Toward a deep convolutional LSTM for eye gaze spatiotemporal data sequence classification

Komi Sodoké
Université du Québec à Montréal
2098 Rue Kimberley
Montréal, QC
sodoke.komi@uqam.ca

Roger Nkambou
Université du Québec à Montréal
2098 Rue Kimberley
Montréal, QC
nkambou.roger@uqam.ca

Aude Dufresne
Université de Montréal
2900 Edouard Montpetit
Montréal, QC
aude.dufresne@umontreal.ca

Issam Tanoubi
Université de Montréal
2900 Edouard Montpetit
Montréal, QC
i.tanoubi@icloud.com

ABSTRACT

Eye gaze movements analysis are being increasingly used in many researches within learning context. Most of those researches analyses the eye movements fixations inside some areas of interest, the saccades trajectory and the scanpath. The eye gaze data are spatiotemporal sequences representing the dynamic of the eye fixations in the visual space over the time. In addition, they contain noises caused by different factors. The task of developing predictive model based on those raw spatiotemporal eye gazes' sequences is challenging. In this research, we present machine learning approaches that we have successfully used to address those challenges with high accuracy mainly with the deep convolutional LSTM architecture.

Keywords

Eye tracking; Deep learning; Spatiotemporal eye gazes sequences classification

1. INTRODUCTION

In some medical field such as anesthesiology, the visual perception is just a tip of the iceberg known as the "situational awareness." In fact, the clinician needs to develop the skills to see adequately the patient vital signs evolution over the time in order to build their understanding and interpretation of the clinical situation to perform their clinical reasoning. In this paper, we explore the following question: Can we tell novice and expert clinicians apart by analyzing only their eye-gaze movements to perform their clinical reasoning? Eye gaze data often contains noise which can be caused by many factors [10]. In addition, the consecutive data points generated by the eye movements trajectory over the time within the area of interest are spatiotemporal considering their order and their positions in the visual space. Ultimately, our experiments aim to understand key differences between novice and expert clinicians eye movements behavior during their clinical reasoning. Taken together, they will provide us insights to build an Intelligent Tutoring System (ITS) aiming to reinforce gradually the learning curve of novice clinicians with some cues from the experts behavioral implicit knowledge in terms of visual attention to perform a clinical reasoning in critical anesthesiology case.

Komi Sodoke, Roger Nkambou, Aude Dufresne and Issam Tanoubi "Toward a deep convolutional LSTM for eye gaze spatiotemporal data sequence classification" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 672 - 676

2. RELATED WORKS

The researches using eye-tracking and ITS can be summaries in two main axes according to Conati et al [6]. The first axe is the investigation of eye-tracking data as source of information for student modelling and personalized instructions. The second axe is leveraging the gaze data to attempt to understand relevant student behaviors. For that purpose, data mining techniques are often used to retrieve similarities, differences, etc. using the eye movements characteristics such us the fixations, the saccades and the scanpaths. Some researches also focus on mining eye-tracking patterns [18] . As a contribution, in this paper we propose predictive models using the sequence of the eye fixations positions over the time. These model will be used by the envisaged ITS to proactively classify eye fixations patterns as Novice vs Expert behavior in order to provide adequate eye movement tutoring services.

3. EXPERIMENTS AND DATASET

3.1 Experiments

An experiment has been conducted to collect eye gaze data for the research using an authentic task involving visual perception and clinical reasoning. Seven Novices and seven experts clinicians were asked to visualize a simulated clinical scenario to perform their clinical reasoning. A [Novice] is a resident clinician within the first or second year of the residency program (PGY1 or PGY2).[1] An [Expert] is a hospital staff member with more than 8 years experience. Each

[1]PGY refers to a North American scheme denoting the progress of postgraduates in their residency programs.

participant looked at a 23" HD monitor (1920x1080 px) on which the simulation was broadcasted. A Tobii TX300 eye tracker was attached to the monitor to record their eye-gaze movements. The simulation is based on the Cannot Intubate/Cannot Oxygenate (CICO) algorithm from the Difficult Airway Society to manage unanticipated difficult intubation in adults [9]. The simulation was scripted to integrate various unanticipated and realistic complications. It was recorded using high-fidelity settings and the video had a total duration of 13 minutes.

As a task, the participants were asked to verbalize their clinical reasoning using a think-aloud protocol (recorded with the eye tracker built-in microphone) while watching the simulation video. Specifically, they had to explain what they see in the different areas of interest (Figure 1) to perform their reasoning. In addition, the participants must explain what they would have done as clinician in charge in some key medical and situational awareness events (Table 1) identified throughout the simulation.

