1-1-2022

Visual Task Classification using Classic Machine Learning and CNNs

Devangi Vilas Chinchankarame
San Jose State University

Noha Elfiky
Saint Mary's College of California

Nada Attar
San Jose State University, nada.attar@sjsu.edu

Recommended Citation

Devangi Vilas Chinchankarame, Noha Elfiky, and Nada Attar. "Visual Task Classification using Classic Machine Learning and CNNs" *Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science* (2022). https://doi.org/10.11159/mhci22.110

Proceedings of the 8th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'22)
Prague, Czech Republic - July 28- 30, 2022
Paper No. MHCI 110
DOI: 10.11159/mhci22.110

Visual Task Classification using Classic Machine Learning and CNNs

Devangi Vilas Chinchankarame[1], Noha Elfiky[2], Nada Attar[1]
[1]San Jose State University
1 Washington Sq, San Jose, CA, USA
devangivilas.chinchankar@sjsu.edu; nada.attar@sjsu.edu
[2]Saint Mary's College of California
1928 St Marys Rd, Moraga, CA, USA
nme5@stmarys-ca.edu

Abstract **-** Our eyes actively perform tasks including, but not limited to, searching, comparing, and counting. This includes tasks in front of a computer, whether it be trivial activities like reading email, or video gaming, or more serious activities like drone management, or flight simulation. Understanding what type of visual task is being performed is important to develop intelligent user interfaces. In this work, we investigated standard machine and deep learning methods to identify the task type using eye-tracking data - including both raw numerical data and the visual representations of the user gaze scan paths and pupil size. To this end, we experimented with computer vision algorithms such as Convolutional Neural Networks (CNNs) and compared the results to classic machine learning algorithms. We found that Machine learning-based methods performed with high accuracy classifying tasks that involve minimal visual search, while CNNs techniques do better in situations where visual search task is included.

Keywords**:** Eye Tracking, Machine Learning, CNN, Vision, Visual search

1. Introduction

Despite humans having five vital sensory organs – eyes, ears, nose, tongue, and skin – and the brain receives and processes this information to activate response mechanisms, studies suggest that sight contributes to about 80% of all the sensory information that the brain processes [25]. One of the common tasks that our eyes perform thousands of times every day is visual search. It has been a major paradigm for studying visual attention [15]. Researchers have used various visual search paradigms to gain insight into attentional selection in the visual system. In a typical visual search task, subjects are asked to report if a visually distinctive target object is present among a set of distractors in each scene. If the target and distractors have similar visual characteristics, observers must sequentially attend to search items to find the target or determine its absence. Many models of visual search have been proposed, aimed at explaining the role of visual attention [5], [9], [12], [15].

Yarbus has provided extensive findings and presented qualitative data showing that eye movement patterns were affected by an observer's visual task. He suggested that complex mental states could be inferred from scan paths [1]. Several studies provide evidence that support Yarbus's claim that eye movement during visual search tasks can be used to predict user behavior [2], [11], [16], [20]. Greene et al. found that Yarbus's findings are questionable, because while it is possible for an observer's mental state to be decoded from some eye movement features, static scan paths alone are not sufficient to classify a visual search task or to infer other complex mental states of the users [13]. They computed seven measures from the eye movement scan paths and fed them into a linear discriminant classifier. They failed to find any support for Yarbus's claim in their study. The eye movement patterns can differ across tasks, but not across images. This could account for the striking difference between Yarbus's result and Greene's.

Several studies have investigated CNN models and other classification methods in visual tasks [3], [21], [23], [24], using fixation patterns to predict scene category [18], or gaze fixations during visual search tasks [22]. Hutt et al. classified tasks related to mind wandering using Bayesian networks [26], while the study by Faber et. al. used logistic regression [27]. Previous efforts on task classification from image representation have been made by Wang et al. [28] to represent data as images using Gramian Angular Fields (GAFs) and Markov Transition Fields (MTF).

The most recent study by Kumar et al. achieved 95.4% on task classification by eliminating highly correlated features before training an SVM and AdaBoost classifier to predict the tasks from filtered eye movements data [6]. However, in their method, they used five correlated variables from the eye fixations, pupil size, and a label to classify the data based on the task. We think this approach can be arbitrary as they only classified one task condition, where subjects were instructed to only look at a fixation mark at the center of the scene. Therefore, classifying every single row of the data with five variables into either of the categories might not predict the visual search tasks [16]. A small set of eye movement variables or a singular example in the scan path could indicate the cognitive level at some state of the task but not necessarily the scanning patterns for an observer examining particular stimuli.

