Investigating Static and Sequential Models for Intervention-Free Selection Using Multimodal Data of EEG and Eye Tracking

Mazen Salous
University of Bremen
Bremen, Germany
salous@uni-bremen.de

Felix Putze
University of Bremen
Bremen, Germany
felix.putze@uni-bremen.de

Tanja Schultz
University of Bremen
Bremen, Germany
tanja.schultz@uni-bremen.de

Jutta Hild
Fraunhofer IOSB
Karlsruhe, Germany
jutta.hild@iosb.fraunhofer.de

Jürgen Beyerer
Fraunhofer IOSB
Karlsruhe, Germany
juergen.beyerer@iosb.fraunhofer.de

ABSTRACT

Multimodal data is increasingly used in cognitive prediction models to better analyze and predict different user cognitive processes. Classifiers based on such data, however, have different performance characteristics. We discuss in this paper an intervention-free selection task using multimodal data of EEG and eye tracking in three different models. We show that a sequential model, LSTM, is more sensitive but less precise than a static model SVM. Moreover, we introduce a confidence-based Competition-Fusion model using both SVM and LSTM. The fusion model further improves the recall compared to either SVM or LSTM alone, without decreasing precision compared to LSTM. According to the results, we recommend SVM for interactive applications which require minimal false positives (high precision), and recommend LSTM and highly recommend Competition-Fusion Model for applications which handle intervention-free selection requests in an additional post-processing step, requiring higher recall than precision.

KEYWORDS

Multimodal data, EEG, Eye tracking, Competition Model, Recall, Precision

ACM Reference Format:
Mazen Salous, Felix Putze, Tanja Schultz, Jutta Hild, and Jürgen Beyerer. 2018. Investigating Static and Sequential Models for Intervention-Free Selection Using Multimodal Data of EEG and Eye Tracking. In *Workshop on Modeling Cognitive Processes from Multimodal Data (MCPMD'18), October 16, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3279810.3279841

1 INTRODUCTION

Humans engage in several cognitive processes while interacting with computers. Observing data streams from those cognitive processes yields heterogeneous but complementary multimodal data for predicting such cognitive processes. For example, during a typical selection task in a GUI, a user generates eye gaze activity, motor activity (mouse clicks), and of course brain activity during the whole process. While eye tracking gaze data describes the spatial domain of such a human action, analyzing brain activity allows models to investigate the temporal course of that action. Thus, eye tracking data can be used to track the spatial trajectory of eye movements. Additionally, brain activity in Human Computer Interaction (HCI) has been widely analyzed by observing Electroencephalography (EEG) data to instantaneously detect Event Related Potentials (ERPs). Consequently, a combination of both data modalities (EEG+Gaze) offers an opportunity to utilize the spatio-temporal properties of human action in general, and especially in a GUI selection task.

We can exploit this multimodal EEG+Gaze data to enable intervention-free selection of targets in a GUI without any explicit user intervention. This concept has been introduced by Putze et al. [11] to enhance a user interface by supporting it with a fallback mechanism to recover missed selections without manual interventions. They used a static classification model (a Support Vector Machine, SVM) to detect occurrences of target objects from a large stream of non-target distractors. Although SVM is recommended for Brain Computer Interfaces (BCIs) [8], both data modalities, EEG and gaze, encapsulate sequential dependencies between subsequent data samples. For example, a data sequence of EEG+Gaze observed while tracking a target on screen differs from a data sequence observed while freely scanning the screen waiting for targets. In this paper, we re-use SVM as a discriminant-static model, and also investigate the feasibility of a sequential model based on Long-short Term Memory [5] (LSTM) networks. LSTM is a recommended model to exploit sequential dependencies in data for classification. Moreover, given that models' miss-classifications of the same task may not necessarily overlap [7], we further investigate a confidence-based fusion model using both SVM and LSTM. We compare between all three investigated models according to different performance measures.

Thus, the main contributions of this paper are: 1) Compare the state-of-the-art static classifier with a sequential model to capture sequential dependencies, 2) Investigate the fusion of the static and the sequential models, and 3) upon investigation, to recommend which model to use for different applications of intervention-free selection.

