Fall 12-17-2021

Task Classification During Visual Search Using Classic Machine Learning and Deep Learning

Devangi Vilas Chinchankar

Task Classification During Visual Search Using Classic Machine Learning and Deep Learning

A Project

Presented to

Department of Computer Science

San José State University

In Partial Fulfillment

Of the Requirements of the degree

Master of Science

By

Devangi Vilas Chinchankar

December 2021

The Designated Project Committee Approves the Project Titled

“Task Classification during Visual Search using Classic Machine Learning and Deep Learning”

By

Devangi Vilas Chinchankar

FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSE STATE UNIVERSITY

December 2021

Dr. Nada Attar, San Jose State University

Dr. Mark Stamp, San Jose State University

Dr. Noha Elfiky, Saint Mary’s College of California

ABSTRACT

In an average human life, the eyes not only passively scan visual scenes, but most times end up actively performing tasks including, but not limited to, searching, comparing, and counting. As a result of the advances in technology, we are observing a boost in the average screen time. Humans are now looking at an increasing number of screens and in turn images and videos. Understanding what scene a user is looking at and what type of visual task is being performed can be useful in developing intelligent user interfaces, and in virtual reality and augmented reality devices. In this research, we run machine learning and deep learning algorithms to identify the task type from eye-tracking data. In addition to looking at raw numerical data, we take a "visual" approach by experimenting on variations of Computer Vision algorithms like Convolutional Neural Networks on the visual representations of the user gaze scan paths. We compare the results of our visual approach to the classic algorithm of random forests.

***Keywords* - Visual Search, Visual Attention, Eye Tracking, Machine Learning, Random Forests, Deep Learning, Convolutional Neural Networks, Computer Vision**

ACKNOWLEDGMENTS

I would like to sincerely thank my project advisor, Dr. Nada Attar, for understanding, guiding, and pushing me in times of doubt. Her inputs and pieces of advice not only proved to be valuable for this project but will always be a guiding force as I go forward.

I am grateful to my committee members, Dr. Mark Stamp and Dr. Noha Elfiky, for showing their faith in me, and for their indispensable time and support.

Finally, I would like to thank my friends, family, and professors for their continuous encouragement.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

Living forms perceive their surroundings primarily with the help of the five vital sensory organs – eyes, ears, nose, tongue, and the skin. The brain then receives and processes this information to determine and activate the response mechanisms. Studies suggest that sight contributes to about 80% of all the sensory information that the brain processes [1].

Eyesight is a physical phenomenon, while vision deals with how the scenes are processed by the brain. Visual exploration and search are routine tasks that humans or non-human living forms with higher cognitive abilities perform in their day-to-day activities. They involve active scanning of the surrounding environment to observe and/or look for objects of interest. A typical visual search experiment asks the participants to search for a distinctive object in an image where the degree of visual attention required is directly affected by the degree of distraction in the provided image.

Understanding where the user is looking at and what interests the user visually can tell a lot about the user as well as the possible surroundings. The inspection behavior and therefore the viewing pattern is significantly impacted by the viewing instruction that is being given [2]. Multiple studies have shown the role of visual attention in visual search and have proposed the possibility to determine the type of visual search task being performed from the experiment data points.

In this research, we plan on identifying the visual task that a user performs by looking at the scan paths by considering it as a computer vision problem. We also try to assess if the size of the pupil is a contributing factor in giving away the type of visual task.

II. Related Work

Yarbus in [2] has provided extensive findings and qualitative data pointing towards the existence of patterns in visual search tasks. Yarbus's observation on how the distribution of fixation gaze points differs with change in the kind of information requested by the task is invaluable. Multiple studies have focused on understanding tasks using pattern analysis from visual search data. Previous works on classifying the tasks have involved using various statistical and learning methods on raw data. Hutt et al. [4] classified tasks related to mind wandering using Bayesian networks while the study by Faber et. al. [5] used logistic regression.