In this research study, (i) we looked at identifying the visual task that a user performs by looking at the visual scan path. We also try to assess if the size of the pupil is a contributing factor in giving away the type of visual task. We tested our model on the free-viewing condition that Kumar et al. eliminated from their study, where subjects move their eyes freely in the stimuli during the visual search and exploration tasks. Our method also differed from previous studies in that (ii) we use RGB images of the scan path instead of extracting eye-tracking features. In our work, (iii) we used CNNs to analyze if incorporating pupil information - visually - can make an impact on the classification of the tasks. We aim to present a new way to classify tasks using state-of-the-art computer vision algorithms and compare it with classic Machine Learning (ML) models. Before testing the image representation, we reproduce results from [6] using Random Forests as a baseline. We used the same dataset that was used by Kumar et al. to run the Random Forest algorithm. Then, we ran various flavors of Convolutional Neural Networks (CNNs) to find the best setting. Finally, (iv) we experimented on Transfer Learning and Data Augmentation, which show significant improvements on 1-layered CNNs.

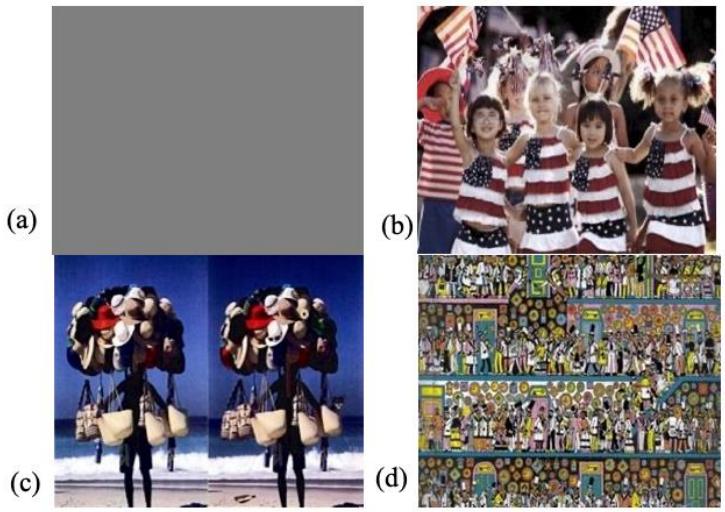

Fig. 1: Sample source for the 4 task types (a) Blank, (b) Natural, (c) Puzzle, and (d) Waldo.

2. Data Collection

We used the dataset in [6] that was collected from an extensive study by Otero-Millan et al. [19]. This dataset is compatible with our experimental design and meets the requirements of our goal of comparing classifiers on different visual tasks. The experiment included 2 conditions and 4 different tasks for a variation of visual tasks and eye movement data. The experimental design was described in [14], [19]. The experiment has fixation conditions and free-viewing conditions. In the fixation conditions, subjects had to fixate a red cross on the center of the screen. In the free-viewing conditions, subjects were free to move their eyes over the visual scene. The four tasks are: 1) Blank screen that only showed a grey screen, 2) Natural scene that has no target and subjects were instructed to explore the image, 3) a Where's Waldo task, where the subjects performed a visual search task to find Waldo, and 4) Picture Puzzle condition, where subjects were required to find

all the differences between two side-by-side, nearly identical images and indicate their locations at the end of the trial. Kumar et al. used the fixation conditions only in their analysis, where the subject's task did not vary as they only needed to look at the fixation mark at the center [6]. In the free-viewing conditions in our study, the subject's task varied according to the visual scene presented. Each task had 15 different visual scenes per the two conditions (except for the blank conditions). The total of the trials was 120 for each subject. The experiment was conducted over 3 sessions of 40 trials. Figure 1 shows a sample of each task.

3. Data Preparation and Experimental Setup

The dataset we included has the total of 8 subjects. The data includes the pupil diameter and x, y coordination of the fixation positions. The data has the variables LP (left pupil), RP (right pupil), LX href (eye velocity for left eye on the x-coordinate), LY href (eye velocity for left eye on the y-coordinate), LX Pix (pixel location for left eye on the x-coordinate) and LY Pix (pixel location for left eye on the y-coordinate). Each stimulus has the original image of 921x630 pixels.