(Pre-print) MCPMD'18 , October 16, 2018, Boulder, CO, USA
2018. ACM ISBN 978-1-4503-6072-2/18/10…$15.00
https://doi.org/10.1145/3279810.3279841

2 RELATED WORK

In this section we discuss related works in two main directions: 1) Combination of EEG+Gaze as multimodal data and 2) prediction models for biosignal data such as EEG and eye tracking gaze data. First, for multimodal data of EEG+Gaze, several studies combined the Event Related Potential ERP of EEG data with the fixation event of Gaze data as fixation-related potentials (FRPs). While some FRPs studies such as [1, 3, 6] intended only to gain insights into cognitive processes without investigating single-trial classification of FRPs, others did investigate FRPs in classification problems. For example, Finke et al. [4] investigated the feasibility of FRPs in HCI contexts by successfully classifying FRPs on a single-trial basis in a natural static scene. Moreover, Shishkin et al. [15] classified FRPs as intentional vs. spontaneous fixations to differentiate between intentional and spontaneous eye movements while looking at objects. Choi et al. [2] discussed enhancing the EEG-based UI speller by incorporating eye tracking data. They used eye tracking data in P300 task to minimize the relevant matrix to 3 x 3 rows and columns determined by the user's gaze position. Thus, they obtained relevant P300 EEG signals when each of the 3 x 3 rows and columns is highlighted, and this increased the accuracy of selecting the correct character. Putze et al. [10] showed that the use of eye tracking and EEG allows spatio-temporal event localization in a simple dynamic scenario.
Second, we discuss prediction models for each data type in our task. For EEG data, Lotte [8] mentioned in his comprehensive EEG processing tutorial that for Brain Computer Interfaces (BCI), the most used classifiers are the discriminant types, and notably the Linear Discriminant Analysis (LDA) and Suport Vector Machine (SVM). As mentioned in the introduction, both EEG data and gaze data encapsulate sequential dependencies within data items in different time stamps. While static models, such as SVM, can model sequential dependencies in data using postprocessing such as smoothing, there exist sequential models which have a specific structure intended to model sequential dependencies. Recurrent Neural Network (RNN), and especially Long-short Term Memory (LSTM) [5] is recommended for modeling such sequential data. A data sequence can be passed to LSTM as consecutive time steps. LSTM treats each time step consecutively by an LSTM cell. LSTM consists of many LSTM cells interconnected to each other, where the output of one cell acts as an additional input for the next cell. The cells are supported with gates to memorize or forget the processed data item. This special structure enables LSTM to store and retain information from previous time steps. LSTM successfully exploits sequential dependencies and tackles the vanishing problem which impedes typical RNNs for long sequential dependencies. Putze and Salous [12, 13] showed that LSTM outperformed discriminant static baseline models in classifying based on sequential behavioral data. For combining different classifiers, an early work by Kittler et al. [7] discussed the importance of such combination because the sets of feature vectors wrongly classified by different models would not necessarily overlap, potentially allowing combination of both models for an improved result.

3 INTERVENTION-FREE SELECTION TASK

In several HCI applications, users have to continuously monitor the screen to select target symbols from large number of moving objects. The intervention-free selection task aims at the implicit selection of target symbols on a screen only from monitoring the user without requiring an explicit response. Selecting targets on the screen includes two main cognitive processes: 1) continuous scanning of the screen with the eyes (which can be captured by eye tracking) and 2) decision making to decide between targets and non-targets (which can be captured by EEG). This section discusses the experiment used to record multimodal data of EEG and eye tracking while a user is counting moving targets on screen and distinguishing them from non-targets.
We process such multimodal data to model this task as a binary classification problem, where we fit classifiers using training data combined from both eye tracking and EEG data to classify moving objects as targets or non-targets. Such multimodal data based classifiers could work online in the future when integrated in real HCI applications.

3.1 Experimental setup

Eight subjects (7 male, 1 female) participated in the experiment. The mean age was 25.8 years, ranged between 21 and 44 years. Using a gray screen with only a stream of objects moving downward vertically at variable speeds (depicted as circles and squares), the task was to count only targets. Figure 1 shows a screenshot of the task. To draw the user's attention to the coming event (target or

