https://doi.org/10.2352/ISSN.2470-1173.2021.3.MOBMU-109

Task Evoked Pupillary Response for Surgical Task Difficulty Prediction via Multitask Learning

Beilei Xu, Wencheng Wu, Lei Lin; Rochester Data Science Consortium, University of Rochester, Rochester, NY
Rachel Melnyk, Ahmed Ghazi; Simulation Innovation Laboratory, University of Rochester Medical Center, Rochester, NY

Abstract

In operating rooms, excessive cognitive stress can impede the performance of a surgeon, while low engagement can lead to unavoidable mistakes due to complacency. As a consequence, there is a strong desire in the surgical community to be able to monitor and quantify the cognitive stress of a surgeon while performing surgical procedures. Quantitative cognitive-load-based feedback can also provide valuable insights during surgical training to optimize training efficiency and effectiveness. Various physiological measures have been evaluated for quantifying cognitive stress for different mental challenges. In this paper, we present a study using the cognitive stress measured by the task evoked pupillary response extracted from the time series eye-tracking measurements to predict task difficulties in a virtual reality based robotic surgery training environment. In particular, we proposed a differential-task-difficulty scale, utilized a comprehensive feature extraction approach, and implemented a multitask learning framework and compared the regression accuracy between the conventional single-task-based and three multitask approaches across subjects.

I. Introduction

Cognitive stress refers to the load on working memory as experienced by subjects when conducting cognitive tasks. It is a critical aspect that can affect human behavior and performance when carrying out complex mission-critical tasks such as aviation and military command and control [1]. It is well known that excessive cognitive load can substantially impede the performance leading to human errors. On the other hand, too little cognitive load can make a person feel complacent, which is error prone as well. In the past, attention in the evaluation of a surgeon's performance has been focused on the outcomes and skill levels, i.e., total task time, instrument pathlength, smoothness and completeness. Studying cognitive stress during surgical procedures has just started to attract attention in the medical community.

A great number of techniques for quantifying cognitive load have been investigated for various tasks. Chen et al. [2] summarized the techniques into four primary methods, i.e., subjective surveys, performance-based, physiological-based, and behavior-based approaches. Zhou et al. [3] provided an analysis and comparisons of those methods. In surgery, wireless sensors have made physiological measures, such as heart rate variability (HRV), galvanic skin response (GSR), electroencephalogram (EEG) based brain activity measures and eye activity measures, more feasible in the operating room to provide objective measurements of surgeons' cognitive state with minimum inference to the tasks at hand [4]-[6]. In this paper, our study is focused on the task evoked pupillary responses (TEPR), i.e., the linear increase in pupil dilation, to quantify a subject's cognitive stress while performing certain procedures. Studies have shown that eye-tracking metrics have strong associations with the perceived memory load [7], [8]. In the medical field [9], the change of pupil diameter as an objective measure of cognitive load has been used in evaluating novices and trained physicians as they answered clinical related questions. Recently, several studies have used gaze-related metrics to compare the differences in cognitive load to the difficulties in surgical tasks [6], [10]. In [10], various statistical tests such as Kruskal–Wallis test, Dunn's test, and Spearman's test were utilized to compare performance differences among different tasks. However, it is unfeasible to rely on those statistical tests continuous measurements. Wu et al. [6] used the Naïve Bayes algorithm to classify the perceived cognitive loads into either high or low load based on the total NASA-Task-Load-Index scores. They employed nine features including two demographic features and seven eye-tracking features (left/right pupil diameter mean, left/right pupil diameter standard deviation, gaze entropy, fixation duration, and percent-of-eyelid-closure) and achieved a reasonably good classification accuracy among the tasks. However, it is not clear whether the cognitive factors were the most relevant features to the classification results. Their study also indicated the weakness in using pupil-diameter mean when predicting fine-grained task difficulties such as the three-level Suture Sponge task.

In this paper, instead of extracting a handful hand-crafted time-series features, e.g., mean, standard deviation, entropy, we employed a multitask approach along with the method that extracts a comprehensive set of features [11] for predicting task difficulties based on the TEPR of a subject while performing a set of exercises on the da Vinci Surgical System (dVSS). The contributions of this paper include: 1) utilized a comprehensive feature extraction and selection step; 2) proposed a differential-task-difficulty scale to convert the subjective ratings to objective metrics; and 3) implemented a multitask learning framework to utilize the relatedness across subjects to improve the performance. The remaining paper is organized as following: Section II briefly introduces the methods used in this study; Section III describes the experimental set up and data collection; Section IV discusses the results; and Section V concludes our work and discusses a few future directions.