Figure 1: Areas of interest in the simulation

The display screen was divided in seven zones; each representing an area of interest (AOI).

3.2 Dataset

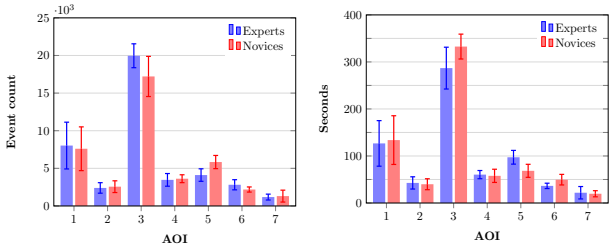
The eye tracker has an accuracy of 0.4 deg and was set to a sampling rate of 60 Hz. This means that a data point is collect each 17 ms. Each "data point" in the dataset is identified with a `{x,y,t}` tuple by the eye tracker. Overall, our eye-tracking dataset contains about 645k data points; i.e., 14 time series of around 46k points each. Each time series $T = \{\boldsymbol{p}^{(1)} \cdots \boldsymbol{p}^{(n)}\}$ is a sequence of 2D vectors, where each vector $\boldsymbol{p}^{(i)} = [x_i, y_i]$ represents the eye-gaze position at a given timestamp t_i.

4. PRELIMINARY ANALYSIS

4.1 Eye movements fixation analysis

First, we conducted preliminary analysis, aimed at providing exploratory insights. For that, we compare novices vs. experts using descriptive statistics on the fixation. For example, the result for the total fixation count and the total fixation duration within each AOI are shown in Figure 2.

These preliminary analysis results showed that both experts and novices have their highest total fixation duration on



(a) Fixation count mean (b) Fixation duration mean

Figure 2: Event count (2a) and mean fixation duration (2b). Error bars denote 95% confidence intervals.

the Technical view (AOI 3) and the General view (AOI 1). This result is further confirmed by the fixation count. Second, novices spent a significantly shorter amount of time at the Saturation view (AOI 5) than the experts ($M = 59$ vs $M = 107$ s, $p = .002$).Inversely, novices spent a significantly higher amount of time at the Technical view (AOI 3) than experts ($M = 382$ vs $M = 266$ s, $p = .042$). All other comparisons were not found to be statistically significant.

4.2 Eye movements behavior around the key events

The video recordings were annotated at different timestamps in terms of clinical keys events. The Table 1 provides an overview of such key event annotations.

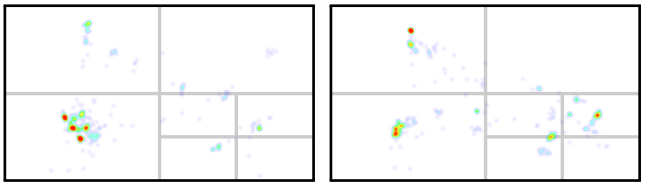
Focus Area	AOIs	Time	Description
Healthcare provider	1,3	02:41	Call for help
	1,2,3	03:35	Mask ventilation
	3	06:41	Installation of oropharyngeal cannula
	3	07:35	Use of video-laryngoscope
	1,3	08:33	Use of supra-glottic device
	1,3	09:39	Blue Code initiation
	3	10:32	Initiation of surgical airway
Patient	2	01:10	Impaired verbal response
	3	01:25	Eye closure
	1	02:09	Hypoventilation
Vital signs monitor	5,7	01:37	Desaturation
	4,6	08:33	Bradycardia
	5	10:22	Loss of the saturation signal

Table 1: Key events through the simulation video, together with their relation to the eye tracker AOIs.

With this video annotations, we rendered the heatmaps from raw eye-gaze coordinates corresponding to each key event. We considered eye movement data corresponding to 2 seconds of duration, 1 second before and 1 second after each key event timestamp, given that both eye fixations and reaction times occur typically around 500 ms [8, 15, 19]. Therefore, it allows to capture the eye gaze behaviour before and after each key event.

With this fine-grained video annotations and the observation of the incremental heatmaps around each key events, we observed more salient differences between novices and experts. For instance at 06:41 (Installation of oropharyngeal cannula) we observed a divergent eye movements behavior: both novices and experts focused on AOIs 3 and 5, but novices also focused in AOI 1 (Figure 3). These observations

suggest that both novices and experts have subtle different eye-gaze movement patterns most of the time, while sometimes they are similar. What is most important, these eye-gaze patterns vary over time, suggesting that both novices and experts tend to focus on different AOIs over time.