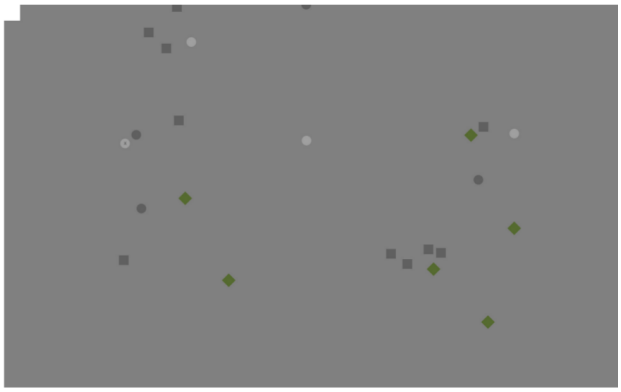

Figure 1: Screenshot of the task. Objects move downward vertically.

non-target), the letter 'x' appears in one random moving object to notify the user that an event is coming soon. Only one object at the same time can be in this state and there is a minimum time of 2 seconds between two appearances of the letter 'x'. After 1500 ms, the 'x' is replaced either by the letter 'e' to denote a target event or by one of four potential other letters ('h', 'n', 'c', 'p') to denote a non-target event. As some non-target letters are very similar to the target letter 'e', targets could only be recognized when focusing directly on the letter. The second letter is visible for additional 500 ms and then replaced by the original object symbol (square or circle). The letters events never occur within the top and bottom 2 cm of the screen. The task is designed this way to collect relevant multimodal data of brain activity (EEG data) and gaze activity (eye tracking data). Thus, We synchronize the collected EEG and gaze events

according to each coming indicator 'x'. Technically, we achieve such synchronization by using a light sensor fixed on the upper left corner of the screen where the screen flashes a light at this area for each coming indicator 'x'. Then, the signals collected from the light sensor will be used to synchronize the collected EEG and gaze events, and thus we pass only relevant data to classifiers.

3.2 Eye Tracking Data: Gaze Features

To collect eye tracking data, we used a Tobii X60 eye tracker with a sampling frequency of 60 Hz. We assume two different gaze activity patterns when tracking the vertical trajectory of a potential target and when scanning the screen waiting for a coming target. To distinguish between those patterns, we calculate four gaze features: gaze residual, gaze direction, piece-wise linear approximation (PLA) and fixation duration.

The gaze residual feature is the residual error of the linear regression line in x and y dimensions of the estimated eye gaze trajectory. While targets' trajectories are expected to follow the linear movement of the object (smaller residual error), the gaze trajectory during a free scan of the screen may not be linear at all (greater residual error). The gaze direction is calculated as the slope of the object trajectory regression line. This feature contributes to target selection because targets move downward and this fact leads to similar slopes for targets' trajectories compared to arbitrary trajectories recorded during free scanning of non-targets. The PLA feature is the number of segments in a sequence before the angle between two segments deviates too much from a straight line for the first time. Similarly, PLA is expected with higher values for targets compared to non-targets due to the downward trajectory of targets. Finally, the fixation feature defines how much time the user follows an object. Thus, fixation times while tracking targets are clearly greater than fixation times of free scanning of non-targets; that is, one moves her/his focus from non-target to another freely waiting for a potential coming target. To calculate this feature, we used the I-VT algorithm [14] to extract the onsets and offsets of fixations.

3.3 Brain Activity Data: EEG Features

For recording EEG data, we used a BrainProducts actiCHamp Recorder with a 32 electrode actiCAP. The electrodes were positioned according to the international 10-20 system. The reference electrode was positioned at the location Fz. First, EEG prepossessing was applied to get only relevant features, avoid artifacts and to filter the relevant frequencies. For example, we applied a 47 Hz to 53 Hz bandstop filter (for line noise artifacts) and a 0.25 Hz high-pass filter (for sweat and myogenetic artifacts). To further concentrate on the relevant low-frequency parts of the signal, we subsequently applied a 15 Hz low-pass filter.

While monitoring objects, we expect a different neural response to the perception of targets' events than to the perception of non-targets. Thus, the final EEG features are designed to capture the typical ERP response following the perception of targets. For this purpose, we use the down-sampled signal at all recorded electrode sites. The EEG data is segmented into windows with a window size of 600 ms, because this is the duration expected while a target event is occurring (See section 4 for more details on windowing and labeling). Finally, The means of all segments for all electrodes form the EEG feature vector.