II. Method: TSFresh & Multitask Learning

With the rise of Internet of Things, methods for time-series analysis become increasingly important for many applications. Traditional approaches relying on extracting a few basic features such as min, max, mean, etc. are no longer sufficient. However, choosing the right features heavily depends on experts' knowledge. The process often becomes a blind search resulting in a set of features that are still not the most relevant representation of the task of interest. Exploring additional features or identifying relevant features is essential to achieve the desired performance. For TEPR-based analysis, even though many features have been employed such as the mean, range, number of peaks of TEPR, etc., for evaluating cognitive stress, there is no consensus of their accuracy or effectiveness. In our research, instead of preselecting a handful features, we utilize the TSFresh, a comprehensive feature extraction method [11] followed by a feature selection step [12].

In [13], Fulcher and Jones built a comprehensive collection of time series features including more than 9000 features from 1000 different feature-generating algorithms used in various fields such as medicine, finance, industrial applications and so on. The proposed algorithm and the framework to extract the features had subsequently been implemented in a Python package called TSFresh [11], which we utilized in this study to extract features from the continuous eye-tracking of each exercise. Those features followed by a feature selection step were then used to predict the task difficulty levels through a multitask learning framework.

Multitask learning seeks to improve the generalizability of a learning task by exploiting the relations (structure) among different, but related tasks. In multitask learning, the related tasks are learnt simultaneously by extracting and utilizing appropriate shared information across tasks. Learning multiple related tasks simultaneously and effectively increases the sample size for each task and improves the prediction performance, especially when the training sample size is small for each task. Multitask learning has demonstrated its successes in many fields such as the disease progression prediction at each time point as shown in [14]. In multitask learning, task relatedness is introduced as a regularization term that can be written in a general form as:

$$\min_{\boldsymbol{W}} \big(\mathcal{L}(\boldsymbol{W}) + \Omega(\boldsymbol{W})\big), \tag{1}$$

where $\boldsymbol{W}$ is the parameter to be estimated from the training samples, $\mathcal{L}(\boldsymbol{W})$ is the empirical loss on the training set and $\Omega(\boldsymbol{W})$ is the regularization term that encodes the task relatedness. Various forms of $\Omega(\boldsymbol{W})$ have been explored as summarized in [15]. In this study, we evaluated three different forms of the regularization term, (i) the trace-norm; (ii) the extension of the Lasso regularized method to multitask learning; and (iii) the robust method.

The trace-norm method tries to capture the multitask relatedness by constraining the models from different tasks to share a low dimensional subspace, i.e., $\boldsymbol{W}$ is of low rank, and to make the problem solvable, the rank function on $\boldsymbol{W}$ can be further reduced to a trace norm, and if we take the loss as a least-square loss between the inputs and outputs, Equation (1) can be rewritten as:

$$\min_{W} \sum_{i=1}^{t} \left\| \boldsymbol{W}_i^T \boldsymbol{X}_i - \boldsymbol{Y}_i \right\|_F^2 + \lambda \|\boldsymbol{W}\|_* , \tag{2}$$

where Xi denotes the input matrix of the ith task, Yi denotes the corresponding output(s), t is the total number of tasks, Wi is the model for task i, and □□is a regularization parameter controls the rank of W [15].

The extension of the Lasso method in multitask learning is to share the parameter controlling the sparsity among all different tasks as written in:

$$\min_{W} \sum_{i=1}^{t} \left\| \boldsymbol{W}_i^T \boldsymbol{X}_i - \boldsymbol{Y}_i \right\|_F^2 + \rho_1 \|\boldsymbol{W}\|_1 + \rho_{L2} \|\boldsymbol{W}\|_F^2, \tag{3}$$

where $\boldsymbol{X}_i$, $\boldsymbol{Y}_i$, $\boldsymbol{W}_i$, i, and t follow the same definition in (2), the regularization parameter ρ_1 controls sparsity, and the optional ρ_{L2} regularization parameter controls the ℓ_1-norm penalty.