(a) Experts (b) Novices

Figure 3: Heatmap of the eye-gaze coordinates taking into account 1 second before and after the key events at 06:41

5. EXPERTISE CLASSIFICATION BASED ON EYE GAZE SEQUENCE COORDINATES

Taken together, the preliminary and the behavioral analysis suggest that we could build a classification model considering the eye-gaze movements coordinates over time. Based the outcome and observations from the preliminary analysis, we wondered if we could automatically learn these eye-gaze behaviors and discriminate clinicians' expertise accordingly; i.e., given a particular sequence of eye movements with their coordinates, can we predict if it is a novice or an expert eye movements behavior? That research objective is a two-class (binary) classification problem on spatiotemporal eye gaze data.

5.1 The challenges of sequential data classification

As discussed by Xing et al. [26], there are three major challenges in sequence classification. First, the vast majority of classifiers can only take input data as a vector of features. However, there are no *explicit* features in sequence data. Second, even with various features selection methods to transform a sequence into a set of features, the feature selection is far from trivial. The dimensionality of the feature space for the sequence data can be very high and the computation can be costly. Third, besides accurate classification results, in some applications, we may also want to get an "interpretable" classifier. As previously stated, building an interpretable sequence classifier is difficult since there are no explicit a priori features.

There are many approaches that have been proposed to address the problem of sequence classification. We will briefly discuss the two main categories: vector-based and model-based classification. In vector-based classification, a data sequence is transformed into a vector of features through feature selections. Then, we need a distance function to measure the similarity between a pair of sequences. The choice of distance measures is critical to the performance of these classifiers. For simple time series classification, Euclidean distance is a widely adopted option [26]. Since Euclidean distance is sensitive to distortions in time dimension, dynamic time warping (DTW) is proposed to overcome this problem and does not require two time series to be of the same length [13]. Dynamic time warping is usually computed by dynamic programming and has the quadratic time complexity. Therefore, it is computationally costly on a large data set. Using that vector representation of the data, sequences can be classified by a conventional classification method, such as support vector machines [24], decision trees [4], etc.

In model-based classification, given a class of sequences, an underlying model learns the probability distribution of each sequence. The simplest approach is the Naive Bayes sequence classifier [7]. It assumes that, given a class, the features in the sequences are independent of each other. However, this assumption is often violated in practice. A hidden Markov model (HMM) can learn the dependence among elements in sequences [1, 22], assuming that the system being modelled is a Markov process with unobserved states, where the state is described by a single discrete random variable. In contrast, neural networks do not have these assumptions. Moreover, HMMs can only deal with a limited number of step dependencies, while LSTMs can deal with long-term dependencies.

5.2 Machine Learning Models

Since the objective is to predict the expertise given a particular eye movements sequence, the full-length eye-gaze sequence are sliced in smaller parts. Each instance is a fixed-size time series consisting of the raw eye-gaze coordinates; i.e., `(x,y)` points (a 2D vector). For our experiment, we used sequence slices of length $s = 1000$, which represent eye-gaze sequences (time series) of about 17 seconds each. Finally, because of the small number of participants, we choose the LOOCV (Leave-one-out Cross Validation) as a resampling technique.

Two machine learning architectures were developed to perform the eye gaze spatiotemporal data classification : a WKM-kNN architecture and a DeepConv-LSTM architecture

5.2.1 WKM-kNN architecture

The WKM-kNN architecture is a composition of warped K-means (WKM) with k-nearest neighbor (k-NN). WKM is a fast algorithm for clustering data sequences based on distances, and has outperformed comparable approaches in the task of sequence classification [17]. In addition to providing a compact representation of data sequences, WKM makes them robust to noise or distortions in such data. The input to this model is a time series (a sequence of 2D vectors), and the output is either novice or expert, according to the k-NN classifier.

The WKM algorithm capitalizes in the sequentiality of the data and starts with a suitable initial partition [16], by using piecewise linear interpolation, which results in a non-linearly distributed initial partition of the data. Then, WKM iterates over the data points using a K-means-like optimization procedure. Finally, the k-NN classifier is a non-parametric instance-based learning method, which is among the simplest of all machine learning algorithms. In this work we use $k = 1$ for classification.

To sum up, the WKM-kNN architecture proceeds as follows: first WKM compresses a time series of length n into c disjoint homogeneous segments (or "elementary units") with $1 < c \ll n$, then the centroid of each segment is used as input to a 1-NN classifier. As in any other clustering algorithm, the number of sequence chunks c should be provided as input. Therefore, because the optimum c for classification is unknown in advance, we tested different values of c, increasingly from 1 (each time series is reduced to a single 2D vector) to 500 (half of the original sequence length).