4 DISCRIMINANT-STATIC AND SEQUENTIAL CLASSIFIERS

Putze et al. [11] used a discriminant-static SVM model with RBF kernel for classification in the intervention-free selection task. The joint data stream (EEG+gaze) is segmented into windows with a window size of 600 ms. This window size was chosen as it corresponds roughly to the duration for which a target event is present and thus it provides enough information to catch an ERP in the EEG signal and to detect meaningful gaze patterns in the eye tracking data. The windows are overlapped with a window shift of 15 ms. This overlapping with short window shift is necessary to well align such overlapping windows to target events which can happen at any point of time during the window period of 600 ms. In other words, if windows were not overlapped, consecutive windows (each of size 600 ms) could be badly aligned to targets because a target event can begin anytime e.g. at the 500th ms of window X (bad alignment to that window). In contrast, for the overlapping windows with such a short shift of 15 ms, whenever a target event happens, it will match a couple of overlapping windows, and thus, it will be assigned to them. Feature vectors of both data types (EEG and Gaze) are normalized to zero mean and unit variance and then combined in a joint feature vector. For the static model SVM, a label (target/non-target) is assigned for each window and this depends on whether a target has been tracked or not during that window.

In this work, we replace the static SVM model with a sequential model based on LSTMs. In contrast to the static model, the LSTM classifier is passed a sequence of s feature vectors and assigns one label to the full sequence. The sequence is moved with a shift of 1 across the sequence of all feature vectors. For comparison, we drop the first s feature vectors from the SVM classification for which no corresponding LSTM result is available.

Putze et al. used a combination of oversampling and undersampling for handling the class imbalance. While undersampling can be applied to the training data of an LSTM, there are no established techniques for oversampling of sequential data. Therefore, we limit the analysis to undersampling with a fixed ratio of 90% for the majority class.

Sequence length is a parameter to be optimized in LSTM. We optimize meta parameters for both models using 10-fold cross-validation on the training data. At the end, we also combined SVM and LSTM based on their confidence and compared the fusion model metrics with those metrics of each individual model, SVM and LSTM. See next Subsection 4.1 for fusion model details and the evaluation in Section 5 for details of metrics and comparison analysis.

4.1 Fusion Model: Confidence-based Competition

SVM and LSTM may agree or disagree on the label (target and non-target) of a given feature vector. We analyzed the predicted labels of both SVM and LSTM and found different patterns in how labels were predicted. For example, we searched for gaps in predicted labels, where a non-target gap is a predicted target in a sequence of

non-targets (probably a false positive) and a target gap is a predicted non-target in a sequence of targets (probably a false negative). We found that target gaps are very rare in both models, however, we found that the SVM model has more non-target gaps than the LSTM model. This is plausible as the SVM does not model temporal dependencies between neighboring feature vectors. Figure 2 illustrates an example of such gaps.

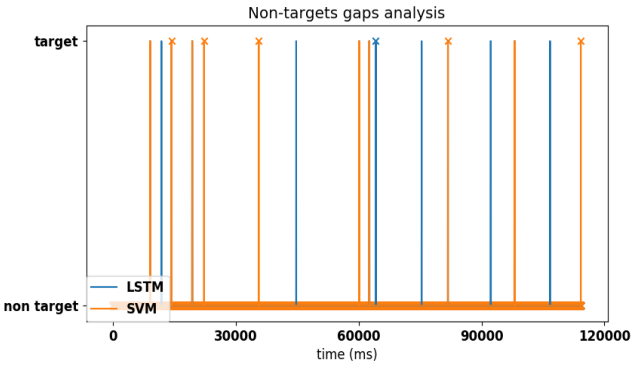

Figure 2: Non-targets' gaps are scattered with "X" to distinguish them from predicted sequences of targets. "X" points are wrongly classified targets (false positives). SVM has more non-targets gaps than LSTM.

While the SVM makes more non-target gaps than LSTM, further analysis shows that in total, the LSTM makes many more false positives than the SVM caused by sequences of false positives. These different characteristics in label assignment indicate that it might be beneficial to combine both models and apply a competition mechanism to judge between their contrary predictions. For a fusion of two information sources, we need a mechanism to estimate which of the sources to trust in case of a prediction mismatch. For this purpose, we use the prediction confidence of each model, i.e. for a given feature vector, if SVM and LSTM predict contrary labels, the predicted label with greater confidence will be pooled as the label for this feature vector.