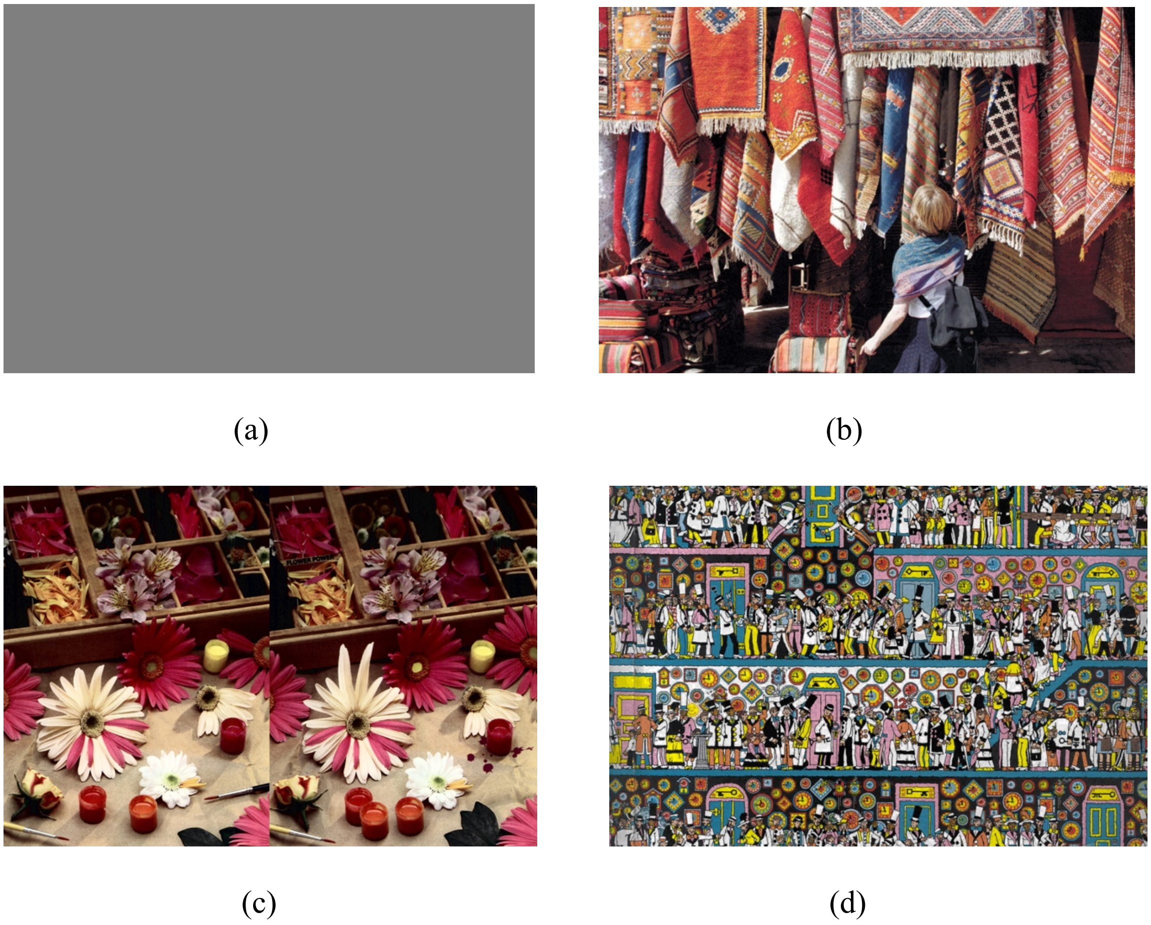
Recent studies of Kumar et al. [3] achieved a classification accuracy of about 95.4% on four types of visual search tasks. Their work performed studies on fixation tasks, where the participant is asked to fixate their gaze in the center of the image, no matter the type of image. Their work was extended through the study by Thentu [8] in classifying free viewing tasks where participants can freely observe/search the image unlike in fixation tasks.