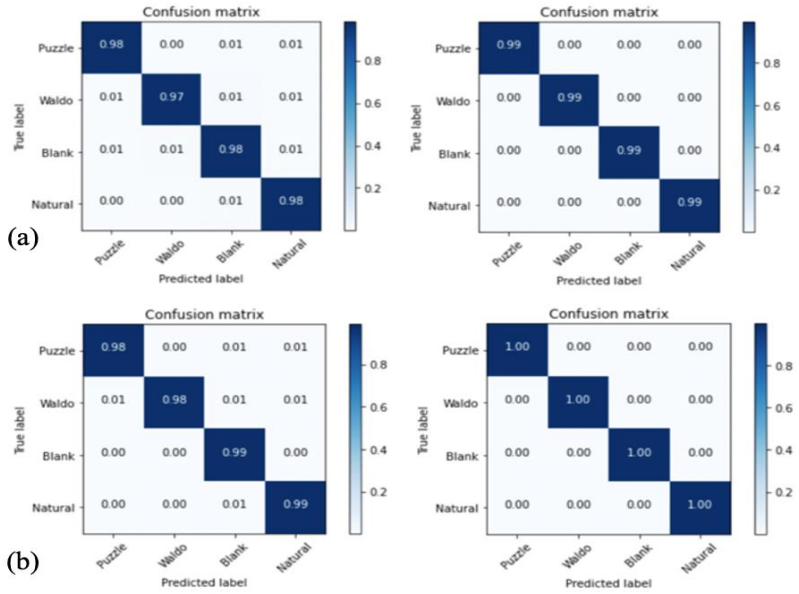

Fig. 2: (a) shows the confusion matrices for free-viewing condition using Random Forests when using (left) a subset of features and (right) all features. The bottom row shows the confusion matrices for fixation condition using Random Forests when using (left) subset of features and (right) all features.

3.1. Working on Raw Data

Kumar et. al.'s work had promising results for the classification of the fixation condition on raw data. Each data point in the observation file for a user denotes a timestep of the observation. The authors considered each datapoint as a separate sample for the training process and conducted experiments considering only the features of the left eye. We designed our experiments on random forests in a similar approach, but by considering features for both left and right eyes. In addition, we also tried to run the algorithms on free-viewing dataset.

The dataset has 10 the features ($LXpix, LYpix, RXpix, RYpix, LXhref, LYhref, RXhref, RYhref, LP, RP$). In this work, we chose a particular subset of features which includes the gaze fixation points (LXpix, LYpix, RXpix, RYpix) and the pupil information of the left and right eyes (LP, RP) since the same subset is used to translate the data into image representations in the latter experiments and serves as a common ground for comparison.

3.1. Machine Learning Experiment: Random Forest

One of the most important requirements for Machine Learning is data quality. Missing data, unnormalized values, outliers, etc. can significantly affect model performance. Our raw data contained few rows with missing values that corresponded to the data points where the user looked outside of the viewing area. Eliminating such rows (observations at a time instant) can improve overall performance. Therefore, if the eye coordinate position calculated was beyond the edge of the image coordinates (621 by 930), the data will be ignored assuming that the user was looking outside the image scene.

Before we ran Random Forest algorithm, we first replicated Kumar et al. [6] using Adaboost model on the same features to see what accuracy we can obtain on free-viewing dataset that Kumar et al. did not include in their study. We followed Kumar et al. in labeling each task as (0-waldo, 1-puzzle, 2-blank, 3-natural) for the 4 tasks in both conditions. The model used each timestamp's five variables (LXpix, LYpix, LXhref, LYhref, LP) and a label as a single exemplar. The accuracy result for the fixation condition corroborates Kumar et al. results of 95%, while the accuracy for classifying the four tasks for the free-viewing condition is 84%.

We designed our experiment on random forest in a similar approach, but by considering features for both left and right eyes. In addition, we also ran the algorithms on the fixation and free-viewing datasets. Table 1 illustrates the accuracies for both conditions using random forests. The results in Figure 2 (a) and (b) show the confusion matrices using the random forest for both conditions, using 10 trees set and the maximum depth is set until all the leaves are pure.