The confidence for the SVM model represents the probability that a given data point belongs to a particular class using Platt scaling [9]. The closer the data-point to the classifier boundary, the lower confidence it has. The confidence in LSTM is calculated also as the probability yielded by the softmax activation function. Confidences are normalized to be comparable between the two different models. To check the validity of such confidences, we plot for each model and each label the distribution of confidences for correctly vs. incorrectly predicted labels in histograms together with a fitted probability density function (PDF). Figure 3 shows that the confidence distribution of the majority class (non-target) clearly discriminates between correctly-predicted and incorrectly-predicted non-targets, and even for the minority class (targets), the confidence distribution discriminates (although less clearly) between correctly-predicted targets and incorrectly-predicted targets. The observable differences indicate the validity of the chosen confidence estimators for the fusion of both models.

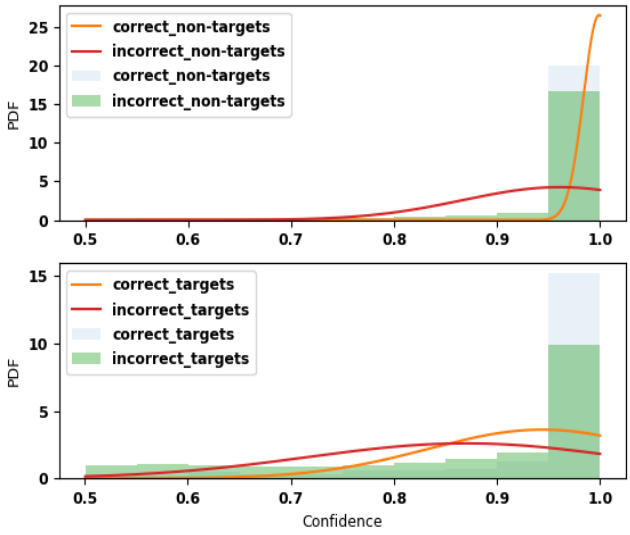

Figure 3: Confidences distribution for correctly-predicted labels is greater than the confidences distribution for incorrectly-predicted labels.

5 EVALUATION

Evaluation of the classification models is performed in a person-dependent 10-fold cross-validation. Since we deal with a highly imbalanced class distribution in this classification task, we evaluate the classifiers according to precision, recall and Fβ-Score rather than accuracy which would be strongly biased toward the majority class (non-targets). We evaluate and compare between the SVM, LSTM and confidence-based competition-fusion model.

Since in this analysis windows are overlapped with a very short window shift of 15 ms, we evaluate the models not only in a window-by-window fashion but also perform an evaluation which allows a certain tolerance to avoid instances where insignificant temporal shifts lead to an exaggerated number of reported errors. Tolerance here is defined as the number of windows allowed in both sides of the ground truth when we check a match between a predicted target and its corresponding label in the ground truth. For example, $tolerance = 0$ means that we should find an exact match between a predicted target and its corresponding label in the ground truth to report a true positive, while $tolerance = 3$ reports a true positive for a predicted target even if the corresponding target in the ground truth is shifted up to 3 windows in any of both sides.

Fβ-Score weights both precision and recall using the parameter β. First, we evaluate for equal weights for both precision and recall, i.e. $\beta = 1$. In further analysis, we will also look at situations in which precision or recall is more important than the respective other metric and change β accordingly.

Figure 4 shows the comparison between SVM, LSTM and confidence-based fusion models for different tolerance values. We see that recall

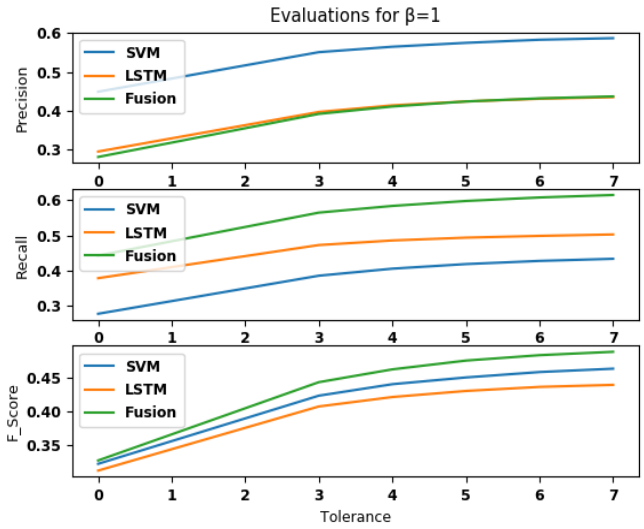

Figure 4: Sequential model LSTM is more sensitive but less precise than the static SVM model. Furthermore, the competition-fusion model is more sensitive than both SVM and LSTM and has a precision score close to that of the LSTM.

