RODENT SLEEP ASSESSMENT WITH A TRAINABLE VIDEO-BASED APPROACH

Van Anh Le[†] *Mitchell Kesler*[⋆] *Jong M. Rho*[⋆] *Ning Cheng*[⋆] *Kartikeya Murari*[†]

[†]Electrical and Computer Engineering, University of Calgary, Calgary, AB
[⋆]Alberta Childrens Hospital Research Institute, Cumming School of Medicine, Calgary, AB

ABSTRACT

Assessment of sleep can reveal healthy physiology and behaviour, which are essential to study diseases and treatment. The primary approaches to quantify sleep in animal models are using invasive methods that require implantation of electroencephalogram (EEG) and electromyogram (EMG) electrodes. Those methods are resource-intensive and less than ideal for high-throughput screening. Several studies proposed using video processing to monitor sleep. Those approaches require high quality videos and an optimal threshold value which can be sensitive to different experiment settings. In this paper, we present a trainable video-based approach that can alleviate those limitations. We have come up with a set of effective features at frame-level which are then put into a recurrent neural network to capture long-term temporal features. The result obtained is highly correlated with EEG/EMG-defined sleep.

Index Terms— Long Short Term Memory, sleep assessment, animal behavior recognition

1. INTRODUCTION

Sleep is universally present in all species studied so far, from fruit flies to humans. Sleep is tightly regulated, and its loss impairs many physiological functions. However, it is disrupted in numerous pathological conditions, such as chronic pain, Parkinson's disease, and autism spectrum disorder. Conversely, its disturbance is considered a risk factor for many diseases, including depression, Alzheimer's disease, type 2 diabetes, cardiovascular disease, and obesity. Among mammals, mice have been increasingly used to characterize behavior for genetic and translational studies because they are small, low cost, and easy to breed. Excellent resources exist for identifying abnormal physiology and any related genes [1, 2]. Common assessments of sleep in mice are the electroencephalogram (EEG) and electromyogram (EMG) which involve invasive surgery followed by time to recover from surgery, which makes them impractical for high throughput screening. To tackle those problems, several methods have been proposed. Storch et al. [3] implanted a small magnet subcutaneously near the neck muscles of mice and their movement is determined through movements of the magnet that were registered via a sensor plate. Althought this method can be used for rapid pre-screen in animal sleep research, it still requires surgical intervention. In a different study, a non-invasive high-throughput system was described that used a single Polyvinylidine Difluoride sensor installed on the cage floor to measure pressure signals caused by movement [4]. Another non-invasive method used highly sensitive piezoelectric motion detectors attached to the cage floor [5]. Those film strips can capture movement due to respiration during sleep and other activities. Similarly, Brown et al. [6] introduced a system to measure activity in mice based upon passive infrared motion sensors. These approaches mentioned above have shown certain advantages over EEG/EMG methods, but they require specialised equipment, careful calibration and custom software, which are the reasons they are not widely used. An alternative high-throughput strategy is video analysis. The first significant work based on video processing [7] concluded that any period of continuous immobility of $\geq$40 s is likely to be sleep and their algorithm could achieve an average agreement of 92% with EEG/EMG recordings. Fisher et al. [8] proposed a complete system integrated with available commercial software and hardware that biologists can easily replicate. Besides using the $\geq$40 seconds of immobility identified in [7], the paper further investigated optimal threshold values determining immobility. Inspired by breakthroughs in the field of machine learning in recent years especially the advantage of Recurrent Neural Network (RNN) to capture sequence features, we propose a trainable method for detecting sleep in mice. In particular, we come up with a set of tracking and posture features and further learn temporal features of those features through a Long Short Term Memory (LSTM) Network. The result obtained is highly matched with EEG/EMG-based assessment.

2. RELATED WORK

Capturing meaningful features is crucial in machine vision tasks. Since videos are considered as time series signals and therefore the most challenging aspect in deep-based models is to deal with the temporal dimension [9]. A number of studies have proposed different approaches to extract spatio-temporal

Thanks to Alberta Innovates Technology Futures (AITF) for funding and NVIDIA corporation for donating the GPU used in this research.