Table 1: Summary of accuracy results using random forests

Random Forest	Subset of Features	All Features
Free-viewing	98%	**99%**
Fixation	99%	**100%**

The experiment using Random Forests seems to work well with high accuracy under both viewing conditions when all the features are considered. However, each timestep of the viewing observation of a user is considered as a separate training sample. Those timestamps are just a part of the whole observation and represent a single gaze point. Data from a single timestamp can be useful to understand cognitive load at a particular task state, but not for the entire scanning pattern across the entire task period. The observation should ideally make up a single training sample to reflect the correct cognitive state. Alternatively, the AdaBoost model in Kumar et al. [6] was successful in classifying tasks in the fixation condition but was not efficient in classifying tasks that require a visual search. The filtered eye movement data used in Kumar et al.'s study has pupil size as one of the features. When a user only fixates at the center of the screen, the pupil size could be affected by the overall brightness or the luminance of the stimulus more than visual search or cognitive load level [15]. Therefore, the classification using the eye movement data will depend mostly on the pupil dilation, given the fact that the x and y positions of the fixations will not vary across the tasks in the fixation condition. Another possibility is that the users could try to do a minimal search during the Puzzle or Waldo tasks in the fixation condition, while maintaining to look at the center. That creates a small scan path around the center of the screen [4], [17].

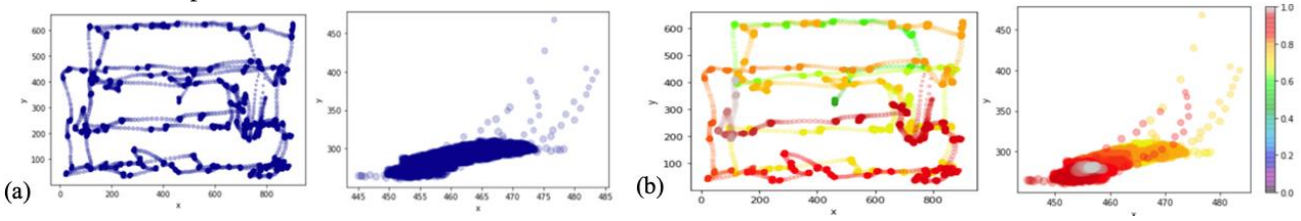

Fig. 3: (a) Scan path of "Waldo" image in (left) Free-viewing and (right) Fixation conditions where darker points show points looked at repeatedly. (b) Scan path of "Waldo" image with colors reflecting pupil dilation in (left) Free-viewing and (right) Fixation conditions.

4. Paper Format CNNs and Image Representation of Scan Path

In order to prepare the eye-movement data to CNNs models, we took a new approach by generating images from the pupil size and the scan path of the eye fixations of both the left and the right eyes. We extracted the position of the eye based on LXpix, LYpix, RXpix, and RYpix. We then calculated the eye position as of the X-coordinate = ceil (LXpix + RXpix)/2. Similarly, for the Y-coordinate = ceil (LYpix + RYpix)/2. This implies that at each timestamp, the user's eye was looking at a specific pixel that has two values. Then, these two values will identify the row and column indices in the matrix that represent a point of the scan path.

Since not all gaze points are distinct from each other, some amount of overlap is inevitable. We adjusted the opacity of the pixels by adding the pupil size information as shown in Figure 3 (a) to find which points are looked at more. The darker colored pixels are points of interest for the user than the ones which are lighter. Studies have shown that pupil size can be an indicator of cognitive load during visual tasks [15], [16]. To use pupil size in task classification, we plotted the pupil value on the scan path indicative by a color range. In addition to the consideration of opacity or "interest" in Figure 3 (b), we adopted a color range to map the range of pupil values found across the dataset. This means the lowest across all and the highest across all determined the range of the pupil values, and hence color maps across all image representations can be emitting information on a uniform level. Thus, we can consider opacity to denote "interest" and color to denote "pupil dilation". Finally, we converted all the matrices into images. We generated the scan paths from the raw data on every user for fixation as well as free-viewing conditions on the 4 types of tasks defined above. The dataset consists of 120 images for each task and hence, 480 images are the total across all subjects. We chose contrasting values for the colors used to distinguish different pupil values, which is important [29]. We ran shallow CNNs on paths represented by Figure 3 (a) shades of blue and contrasting red and green shades as represented in Figure 3 (b). The first approach yielded accuracies of 62%, while the second approach yielded accuracies of 66% for classification on 4 classes.

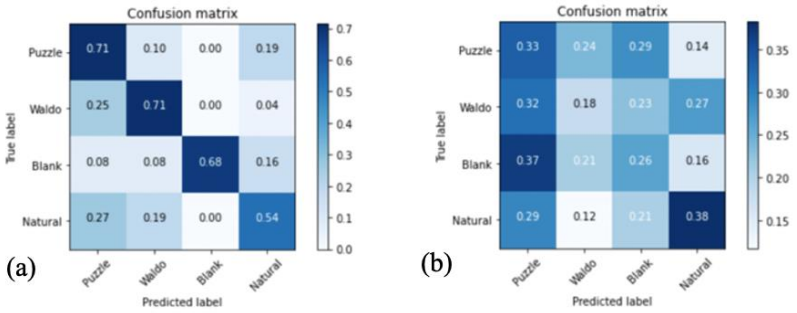

Fig. 4: Confusion Matrices (4 classes) using 1 layered CNN in (a) Free-viewing and (b) Fixation conditions.