Most multi-task learning formulations, such as (2) and (3), assume that all tasks are relevant, which is however not the case in many real-world applications. Robust multitask learning (RMTL) was proposed aiming at identifying irrelevant (outlier) tasks when learning from multiple tasks. One such method proposed by [16] formulated the RMTL by decomposing the W in (1) into two components as written in:

$$\min_{W} \sum_{i=1}^{t} \left\| \boldsymbol{W}_i^T \boldsymbol{X}_i - \boldsymbol{Y}_i \right\|_F^2 + \rho_1 \|\boldsymbol{L}\|_* + \rho_2 \|\boldsymbol{S}\|_2, \tag{4}$$

subject to W = L + S, where Xi, Yi, Wi, i, and t follow the same definition in (2), the introduced regularization parameter ρ1 controls the low rank regularization on the structure L similar to (1) and the ρ2 regularization parameter controls the penalty on S. The low rank structure L in (4) captures task-relatedness and the group-sparse structure S detects outliers, i.e., if a task is not an outlier, then it falls into the low rank structure L with its corresponding column in S being a zero vector; if not, then the S matrix has non-zero entries at the corresponding column. Although many other formulism such as the one discussed in [17] have also been proposed, in this paper, we evaluated the three approaches defined in (2)-(4). It is noted that the choice of the regularization parameters can be learned through a small-training dataset during model training.

III. Experiments

Participants: This study was approved by the university's institutional review board. Our dataset consists of five subjects from three different experience levels, i.e., two expert surgeons (Expert-1 and Expert-2), two novices with no prior experience (Untrained-1 and Untrained-2) and one novice (Trained-1) with prior exposure of the instrument/exercise before the study. We had to drop one trained subject due to too few exercises he/she was able to accomplish.

Robotic System and Tasks: The da Vinci Skills Simulator (dVSS) was used in our experiments. Among the available simulated exercises, we selected the Suture Sponge exercise, in which the subject is required to thread a curved needle through a specific pair of entry and exit holes on a piece of sponge. There are multiple steps in the exercise including, Needle Handling, Needle Positioning, Needle Entry, Needle Curving, Needle Removal and Needle Exit. As it was pointed out in [6], these steps impose different visual, cognitive, and manual demands of the subject by involving many skills such as camera control, endowrist manipulation, and needle control and driving. Particularly, depending on the relative position and distance between the entry and exit holes on the sponge, there are 36 different exercises with various difficulty levels. Figure 1 shows four examples of the Suture Sponge exercises with different entry (yellow dots) and exit (black dots) positions relative to the sponge edge and each other. For example, comparing A with B, the distance between the entry and exit holes are different and the task to suture through shorter distance is easier than the one requires longer distance. Comparing A and B with C and D, it is noticed that they require different front-and-back hand operation, i.e., entry on top or exit on top. In general, the back-hand operation is much harder than the front-hand operation.

Data Collection: A wearable eye-tracking system, Pupil Core (Pupil Labs©, Berlin, Germany), was used to binocularly sample eye movements at 60 Hz. A camera located in the middle of the glass frame (outer side) records the scene while sensors mounted in the inner side of the glass frame capture two videos of the eyes. Recordings were then annotated using Pupil Player [18] from Pupil Labs©. TEPR (i.e., pupil diameter) was calculated from the raw data eye-tracking measurements.

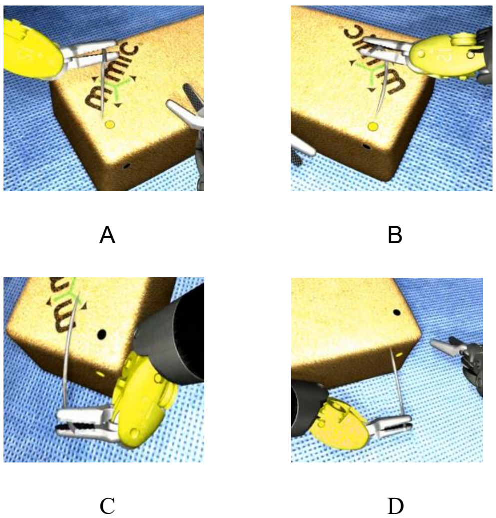

Figure 1: Examples of the Suture Sponge exercise

Exercise Difficulty Labeling: As aforementioned, the 36 exercises vary from each other in terms of the relative locations of the entry-exit holes on the sponge, distance traveled within the sponge and front-or-back hand manipulation required. The naming scheme of the suture exercises uses four letters, where each letter represents one corresponding attribute describing the exercise as shown in Fig. 2. The difficulty level of each exercise was rated by a panel of five expert surgeons on a scale of one to five. We then used the average of the five ratings as the final rating of each exercise as shown in Fig. 3 (labelled using the four-letter-labelling scheme in Fig. 2). We noticed that, even though the orders of the ratings were mostly consistent among the raters, the exact rating of each exercise was not the same across all raters. It was realized that this subjective rating scheme could introduce labelling noise into the training and testing and the proposed differential scale could reduce this noise by using the relative scales between exercises.