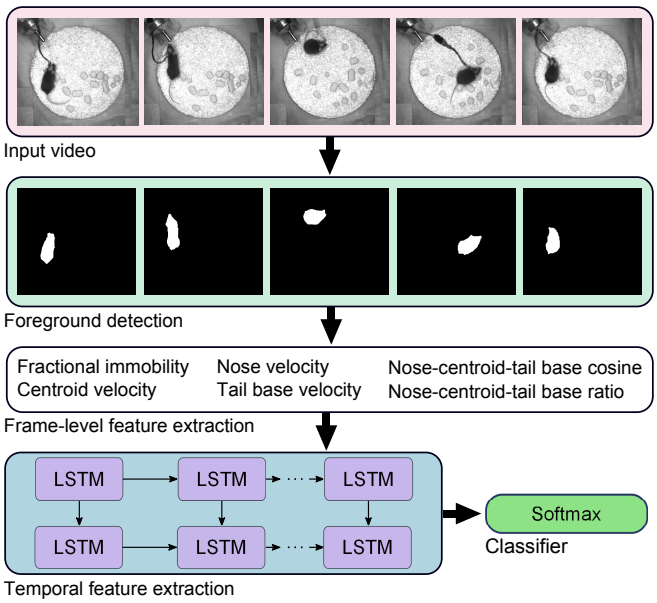

Fig. 1: The overview of proposed approach with 3 main stages: foreground detection, frame and temporal feature extraction and classification.

features for video processing. Recently, deep learning techniques have outperformed hand-crafted spatio-temporal feature extractors in various applications from tracking to gesture and action recogntion [9, 10]. One approach is to extend the convolution network [11, 12] along the temporal axis in order to capture temporal information, [13, 14]. However, the main downside is that the temporal axis is limited in order to fit existing memory. A second approach proposed a combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to build a model in which features extracted from a CNN are put in an RNN in sequence to learn temporal features [15, 16]. Despite of the dominance of CNN features in computer vision applications, almost all approaches in rodent behavior recognition use hand-crafted and tracking features [17, 18]. This can be explained by the fact that rodents do not have distinctive body features. Their limbs are too small compared with their bodies and there are lot of variations in terms of postures, which makes it expensive to produce a quality dataset for CNN to learn. Therefore, we believe that extracting describable features instead of CNN features and learning temporal features throught RNN can save a great amount of time and be easily replicated in different experiment settings.

3. PROPOSED METHOD

Fig. 1 shows the pipeline of the proposed approach which consists of 3 modules. The input video is first subtracted from estimated background to obtain foreground mask for pixels belonging to the mouse and then features at frame level are computed from foreground. Those features are fed into an LSTM network to extract long-term temporal features which are then classified in the last stage. We provide details of stages in following subsections.

3.1. Experimental setup

All procedures in this study were performed in accordance with the recommendations in the Canadian Council for Animal Care. The protocol of this study was approved by the Health Sciences Animal Care Committee of the University of Calgary. C57BL mice were housed in a humidity- and temperature-controlled room with a 12-h light/dark cycle and were fed ad libitum. On postnatal day 28, stainless steel screw EEG electrodes (Pinnacle Technology Inc., #8247) were implanted. For dorsal neck muscle EMG recordings, the nuchal muscles were surgically exposed, EMG electrodes were inserted underneath and sutured in place. Animals were allowed one week of recovery. EEG and EMG was acquired at 250 Hz. A camera was installed above the cage and pointed towards the bottom of the cage. Infrared lights and a visible light blocking filter were used to ensure constant illumination regardless of the light/dark cycle. Videos were synchronized with EEG and EMG, and recorded at 15 frames per second with a resolution of 480×540. Sleep epochs were identified when EMG power dropped below 50% of the median for more than 30 seconds and EEG 1-4 Hz delta activity was elevated [19].

3.2. Foreground detection

As long as the reasonable contrast between background and mice is maintained, the background can be considered stationary and is estimated by a temporal median filter in the very first frames. However, the primary challenge of our data is the movement of EEG tether and the bedding material in the field of view. These are sometimes significant in comparison with the mouse, and therefore we could not obtain the foregound mask by subtracting the video background directly. We first locate a region of interest throughout applying morphological processing, and then the correct foreground can be obtained as shown in Fig. 2.

3.3. Frame-level feature extraction

After obtaining the mouse body mask, the coordinates of key points are determined as in Fig. 3. Firstly, the body center ($\mathbf{C}$), the centroid of the foreground, is calculated as follows:

$$C_x = \frac{\Sigma_x x \Sigma_y I(x,y)}{\Sigma_{x,y} I(x,y)}; C_y = \frac{\Sigma_y y \Sigma_x I(x,y)}{\Sigma_{x,y} I(x,y)} \quad (1)$$

where $I(x,y)$ equals 1 if it belongs to the foreground F and 0 otherwise. Then the nose ($\mathbf{N}$) is located as the point on the

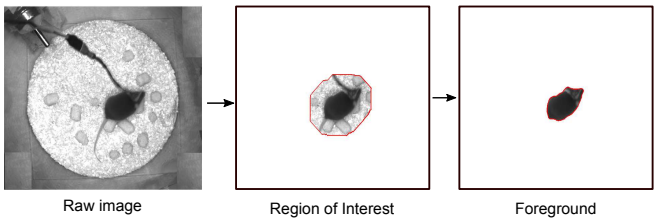

Fig. 2: Foreground detection: a region of interest is obtained by applying morphological operations on the raw image, background subtraction then gives the foreground.