4.1. Shallow CNNs

To establish a baseline, we first ran 1-layered CNNs with 16 (3*3) filters on images of size 64*64. The classification accuracy was 66% for the free-viewing condition, while it was 30% for the Fixation condition. The confusion matrices in Figure 4 show that the classification for the "Natural" scene is the least accurate for the free-viewing condition. The Natural scene has the lowest accuracy due to the similarity in the viewing patterns in Blank and Natural tasks. Another consideration is that the task of exploration depends significantly on the user's cognitive state. The process of exploration can vary from user to user. On the contrary, searching tasks of either finding an object (Waldo) or the difference (Puzzle) demand specific information from the user. Visual exploration in the natural task does not ask the user to perform a specific activity, hence classifying these tasks can be challenging due to the lack of evident and uniform patterns.

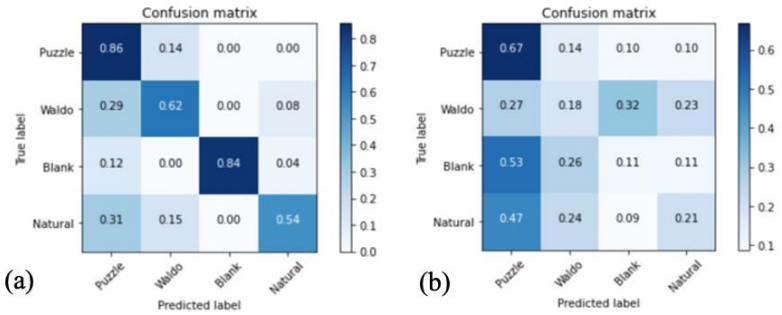

Fig. 5: Confusion Matrices (4 classes) - Deep CNNs on (a) Free-viewing and (b) Fixation conditions.

4.2. Deep CNNs

We conducted experiments on Deeper CNNs on different image sizes from 64*64, 128*128, and 256*25 to see how the performance of the CNN was affected as the image size changed. Our model had a depth of 3 and we tested the number of filters to be 16/32, and the filter size to be of the popular size of 3*3 units. We split our data 80-20 for the training and testing processes, and further split the data into 80-20 to be the actual training samples and the validation samples. We employed 5-fold cross-validation and averaged the results over the 5 iterations. Tables 2 (a) summarizes results by using CNNs on different image sizes, and Figure 5 shows the confusion matrices. We experimented with image sizes of 256*256 and the results did not add any improvement. Our results indicated that higher image size can be helpful but going further is not beneficial. CNNs can suffer from smaller datasets and perform their best when presented with huge training samples. To solve this problem, we employed next the techniques of transfer learning and data augmentation.

5. Transfer Learning Experiment

Transfer Learning is the process by which the "learning" achieved in one problem is used to solve another problem [30], [31]. It is a useful technique when the data available is small for the problem at hand. In this study, we explored the usage of the MobileNetV2 pre-trained model and added a fully connected layer ahead. MobileNetV2 uses depth-wise separable convolutions as building blocks [32]. Transfer learning had an improved accuracy compared to CNNs, even with an image size of 256*256. The results are shown in row (b) of Table 2 and the confusion matrix of Figure 6.

Table 1: Summary of accuracy results using random forests

Classification Model	Condition	64 x 64	128 x 128	256 x 256
(a) DEEP CNN	Free-viewing	60%	**70%**	54%
	Fixation	26%	**28%**	25%
(b) CNN+ Transfer Learning	Free-viewing	65%	**81%**	80%
	Fixation	**36%**	29%	30%
(c) CNN + Data Aug.	Free-viewing	**86%**	25%	
	Fixation	**43%**	23%	

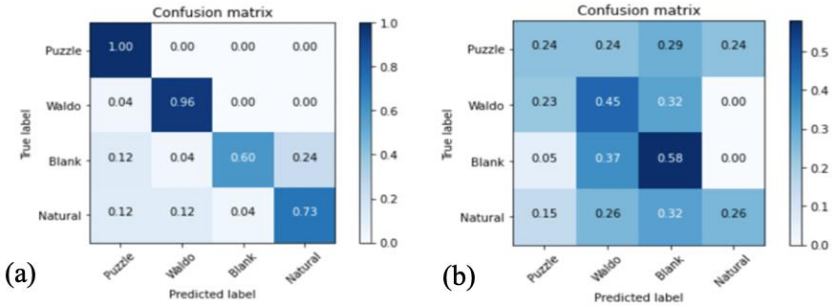

Fig. 6: Confusion Matrices (4 classes): Transfer Learning on (a) Free-viewing and (b) Fixation conditions.

