance between the network state in each recorded layer during observation of the current image and the previously-recorded salient frame (blue curves) and the state of the attention threshold (green curves). **B:** Number of features that have been accumulated in each network layer. **C:** Frames of the experienced video that were considered salient in the highest network layer. **D:** Estimated time (blue curve) and ground truth (red line).

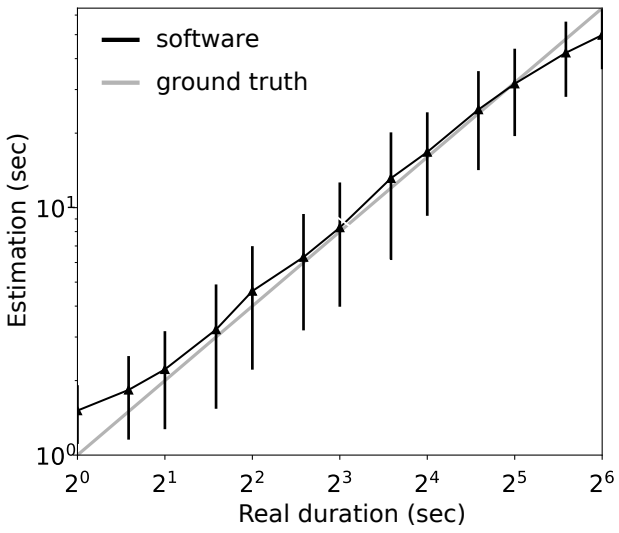

Figure 2: Average performance of the system. The error bars show standard deviation over a sample of 4290 trials.

Results & Discussion

Time estimates produced by the system replicate key qualitative aspects of human time perception such as a proportional increase in response variability with duration (scalar variability) and regression of responses towards the mean (Vierordt's law) (Fig. 2). Our findings show that an internal clock is not needed for human time perception, and provide a new approach to understanding this central aspect of experience.

Although multiple levels of biological abstraction have been used, the proposed architecture is designed in a modular manner where each module resembles a neural mechanism in the brain. An implementation based solely on a network of spiking neurons would not only be feasible (Ma, Beck, Latham, & Pouget, 2006; O'Connor, Neil, Liu, Delbruck, & Pfeiffer, 2013), but would also naturally introduce a number of desirable features via the fusion of multiple modules. For instance, the normalization induced by local inhibitory processes could potentially replace the attention mechanism described above.

Acknowledgments

This work has been supported by the EU FET grant (GA:641100) TIMESTORM – Mind and Time:Investigation of the Temporal Traits of Human-Machine Convergence.

References

Bhowmik, D., Nikiforou, K., Shanahan, M., Maniadakis, M., & Trahanias, P. (2016). A reservoir computing model of episodic memory. In *IJCNN* (pp. 5202–5209).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *NIPS* (pp. 1097–1105).

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

O'Connor, P., Neil, D., Liu, S.-C., Delbruck, T., & Pfeiffer, M. (2013). Real-time classification and sensor fusion with a spiking deep belief network. *Frontiers in Neuroscience*, *7*.