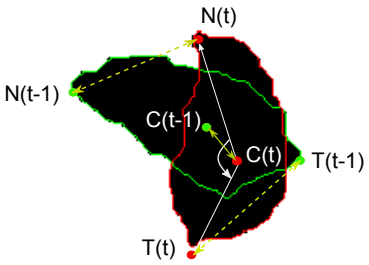

Fig. 3: Extracted features: the nose (N), centroid (C) and tail base (T) are used to find velocities. The area of overlap divided by the current foregound area is the fractional immobility. The angle subtended by N and T at C and the ratio of CN to CT are also indicated.

boundary that is the farthest from the body center. Similarly, tail base ($\mathbf{T}$) is the point on the boundary that is farthest from the nose. We then computed six features that represent the movement and posture of the mouse. The features include 'fractional immobility' between current and previous frames (FI); velocities of the centroid (V_C), the nose (V_N) and the tail base (V_T); cosine of the angle subtended by the nose $\mathbf{N}$ and the tail base $\mathbf{T}$ at the centroid $\mathbf{C}$ ($\cos\theta$); and the ratio of the centroid-nose to the centroid-tail base distance (R).

$$FI(t) = \frac{F(t) \cap F(t-1)}{F(t-1)} \tag{2}$$

$$V_C(t) = \|\mathbf{C(t)} - \mathbf{C(t-1)}\| \tag{3}$$

$$V_N(t) = \|\mathbf{N(t)} - \mathbf{N(t-1)}\| \tag{4}$$

$$V_T(t) = \|\mathbf{T(t)} - \mathbf{T(t-1)}\| \tag{5}$$

$$\cos\theta(t) = \frac{\mathbf{CN(t)} \cdot \mathbf{CT(t)}}{\|\mathbf{CN(t)}\|\|\mathbf{CT(t)}\|} \tag{6}$$

$$R(t) = \frac{\|\mathbf{CN(t)}\|}{\|\mathbf{CT(t)}\|} \tag{7}$$

The first four features are designed to capture movements while the last two designed to capture posture. The velocity features should be divided by the frame interval, however that is a constant scaling factor and can be dropped.

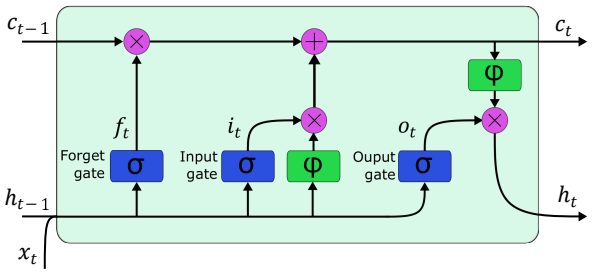

Fig. 4: Structure of LSTM cell

3.4. LSTM network for learning temporal features

3.4.1. LSTM cell

Recurrent Neural Networks have been used to capture sequential information, especially complex temporal dynamics. They are widely used in Natural Language Processing (NLP) such as speech recognition [20] and machine translation [21], video processing [22], and image captioning [23]. Given an input sequence $\mathbf{x} = (x_1, ..., x_T)$ to a conventional RNN, the hidden vector sequence $\mathbf{h} = (h_1, ..., h_T)$ and output vector sequence $\mathbf{y} = (y_1, ..., y_T)$ are computed by iterating the following equations from $t = 1$ to T.

$$h_t = \phi(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{8}$$

$$y_t = \phi(W_{hy}h_t + b_y) \tag{9}$$

where W terms are weight matrices, b terms are bias vectors and ϕ is an activation function.

However, training initial RNNs is difficult due to the problems of vanishing and exploding gradients [24]. Accordingly, many variants of RNN have been proposed and among the most widely used models is Long Short Term Memory (LSTM) shown in Fig. 4. The key to LSTMs [25] are memory cells for storing and outputing information. These cells are carefully regulated by structures called gates. LSTM updates memory cell c_t and hidden layer h_t as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{10}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \tag{11}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{12}$$

$$c_t = f_t c_{t-1} + i_t \varphi(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{13}$$

$$h_t = o_t \varphi(c_t) \tag{14}$$

where σ and φ are sigmoid and hyperbolic tangent functions. i, f, o, h and c are respectively the *input gate*, *forget gate*, *output gate*, *hidden state* and *cell state*.