6. Data Augmentation Experiment

Machine learning tasks in computer vision work best with huge amounts of data [33]. However, the data is not always available in such amounts. Hence, tasks call for the need to synthetically produce data. Most dataset that includes eye-movement is usually not sufficient for deep learning models. In our study, we had a dataset of 480 images for both Fixation and Free-viewing conditions. To augment it 3-fold, for each image, we generated 2 new images by randomly shifting each pixel upward, downward, leftward, rightward. This leads to 2 synthetic images from every image and increasing our dataset to 1440 images. This approach of pixel-wise shifting is highly positional and makes drastic changes to the original images to generate new ones and is not the best way of creating artificial data. Alternatively, we simulated different eye positions and generated new images accordingly. Figure 7 (b) shows a sample of how a synthetically produced scan path image looks from its original scan path in (a). Data Augmentation helped the Fixation condition. Transfer Learning could only improve so much, but with this approach accuracy increased from 28% to 43%. Summary of the results are shown in Table 2 (c) and confusion matrices in Figure 8.

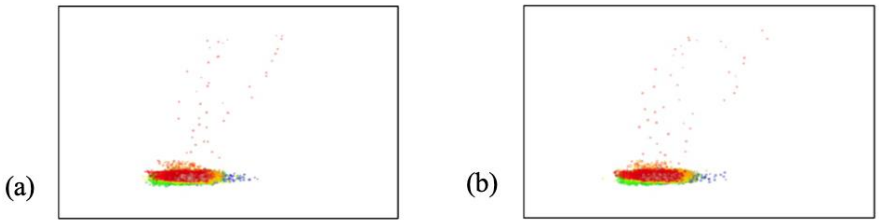

Fig. 7: (a) Original Scan path of a random user looking at a Puzzle image for Fixation condition, (b) Synthetically produced scan path for a Puzzle image for fixation condition.

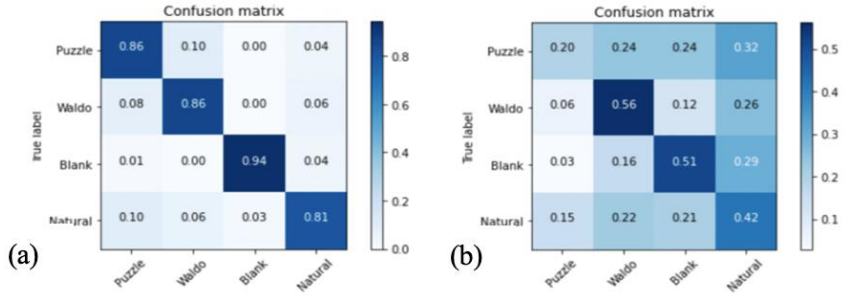

Fig. 8: Confusion Matrices (4 classes): CNN on Augmented Data on (a) Free-viewing and (b) Fixation conditions.

7. Discussion

The AdaBoost model in Kumar et al. successfully classified tasks in the fixation condition but was not efficient in classifying visual search tasks. When a user fixates at the center of the screen, the pupil size features could be affected by the overall brightness or the luminance of the stimulus more than visual search or cognitive load level [15]. Therefore, the classification using the eye movement data will depend mostly on the pupil size, given the fact that the x and y positions of the fixations will not vary across the tasks in the fixation condition. Another possibility is that the users could try to do a minimal search during the Puzzle or Waldo tasks in the fixation condition, while maintaining to look at the center. That creates a small scan path around the center of the screen [4], [17].

Alternatively, adding pupil size to machine learning algorithms improves the classification accuracy. In our experiment using Random Forests seems to work well with high accuracy under both viewing conditions when all the features are considered. However, each timestep of the viewing observation of a user is considered as a separate training sample. Those timestamps are just a part of the whole observation and represent a single gaze point. Data from a single timestamp can be useful to understand cognitive load at a particular task state, but not for the entire scanning pattern across the entire task period. The observation should ideally make up a single training sample to reflect the correct cognitive state.