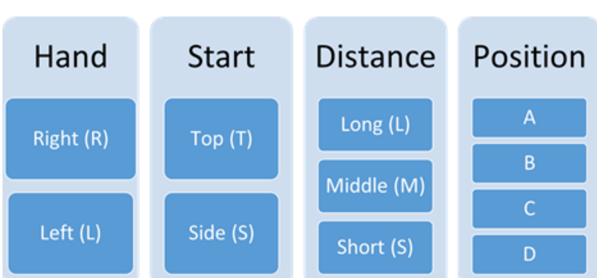

Figure 2: Naming scheme based on suture attributes

For each exercise, one of the suture exercises listed in Fig. 3 was selected randomly by the simulator (note that not all subjects performed the same number or same type of suture exercises). Overall, the two experts accomplished 80 exercises each. For the novices, the Untrained-1, Untrained-2, and Trained-1 accomplished 50, 90, and 120 exercises, respectively.

LSLA	3.6	LSSB	1.6	LTSB	1.2	RSSA	2.8
LSLB	3.8	LSSC	1.8	LTSC	1	RSSB	2.2
LSLC	4	LSSD	2.2	LTSD	1.2	RSSD	2.2
LSLD	4.4	LTLB	2.8	RSLA	4.6	RTLA	4
LSMA	2.6	LTLC	2.6	RSLB	3.6	RTLB	3
LSMB	2.8	LTLD	3	RSLD	4	RTMA	3.4
LSMC	3	LTMB	1.8	RSMA	3.8	RTMB	2
LSMD	3.4	LTMC	1.6	RSMB	3	RTSA	2.2
LSSA	1.6	LTMD	2	RSMD	3.2	RTSB	1.2

Figure 3: Difficulty ratings of the exercise

Experiments: Because the differences in the experience and skill levels between the experts and novices, it is realized that the difficulty levels rated by the experts in Fig. 3 are subjective and might not be able to fully reflect what the novices could experience in performing the same exercise. Therefore, to use these ratings directly for training machine learning models to predict the task difficulties experienced by different subjects is not practical. On the other hand, we observe that the difference between a hard and an easy exercise could be felt similarly by both the experts and novices. Based on this observation, we proposed a differential-difficulty scale, which is a score difference between any two exercises performed by the same subject. For example, the absolute scales for exercises LSLA and LSSA are 3.6 and 1.6, respectively, from Fig. 3. Thus, the differential scale is 2.0 when they are performed by the same subject. This differential-difficulty scale takes out the subjectivity in the absolute rating given by a few experts. It not only allows us to account for the skill and mental differences between subjects, but also significantly increases the number of training and testing samples through randomly pairing of exercises. Furthermore, because each subject could experience different cognitive challenges depending on their skill level and mental capacity, training a single model to predict the exercise difficulties felt by each subject using not only the subject's, but also other's cognitive response does not seem to make logical sense.

In this paper, we implemented a multitask learning approach where each task in the multitask framework is defined as to a regression task to predict the differential scale for each subject. Although each subject is different in their cognitive response to the exercises, nevertheless, they faced the same mental challenges from practicing on the same set of exercises. Therefore, we believe that the multiple tasks defined in our multitask regression model are related to each other and the relatedness could help with the prediction by leveraging the information across subjects. When subject-specific regression tasks are learnt simultaneously, the method not only effectively increases the sample size for each individual task (especially when a subject who completed fewer exercises), but also potentially improves the prediction performance as we will show in the next section. Several experiments have been performed using the differential-difficulty scale: 1) comparison of the multitask methods with the Support Vector Machine regression model (SVR) method; 2) across different number of features; 3) across different number of training samples.