3.4.2. Network architecture

Our architecture consists of two stacked LSTM layers, each layer with 64 memory cells and then a softmax classifier as shown in Fig. 1. The softmax classifier outputs the probability

of the behavior with index k at time t as follows:

$$P(k,t|\mathbf{x}) = \frac{exp(y_t^k)}{\Sigma_{k'} exp(y_t^{k'})} \quad (15)$$

4. EXPERIMENTAL RESULTS

EEG, EMG and videos were recorded simultaneously for 3 days in one mouse. The data of the first day was used to train the model and then the model was tested on the last two days. We evaluated our proposed method in the same way as [7], scoring the continuous recordings in 10-s epochs (8,640 epochs across 24 h). The average accuracy over 8,640 10-s epochs in 24 h is 95.4% with 95.2% for the first testing day and 95.6% for the second testing day in comparison with 92% in [7]. We also achieved an accuracy at a single frame level of 95.3%. To compare with the approach proposed in [8], we used the first day's data which is the training data of our model to determine the optimum immobility threshold value for estimating sleep. The sensitivity of immobility detection was varied from 50% to 100% and a duration of immobility of 40 s or greater was considered as sleep [7]. We identified 95.5% as the optimum sensitivity for immobility detection which is quite similar to the value of 95% determined in [8]. We then used this value to assess sleep in the testing data. In terms of accuracy over 8,640 10-s epochs in 24 h, the immobility-based method got 90.5% and 92.7% for the two testing days. Also, this method achieved an accuracy at a single frame level of 91.2%. Fig. 5 shows a comparison of our method, immobility-defined sleep and EEG-defined sleep over 48 hours. To investigate the effects of chosen features, we repeatedly leave one feature out and train the model again. Fig. 6 shows the accuracy of models corresponding each feature dropped out. Among six defined features, centroid velocity affects the overall accuracy the most.

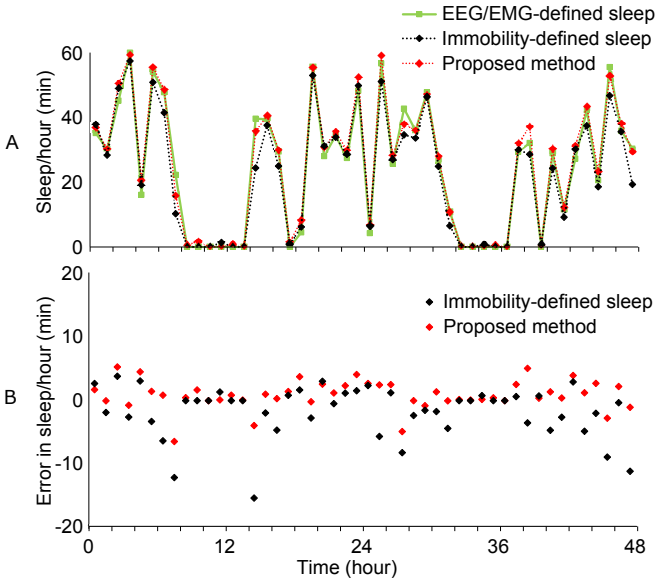

Fig. 5: (A) Comparison of sleep amounts in 1-h intervals across 48 hours. (B) Mean difference between two methods: The RMS errors in the proposed method and immobility-based method are 2.4 minutes and 4.4 minutes, respectively.

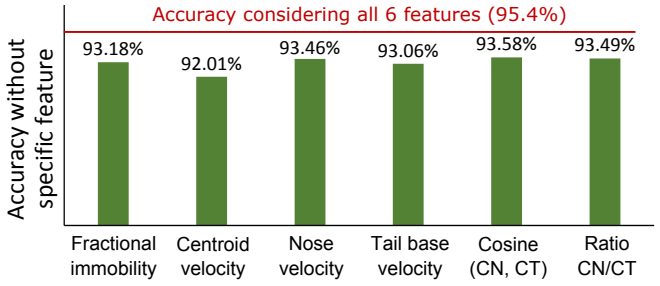

Fig. 6: Contribution of each feature to the overall accuracy

5. CONCLUSION

We have described a high throughput, trainable, video-based method to assess sleep in mice. Inspite of a tether for wired EEG/EMG recording, which posed significant barriers in background subtraction stages leading to sub-optimal feature extraction, the results were highly correlated with EEG/EMG-based assessment in both cumulative and instantaneous measures. This makes the system well suited for experiments requiring other tethered manipulations such as optogenetics or fiber photometry. With higher resolution and higher frame rate imaging, our approach can get even better, but at the cost of increased memory for image acquisition.