8. Conclusions and Future Work

We viewed the problem of task classification from eye-tracking information using machine learning and CNN. We first explored shallow CNNs, as a baseline. Then, we examined a Deeper CNNs that improved the accuracy as the added layers were able to extract more useful information. We further implemented the technique of Transfer Learning as we observed significant improvements in accuracy for the free-viewing condition. Another solution that we employed was data augmentation; by synthetically reproducing data, we achieved improved accuracies (Free-viewing: 86%, Fixation: %43), especially for the fixation condition.

Vision tasks are greatly solved with CNNs. However, our study shows that it is not always true due to the small dataset size, which is common when dealing with eye movement data with the limited number of trials a user performs. In this case, we found that machine learning is the best approach when dealing with the pupil size. Furthermore, we compared the performance gain from CNNs in both conditions using visual data. We found that CNNs techniques do better in situations where visual search task is included. Free-viewing conditions allow the user for active participation. Whereas the CNNs struggle when visual search is minimal and when the visual representation fails to deliver all the information, which we have seen in the Fixation condition. We focused on the vision aspects of this problem as part of this study, but other studies can focus on leveraging sequential information and employing algorithms like HMMs, LSTMs and ML ensemble methods.

The ability to understand the task type of a visual search experiment can open many doors in the field of Human-Computer Interaction. User service can be customized if we can know where the user is looking at and what interests them. In the world of virtual and augmented reality, tracking what the user sees and tailoring experiences based on it can significantly improve user satisfaction. Studies can be carried out on not only identifying the task type from the observations, but also on identifying user attributes like age from the viewing patterns. The domains of user behaviour and user psychology combined with computational techniques open the paths to many possible research ideas.

References

[1] Alfred L Yarbus. 2013. Eye movements and vision. Springer.

[2] Ali Borji, Laurent Itti. 2014. Defending Yarbus: eye movements reveal observers' task. *Journal of Vision*, 14(3):29.

[3] Ali Borji, Laurent Itti. 2013. State-of-the-art in modeling visual attention. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 35, 185–207

[4] Brandt S. A. Stark L. W. 1997. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, 9, 27–38.

[5] Anne M. Treisman. 1980. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136.