5.2.2 *DeepConv-LSTM architecture*

The DeepConv-LSTM architecture is a neural network consisting of a convolutional block followed by a recurrent block (Figure 4).

The recurrent block is a deep long short-term memory (LSTM) network. LSTMs are a type of recurrent neural networks (RNNs) capable of learning long-term dependencies in time series by selectively remembering patterns for long duration and were developed to deal with the *exploding* and *vanishing gradient* problems of traditional RNNs [2, 20]. LSTMs have outperformed many other approaches in a variety of tasks, such as handwriting [11] and speech recognition [12], therefore we adopted this model to analyze eye-gaze sequences. In addition, inspired by recent work that has applied convolutional neural networks (CNNs) to sequence modeling with great success [3], we add a one-dimensional convolutional layer (temporal convolution) to the network input followed by a max pooling layer, which then feed the consolidated features to the LSTM. In other words, a CNN layer learns spatial features which are then learned as sequences by an LSTM layer. This way, we combine the spatial structure learning properties of CNNs with the sequence learning of LSTMs. On the other hand, the max pooling layer is a sample-based discretization process, with 3 goals in mind: (1) reduce the input dimensionality, by filtering the initial data representation; (2) avoid over-fitting, by providing an abstracted form of the data representation; and (3) lower the computational cost, by reducing the number of parameters to learn.

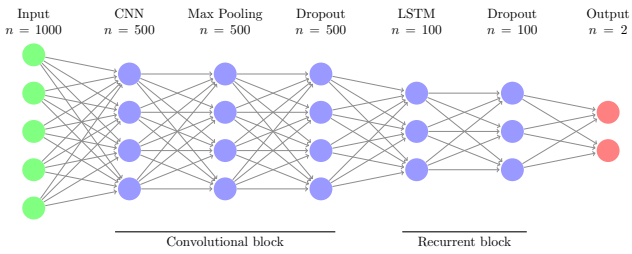


Figure 4: Deep learning network topology. Notes: The drawing is simplified to avoid visual clutter. Each layer dimensionality (n) is denoted below their title.

Overall, the chosen network has 41311 trainable parameters with the topology shown in Figure 4. The network input is the sequence slices (a sequence of 2D vectors), whereas the network output is either novice or expert. Both the CNN and max pooling layers have a kernel size of 2. The LSTM layer is fully connected with 100 neurons. The dropout layers have a probability of 0.2, since it is the recommended value for most machine learning scenarios; see e.g. [21, 23]. These layers have the effect of reducing overfitting and improving model performance.

We trained the neural network with 60 epochs and a batch size of 256 (mini-batch training) on an i5 CPU @ 3.30 GHz with 16 GB of RAM. After each epoch, the model is evaluated against the testing partition, to get an idea of how well the model is performing during training, after which the data is shuffled for the next epoch. The model was fit using the efficient ADAM optimization algorithm [14] with binary crossentropy as loss function.

5.3 Results

The Table 2 summarizes the results, in terms of classification accuracy. Together with the confidence intervals, we report the Area Under the ROC Curve (AUC), which is a one of the standardized measure of a classifier's performance. Since the WKM-kNN architecture was tested at different segmentation values c, we report the best classification accuracy result, which was achieved with $c = 4$ segments.

Model	Accuracy (%)	95% Conf. Int.	AUC
WKM-kNN	72.6	[71.1, 74.2]	0.74
DeepConv-LSTM	84.2	[84.9, 86.4]	0.86

Table 2: Summary of the classification results. Confidence intervals are calculated according to the Wilson method for binomial distributions [25].

6. CONCLUSION AND FUTURE WORKS

This research objective is to collect factual eye gaze data from clinicians during a clinical reasoning task. Given a particular sequence of eye movements, with their coordinates; can we predict if it is a novice or an expert clinician eye movements behavior ? To answer that question, we built two machine learning models for the binary classification. The deep learning architecture provides an overall better results achieving a very competitive level of accuracy (84.2%) on eye-gaze spatiotemporal data. These results are particularly striking given the fact that we used the *raw* gaze coordinates coming from the eye tracker. The key for the success of a deep neural network classifier is the ability to automatically learn hidden features or intermediates representations in the input data.