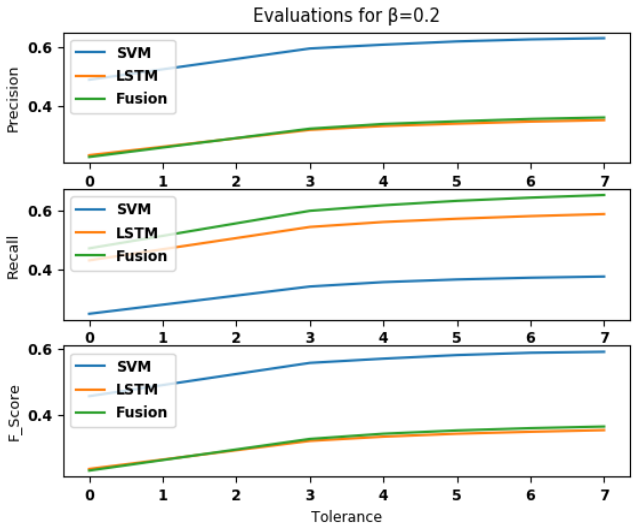

Figure 5: SVM outperforms LSTM and Competition-Fusion when precision has higher priority than recall.

increases from SVM to LSTM and further increases in confidence-based fusion. The recall improvements are statistically significant ($p < 0.05$, calculated using a paired t-test on the result of individual iterations). A classifier with a higher recall is more sensitive to find the positive samples (targets). Precision results, however, show that the SVM outperforms LSTM and confidence-based fusion. The difference in precision is also significant (paired t-test result: $p < 0.05$). The Fβ-score is similar between all three models but with slight advantage for the competition-based fusion model over the individual models. Although figure 4 shows the improvement of Fβ-score as a slight one, a paired t-test proves it is a significant improvement with $p = 0.001$. The advantage of the competition fusion model can be explained by the fact that it further improves the recall compared to both SVM and LSTM without decreasing the precision comparing with LSTM. This result shows that the confidence-based fusion does not merely provide an average between the precision and recall values for the two individual models but is able to outperform them. To study situations in which precision should be considered more important than recall, we optimized the models again for $\beta = 0.2$ ($\beta < 1$ gives higher priority to precision). The results show, as expected, higher Fβ-scores for the SVM over both LSTM and competition-fusion due to its high precision (see Figure 5). The difference in Fβ-score is also proven by a paired t-test as a significant one ($p < 0.05$).

Similarly, we optimized the models again for $\beta = 1.5$ (giving priority to recall over precision) and the results in Figure 6 show, as expected, a pronounced benefit (statistically significant: $p < 0.05$ in a pair t-test) of the competition-fusion model compared to the individual models due to its advantage in recall.

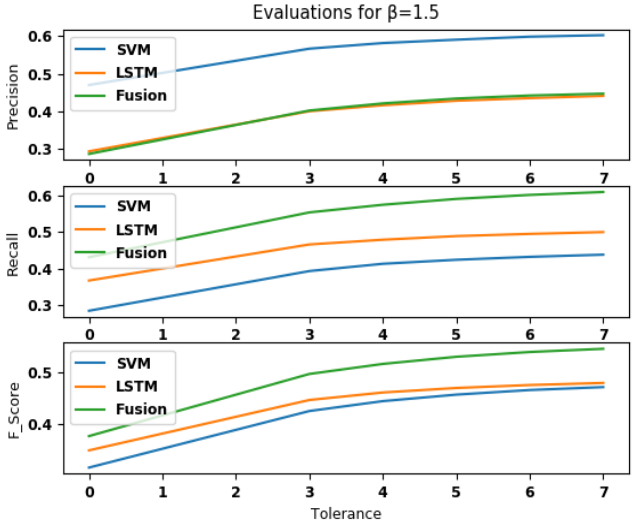

Figure 6: LSTM outperforms SVM when recall has higher priority than precision. Competition-fusion outperforms both LSTM and SVM in this case.