[6] Ayush Kumar, Anjul Tyagi, Michael Burch, Daniel Weiskopf, and Klaus Mueller. 2019. Task classification model for visual fixation, exploration, and search. *In Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications (ETRA '19)*. Association for Computing Machinery, New York, NY, USA, Article 65, 1–4.
[7] Daniel Sonntag. 2015 "Kognit: Intelligent Cognitive Enhancement Technology by Cognitive Models and Mixed Reality for Dementia Patients." AAAI Fall Symposia
[8] Hosnieh Sattar, Sabine Müller, Mario Fritz, Andreas Bulling, "Prediction of search targets from fixations in open-world settings," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015 pp. 981-990.
[9] Jeremy M. Wolfe. 1994. Visual Search in Continuous, Naturalistic Stimuli." *Vision Research*, 34(9):1187–1195,
[10] Julian Steil, Philipp Müller, Yusuke Sugano, Andreas Bulling. 2018. "Forecasting User Attention During Everyday Mobile Interactions Using Device-Integrated and Wearable Sensors". *Proc. ACM International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*, pp. 1:1–1:13
[11] John M. Henderson, * Svetlana V. Shinkareva, Jing Wang, Steven G. Luke, and Jenn Olejarczyk (2013) Predicting Cognitive State from Eye Movements. PLOS ONE 8(5): e64937.
[12] Kyle R. Cave and Jeremy M. Wolfe. "Modeling the role of parallel processing in visual search." Cognitive Psychology, 22:225–271, 1990.
[13] Michelle R. Greene, Tommy Liu, Jeremy M Wolfe. Reconsidering Yarbus: a failure to predict observers' task from eye movement patterns. Vision Res. 2012 Jun 1;62:1-8. doi: 10.1016/j.visres.2012.03.019. Epub 2012 Apr 2. PMID: 22487718; PMCID: PMC3526937.
[14] Michael B. McCamy, Jorge Otero-Millan, Leandro Luigi Di Stasi, Stephen L. Macknik and Susana Martinez-Conde. Highly informative natural scene regions increase microsaccade production during visual scanning. J Neurosci. 2014 Feb 19;34(8):2956-66. doi: 10.1523/JNEUROSCI.4448-13.2014. PMID: 24553936; PMCID: PMC6608512.
[15] Nada Attar, Matthew H. Schneps, and Marc Pomplun. "Working memory load predicts visual search efficiency: Evidence from a novel pupillary response paradigm." *Memory & Cognition*, 44(7):1038–1049, 2016.
[16] Nada Attar, Paul Fomenky, Wei Ding, and Marc Pomplun. "Improving Cognitive Load Level Measurement through Preprocessing of Psychophysical Data by Random Subspace Time-Series Method." *In The IEEE International Conference in Human Computer Interaction (ICHCI 2016)*, 2016.
[17] Nada Attar, Matthew H. Schneps, and Marc Pomplun Pupil size as a measure of working memory load during a complex visual search task. *Journal of Vision* 13 (9), 160-160 (2), 2013
[18] O'Connell T. Walther D. (2012). Fixation patterns predict scene category. *Journal of Vision*, 12 (9): 801
[19] Otero-Millan J, Troncoso XG, Macknik SL, Serrano-Pedraza I, Martinez-Conde S. Saccades and microsaccades during visual fixation, exploration, and search: foundations for a common saccadic generator. J Vis. 2008 Dec 18;8(14):21.1-18. doi: 10.1167/8.14.21. PMID: 19146322.
[20] Shamsi T. Iqbal, Xianjun Sam Zheng, and Brian P. Bailey. 2004. Task-evoked pupillary response to mental workload in human-computer interaction. *In CHI '04 Extended Abstracts on Human Factors in Computing Systems (CHI EA '04*). Association for Computing Machinery, New York, NY, USA, 1477–1480.
[21] Yao Zhou, Jiamin Ren, Jingyu Li, Litong Feng, Shi Qiu, and Ping Luo. 2017. Video Classification via Relational Feature Encoding Networks. *In Proceedings of the Workshop on Large-Scale Video Classification Challenge (LSVC '17)*. Association for Computing Machinery, New York, NY, USA, 9–13.
[22] Zelinsky G. Peng Y. Samaras D. (2013). Eye can read your mind: Decoding gaze fixations to reveal categorical search targets. *Journal of Vision*, 13 (14): 10, 1–13,
[23] Zhang L. Tong M. H. Marks T. K. Shan H. Cottrell G. W. (2008). Sun: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8 (7): 32, 1–20,
[24] Zhao Q. Koch C. (2012). Learning visual saliency by combining feature maps in a nonlinear manner using adaboost. *Journal of Vision*, 12 (6): 22, 1–15
[25] D. Ripley, T. Politzer. 2010. "Vision Disturbance after TBI", NeuroRehabilitation. vol. 27 pp215–216 doi 10.3233/NRE-2010-0599

[26] S. Hutt, J. Hardey, R. Bixler, A. Stewart, E. Risko, and S. D'Mello.. "Gaze-Based Detection of Mind Wandering During Lecture Viewing". *International Educational Data Mining Society.* 2017
[27] M. Faber, R. Bixler, and S. D'Mello. (2018). "An automated behavioral measure of mind wandering during computerized reading" Behav. Res. Methods 50, 134–150.
[28] Z. Wang, T. Oates. "Encoding time series as images for visual inspection and classification using tiled convolutional neural networks", *in Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015
[29] A. Gomez-Villa, A. Martín, J. Vazquez-Corral, M. Bertalmío, J. Malo, "Color illusions also deceive CNNs for low-level vision tasks: Analysis and Implications", *Vision Research Oxford*, 2020-11, Vol.176, p.156-174
[30] Michael B. McCamy, Jorge Otero-Millan, Leandro Luigi Di Stasi, Stephen L. Macknik and Susana Martinez-Conde "Highly informative natural scene regions increase microsaccade production during visual scanning", Journal of Neuroscience. 2014 Feb 19;34(8):2956-66. PMCID: PMC6608512.
[31] Xuhong Li, Yves Grandvalet, Franck Davoine, Jingchun Cheng, Yin Cui, Hang Zhang, Serge Belongie, Yi-Hsuan Tsai, and Ming-Hsuan Yang. "Transfer learning in computer vision tasks: Remember where you come from", *Image and vision computing*, 2020-01, Vol.93 (103853), p. 103853
[32] https://ai.googleblog.com/2018/04/mobilenetv2-next-generation-ofon.html
[33] V. Kothari et al., "Automated image classification for heritage photographs using Transfer Learning of Computer Vision in Artificial Intelligence" *Turkish journal of computer and mathematics education*, 2021-10, Vol. 12 (11), p.1940-1953