6. REFERENCES

[1] Thierry F Vandamme, "Use of rodents as models of human diseases," *Journal of Pharmacy & Bioallied Sciences*, vol. 6, no. 1, pp. 2, 2014.

[2] Danielle Simmons, "The use of animal models in studying genetic disease: transgenesis and induced mutation," *Nature Education*, vol. 1, no. 1, pp. 70, 2008.

[3] Corinna Storch, Arnold Höhne, Florian Holsboer, and Frauke Ohl, "Activity patterns as a correlate for sleep–wake behaviour in mice," *Journal of Neuroscience Methods*, vol. 133, no. 1-2, pp. 173–179, 2004.

[4] Kevin D Donohue, Dharshan C Medonza, Eli R Crane, and Bruce F O'Hara, "Assessment of a non-invasive high-throughput classifier for behaviours associated with sleep and wake in mice," *Biomedical Engineering Online*, vol. 7, no. 1, pp. 14, 2008.

[5] Aaron E Flores, Judith E Flores, Hrishikesh Deshpande, Jorge A Picazo, Xinmin Xie, Paul Franken, H Craig Heller, Dennis A Grahn, and Bruce F O'Hara, "Pattern

recognition of sleep in rodents using piezoelectric signals generated by gross body movements," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 2, pp. 225–233, 2007.

[6] Laurence A Brown, Sibah Hasan, Russell G Foster, and Stuart N Peirson, "Compass: continuous open mouse phenotyping of activity and sleep status," *Wellcome Open Research*, vol. 1, 2016.

[7] Allan I Pack, Raymond J Galante, Greg Maislin, Jacqueline Cater, Dimitris Metaxas, Shan Lu, Lin Zhang, Randy Von Smith, Timothy Kay, Jie Lian, et al., "Novel method for high-throughput phenotyping of sleep in mice," *Physiological Genomics*, vol. 28, no. 2, pp. 232–238, 2007.

[8] Simon P Fisher, Sofia IH Godinho, Carina A Pothecary, Mark W Hankins, Russell G Foster, and Stuart N Peirson, "Rapid assessment of sleep-wake behavior in mice," *Journal of Biological Rhythms*, vol. 27, no. 1, pp. 48–58, 2012.

[9] Maryam Asadi-Aghbolaghi, Albert Clapes, Marco Bellantonio, Hugo Jair Escalante, Víctor Ponce-López, Xavier Baró, Isabelle Guyon, Shohreh Kasaei, and Sergio Escalera, "A survey on deep learning based approaches for action and gesture recognition in image sequences," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 476–483.

[10] Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang, "Hierarchical convolutional features for visual tracking," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3074–3082.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[12] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri, "Convnet architecture search for spatiotemporal feature learning," *arXiv preprint arXiv:1708.05038*, 2017.

[14] Gul Varol, Ivan Laptev, and Cordelia Schmid, "Long-term temporal convolutions for action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[15] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1961–1970.

[16] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2678–2687.

[17] Zheyuan Wang, S Abdollah Mirbozorgi, and Maysam Ghovanloo, "An automated behavior analysis system for freely moving rodents using depth image," *Medical & Biological Engineering & Computing*, pp. 1–15, 2018.

[18] Elsbeth A van Dam, Johanneke E van der Harst, Cajo JF ter Braak, Ruud AJ Tegelenbosch, Berry M Spruijt, and Lucas PJJ Noldus, "An automated system for the recognition of various specific rat behaviours," *Journal of Neuroscience Methods*, vol. 218, no. 2, pp. 214–224, 2013.

[19] Hendrik W Steenland and Min Zhuo, "Neck electromyography is an effective measure of fear behavior," *Journal of Neuroscience Methods*, vol. 177, no. 2, pp. 355–360, 2009.

[20] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur, "Recurrent neural network based language model," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[21] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 3104–3112.

[22] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015, pp. 2625–2634.

[23] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 3156–3164.

[24] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] Felix A Gers, Nicol N Schraudolph, and Jürgen Schmidhuber, "Learning precise timing with lstm recurrent networks," *Journal of Machine Learning Research*, vol. 3, no. Aug, pp. 115–143, 2002.