6 CONCLUSIONS

This paper investigated different models –discriminant-static and sequential models– for intervention-free selection task using multimodal data of EEG+Gaze. A discriminant-static model has been

investigated since it is recommended for EEG classification, and a sequential model has been investigated because both data modalities, EEG and gaze, encapsulate sequential dependencies between neighboring data points. Since SVM and LSTM may not overlap in their classifications, a confidence-based fusion model is also investigated. Based on the results, we can judge which model is to be used based on the requirements of the intended application: For an application which uses the intervention-free selection technique to immediately and interactively handle the selection requests, precision should have higher priority because many false positives can drastically impede the interaction. Thus, the SVM model would be recommended for such applications as it demonstrates the highest precision. On the other side, if an application handles intervention-free selection requests in a post-processing step or through another user, then recall would be more relevant than precision as we want to miss as few targets as possible in the initial selection. For this case, the SVM model should be completed by the LSTM model in the competition-based fusion to achieve the optimal performance. For future work based on such results, we would develop two different HCI applications accordingly (each application prioritizes either precision or recall) and investigate our models in real-time.

7 ACKNOWLEDGEMENT

This work has been done within the project DINCO "Detection of Interaction Competencies and Obstacles". We thank the German Research Foundation (DFG) for funding DINCO project under the reference number PU 613/1-1.

REFERENCES

[1] Thierry Baccino and Yves Manunta. 2005. Eye-fixation-related potentials: Insight into parafoveal processing. *Journal of Psychophysiology* 19, 3 (2005), 204–215.
[2] Jong-Suk Choi, Jae Won Bang, Kang Ryoung Park, and Mincheol Whang. 2013. Enhanced perception of user intention by combining EEG and gaze-tracking for brain-computer interfaces (BCIs). *Sensors* 13, 3 (2013), 3454–3472.
[3] Olaf Dimigen, Werner Sommer, Annette Hohlfeld, Arthur M Jacobs, and Reinhold Kliegl. 2011. Coregistration of eye movements and EEG in natural reading: analyses and review. *Journal of Experimental Psychology: General* 140, 4 (2011), 552.
[4] Andrea Finke, Kai Essig, Giuseppe Marchioro, and Helge Ritter. 2016. Toward FRP-based brain-machine interfacesâĂŤsingle-trial classification of fixation-related potentials. *PloS one* 11, 1 (2016), e0146848.
[5] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[6] Florian Hutzler, Mario Braun, Melissa L-H Võ, Verena Engl, Markus Hofmann, Michael Dambacher, Helmut Leder, and Arthur M Jacobs. 2007. Welcome to the real world: validating fixation-related brain potentials for ecologically valid settings. *Brain Research* 1172 (2007), 124–129.
[7] Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas. 1998. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence* 20, 3 (1998), 226–239.
[8] Fabien Lotte. 2014. A tutorial on EEG signal-processing techniques for mental-state recognition in brain–computer interfaces. In *Guide to Brain-Computer Music Interfacing*. Springer, 133–161.
[9] John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10, 3 (1999), 61–74.
[10] Felix Putze, Jutta Hild, Rainer Kärgel, Christian Herff, Alexander Redmann, Jürgen Beyerer, and Tanja Schultz. 2013. Locating user attention using eye tracking and EEG for spatio-temporal event selection. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 129–136.
[11] Felix Putze, Johannes Popp, Jutta Hild, Jürgen Beyerer, and Tanja Schultz. 2016. Intervention-free selection using EEG and eye tracking. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 153–160.
[12] Felix Putze, Mazen Salous, and Tanja Schultz. 2018. Detecting Memory-Based Interaction Obstacles with a Recurrent Neural Model of User Behavior. In *23rd International Conference on Intelligent User Interfaces*. ACM, 205–209.
[13] Mazen Salous and Felix Putze. [n. d.]. Behaviour-Based Working Memory Capacity Classification Using Recurrent Neural Networks. ([n. d.]).
[14] Dario D Salvucci and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*. ACM, 71–78.
[15] Sergei L Shishkin, Yuri O Nuzhdin, Evgeny P Svirin, Alexander G Trofimov, Anastasia A Fedorova, Bogdan L Kozyrskiy, and Boris M Velichkovsky. 2016. EEG negativity in fixations used for gaze-based control: Toward converting intentions into actions with an eye-brain-computer interface. *Frontiers in neuroscience* 10 (2016), 528.