IV. Results

In our experiment, 713 features were initially extracted using TSFresh from the time series of the pupil diameter, i.e., TEPR of each subject. The number of features were than reduced using [12] to different levels. Feature reduction was performed using data from all subjects. One challenge when comparing TEPR across subjects or subject under different situations is the baseline calibration as pupil dilation can change significantly from subject-to-subject or even the same subject under different illumination conditions. Because each suture exercise consists of multiple steps, i.e., needle positioning, entering, pulling and exiting, we used the mean diameter during needle positioning as our baseline since it is a common step across all subjects/exercises under the same illumination from the simulator and moreover, the stress level could be similar as the subjects are expected to be mentally getting ready for the exercise.

To implement the differential scale, we randomly paired two exercises for each subject by concatenating their features and the rating difference between the two exercises as the group truth differential scale. This pairing also allowed us to increase the sample size by a factor of 10. Because of the different experience levels across subjects, we avoided creating the differential scales across subjects.

We evaluated the performance based on two training sizes, 70% and 50%, and the results were averaged across five runs with random split of 70-30 and 50-50, respectively. In our study, we used the *R*2 from fitting a linear regression model between the actual difficulty levels and the predicted difficulty levels to evaluate our model accuracy of the predictions. The *R*2 results using TEPR with different number of features using the differential-difficulty scale are summarized in Tables I and II, where the columns correspond to different learning models, i.e., the single task SVR and the multitask methods, Trace-Norm (2), Lasso (3), and Robust (4), respectively. The rows correspond to the subjects. Table I shows that, even though the performances using the multitask methods are on par with the SVR for four out of the five subjects, they outperformed SVR for Subject, Trained-1.

TABLE I: R2 Results Using 112 TEPR Features

70%	SVR	Trace-Norm	Lasso	Robust
Expert-1	0.986	1.000	0.982	0.99
Expert-2	0.988	0.997	0.971	0.986
Untrain-1	0.992	1.000	0.996	1.000
Trained-1	0.693	0.836	0.787	0.824
Untrain-2	0.988	1.000	0.985	0.994

TABLE II: R2 Results Using 88 TEPR Features

70%	SVR	Trace-Norm	Lasso	Robust
Expert-1	0.983	0.990	0.93	0.967
Expert-2	0.964	0.977	0.901	0.955
Untrain-1	0.99	1.000	0.976	0.998
Trained-1	0.566	0.673	0.644	0.665
Untrain-2	0.909	0.950	0.918	0.944

The Lasso method performs worse than others (this might be due to the pre-feature selection we performed). From Table II, it is noted that the performances of Trained-1 and Untrained-2 dropped significantly with less features. This might be an indication that the best features for those two subjects might be different from the others. However, the multitask methods still outperformed SVR.

TABLE III: R2 Results Using 112 TEPR Features with 50% Training Data

50%	SVR	Trace-Norm	Lasso	Robust
Expert-1	0.965	0.965	0.955	0.954
Expert-2	0.982	0.964	0.935	0.969
Untrain-1	0.981	0.991	0.988	0.960
Trained-1	0.638	0.735	0.715	0.756
Untrain-2	0.983	0.989	0.956	0.985

Table III shows the results using only 50% of training data, the multitask methods still performed better than SVR for Trained-1 while the performances of other subjects were on par with SVR. Although Subject Trained-1 accomplished the largest number of exercises, using his/her TEPR, i.e., cognitive response to predict the task difficulties had the worst performance. The cause of this difference in prediction performance has not been well understood in this study and will need further investigation when we can have more subjects.

V. Discussion and Conclusion

In this paper, we proposed a multitask learning approach for exercise difficulty prediction based on TEPR in a simulated surgical setting. We proposed a differential scale that is shown to be more effective when evaluating cognitive-load-based performances across subjects with different experience and skill levels. Our experiments show that multitask learning can achieve high regression accuracy for challenging individual. Although the conventional wisdom believes that the cognitive stress might be experience-skill level dependent, it is noted that the cognitive stress can also depend on other hidden factors such as a person's mental capability, health-status, or the specific types of exercises practices, etc. Even though our proposed differential difficulty scale was based on the same types of exercises in our study, in reality, when the same type procedures are not available, the scale can also be calculated against a standard set of procedures in surgery or surgical trainings. As a next step, we would like to extend this single modality multitask learning framework to multimodal multitask learning to include other physiological measures such as GSV, HRV, and EEG. Another direction to explore is to extend the current multitask learning framework to multitask and multi-label in which we would also like to predict the skill level such as expert vs. novice. Furthermore, because the limitation of our current dataset, we were restricted to the traditional approaches. In the future, we would also like to expand the model to deep learning approaches.