The future work is to use the eye-gaze spatiotemporal data classifier outcome and the recorded expert clinical reasoning during the key events as one of the key milestone for the ITS domain model. Also, we have not studied the impact that eye-gaze sequence length may have on model accuracy, though in general shorter sequences should be harder to classify. Some studies argue that humans make informed decisions in a matter of milliseconds [5] although we suspect this is strongly correlated to the application at hand. Therefore, analyzing this possible impact of sequence length on accuracy is another interesting avenue for future work, which in turn opens many research questions. For example: What is the minimum sequence length that maximizes classification accuracy? Is there any upper bound from which we can devise useful eye-gaze information? Does more segment context overlap lead to better model generalization?

7. REFERENCES

[1] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Stat.*, 37(6):1554–1563, 1966.

[2] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.*, 5(2), 1994.

[3] J. Bradbury, S. Merity, C. Xiong, and R. Socher. Quasi-recurrent neural networks. *CoRR*, 1611.01576, 2016.

[4] N. A. Chuzhanova, A. J. Jones, and S. Margetts. Feature selection for genetic sequence classification. *Bioinformatics*, 14(2):139–43, 1998.

[5] M. Cohen, E. C.E., and F. J. Oscillatory activity and phase-amplitude coupling in the human medial frontal cortex during decision making. *J. Cogn. Neurosci.*, 21(2):390–402, 2009.

[6] C. Conati, N. Jaques, and M. Muir. Understanding attention to adaptive hints in educational games: An eye-tracking study. *International Journal of Artificial Intelligence in Education*, 23:136–161, 2013.

[7] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian"on the optimality of the simple bayesian classifier under zero-one loss" classifier under zero-one loss. *Machine Learning*, 29(2–3):103–130, 1997.

[8] A. H. Duc, P. Bays, and M. Husain. Eye movements as a probe of attention. *Prog. Brain Res.*, 171:403–11, 2008.

[9] C. Frerk, V. Mitchell, A. McNarry, C. Mendonca, R. Bhagrath, A. Patel, E. O'Sullivan, N. Woodall, and I. Ahmad. Difficult airway society 2015 guidelines for management of unanticipated difficult intubation in adults. *Br. J. Anaesth.*, 115(6):827–48, 2015.

[10] J. Goldberg and J. Helfman. Comparing information graphics: A critical look at eye tracking. *Conference on Human Factors in Computing Systems - Proceedings*, 04 2010.

[11] A. Graves, S. Fernández, M. Liwicki, H. Bunke, and J. Schmidhuber. Unconstrained online handwriting recognition with recurrent neural networks. In *Proc. Intl. Conf. on Neural Information Processing Systems*, NIPS'07, pages 577–584. Curran Associates Inc., 2007.

[12] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, ICASSP'13, pages 6645–6649. IEEE Press, 2013.

[13] E. J. Keogh and M. J. Pazzani. Scaling up dynamic time warping for datamining applications. In *Proc. Intl. Conf. on Knowledge Discovery and Data Mining*, KDD'00, pages 285–289. ACM Press, 2000.

[14] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *Intl. Conf. on Learning Representations*, ICLR'15. arXiv, 2015.

[15] R. J. Krauzlis, L. Goffart, and Z. M. Hafed. Neuronal control of fixation and fixational eye movements. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 372(1718), 2017.

[16] M. H. Kuhn, H. Tomaschewski, and H. Ney. Fast nonlinear time alignment for isolated word recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, ICASSP'81, pages 736–740, 1981.

[17] L. A. Leiva and E. Vidal. Warped k-means: An algorithm to cluster sequentially-distributed data. *Inf. Sci.*, 237(10):196–210, 2013.

[18] A. Li, Y. Zhang, and Z. Chen. Scanpath mining of eye movement trajectories for visual attention analysis. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 535–540, 2017.

[19] J. L. Orquin and S. M. Loose. Attention and choice: A review on eye movements in decision making. *Acta Psychol.*, 144:190–206, 2013.

[20] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *Proc. Intl. Conf. on Machine Learning*, ICML'13, pages 1310–1318. JMLR.org, 2013.

[21] V. Pham, T. Bluche, C. Kermorvant, and J. Louradour. Dropout improves recurrent neural networks for handwriting recognition. *CoRR*, 1312.4569, 2013.

[22] L. R. Rabiner and B. H. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

[23] N. Srivastava. *Improving neural networks with dropout.* PhD thesis, University of Toronto, 2013.

[24] V. Vapnik. *The nature of statistical learning theory.* Springer, 1995.

[25] E. B. Wilson. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.*, 22:209–212, 1927.

[26] Z. Xing, J. Pei, and E. Keogh. A brief survey on sequence classification. *SIGKDD Explor. Newsl.*, 12(1):40–48, 2010.