Acknowledgments

We would like to thank Mr. William Sehnert for helping with the labeling of the data and the volunteers for participating in the experiments.

References

[1] M. A. Staal, "Stress, cognition, and human performance: A literature review and conceptual framework," in National Aeronautics & Space Administration, 2004.

[2] F. Chen et al., Robust Multimodal Cognitive Load Measurement. Cham: Springer International Publishing, 2016.

[3] J. Zhou, K. Yu, F. Chen, Y. Wang, and S. Z. Arshad, "Multimodal behavioral and physiological signals as indicators of cognitive load," in The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations - Volume 2, Association for Computing Machinery, 2018, pp. 287–329.

[4] R. D. Dias, M. C. Ngo-Howard, M. T. Boskovski, M. A. Zenati, and S. J. Yule, "Systematic review of measurement tools to assess surgeons' intraoperative cognitive workload," Br J Surg, vol. 105, no. 5, pp. 491–501, 2018, doi: 10.1002/bjs.10795.

[5] G. S. Avrunin et al., "Toward Improving Surgical Outcomes by Incorporating Cognitive Load Measurement into Process-Driven Guidance," Softw Eng Healthc Syst SEHS IEEE ACM Int Workshop, vol. 2018, pp. 2–9, May 2018, doi: 10.1145/3194696.3194705.

[6] C. Wu et al., "Eye-Tracking Metrics Predict Perceived Workload in Robotic Surgical Skills Training," Hum Factors, p. 18720819874544, Sep. 2019, doi: 10.1177/0018720819874544.

[7] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources," Psychological Bulletin, vol. 91, no. 2, pp. 276–292, 1982.

[8] M. Pedrotti et al., "Automatic Stress Classification with Pupil Diameter Analysis," International Journal of Human-Computer Interaction, vol. 30, no. 3, pp. 220–236, Jan. 2014, doi: 10.1080/10447318.2013.848320.

[9] A. Szulewski, N. Roth, and D. Howes, "The Use of Task-Evoked Pupillary Response as an Objective Measure of Cognitive Load in Novices and Trained Physicians: A New Tool for the Assessment of Expertise," Acad Med, vol. 90, no. 7, pp. 981–987, Jul. 2015, doi: 10.1097/ACM.0000000000000677.

[10] J. H. Nguyen et al., "Using objective robotic automated performance metrics and task-evoked pupillary response to distinguish surgeon expertise," World J Urol, Jul. 2019, doi: 10.1007/s00345-019-02881-w.

[11] M. Christ, A. W. Kempa-Liehr, and M. Feindt, "Distributed and parallel time series feature extraction for industrial big data applications," arXiv:1610.07717 (cs), May 2017, (Online). Available: http://arxiv.org/abs/1610.07717.

[12] M. Robnik-Šikonja and I. Kononenko, "An adaptation of Relief for attribute estimation in regression," in Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97), 1997, pp. p296-304.

[13] B. D. Fulcher and N. S. Jones, "Highly comparative feature-based time-series classification," IEEE Trans. Knowl. Data Eng., vol. 26, no. 12, pp. 3026–3037, Dec. 2014, doi: 10.1109/TKDE.2014.2316504.

[14] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11, San Diego, California, USA, 2011, p. 814, doi: 10.1145/2020408.2020549.

[15] J. Zhou, J. Chen, and J. Ye, "MALSAR: Multi-tAsk Learning via StructurAl Regularization," (Online). Available: http://www.MALSAR.org.

[16] J. Chen, J. Zhou, and J. Ye, "Integrating low-rank and group-sparse structures for robust multi-task learning," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11, San Diego, California, USA, 2011, p. 42, doi: 10.1145/2020408.2020423.

[17] Y. Zhang and D.-Y. Yeung, "A convex formulation for learning task relationships in multi-task learning," in Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, Jul. 2010, pp. 733–742, Accessed: May 19, 2020. (Online).

[18] "Pupil Player | Pupil Labs." https://docs.pupil-labs.com/core/software/pupil-player/ (accessed May 19, 2020).