Previous efforts on task classification from image representation have been made by Wang et al. [6] to represent data as images using Gramian Angular Fields (GAFs) and Markov Transition Fields (MTF). Thentu [8] represented the scan paths as RGB images and explored computer vision algorithms for the classification task. Efforts in [8] focus solely on the scan path of the eye. An important feature that would otherwise present itself in raw data is missing - pupil dilation. Could a user's pupil response (pupil dilation) to a given task potentially vary with different tasks [14, 15]? Given the varying level of attention different tasks need, could pupil dilation give away the type of task? In this report, we try to use Deep Learning 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 these tasks using classic machine learning models as well as state-of-the-art computer vision algorithms.

Before testing the image representation, we reproduce results from [3] using Random Forests as a baseline. Then we run various flavors of Convolutional Neural Networks (CNNs) to find the best setting. Experiments on Transfer Learning and Data Augmentation show significants improvements on 1-layered CNNs.

III. Dataset

We use the same dataset which was used by Kumar et al. [3] from Otero-Millan et al's study [7]. We have a total of 480 observations from 8 subjects where each trial lasts for about 45 seconds and results in 22000 data points. The type of target stimuli images used are a blank scene, natural scene (observe), picture puzzle (find the difference), and "Where's Waldo" (find an object); samples of which are shown in Figure 1. The stimuli images have a resolution of 921 x 630 pixels.



(a) (b) (c) (d)

Figure 1 Sample source for the 4 task types (a) Blank, (b) Natural, (c) Puzzle, (d) Waldo

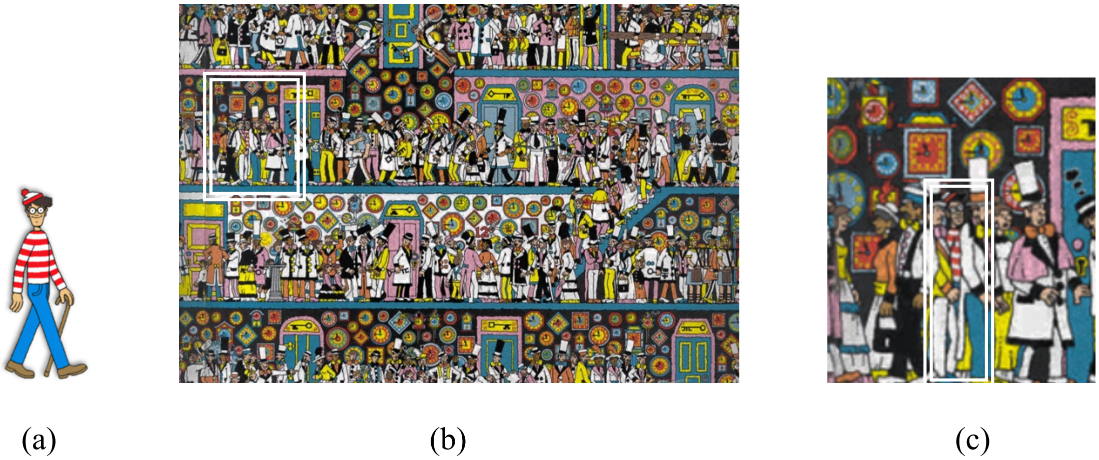


Figure 2 "Where's Waldo?"; (a) the character 'Waldo', (b) location of Waldo in the image, (c) enlarged part of the location

A. Viewing Conditions

As part of the experiments conducted to collect this data, users were asked to view these tasks in two viewing conditions – Free-viewing and Fixation. In the Free-viewing condition, users were asked to freely look at the image in the "Natural" and "Blank" type tasks, actively find the difference in the "Puzzle" tasks, and locate an object (find Waldo) in the "Waldo" images. The task of viewing "Natural" scenes comes under the visual exploration category while viewing the "Puzzle" and "Waldo" scenes comes under the visual search category. In contrast to this method of observing patterns, users were asked to fixate at the center of the target image in the Fixation condition irrespective of the type of image,

B. Data Distribution

Experiments were conducted by running the images through 8 subjects. Each subject was asked to look at 60 different images in both Free-viewing and Fixation conditions yielding a total of 480 observations for both Free-viewing and Fixation. Each subject looked at 15 images each of type Puzzle, Waldo, Blank, and Natural.

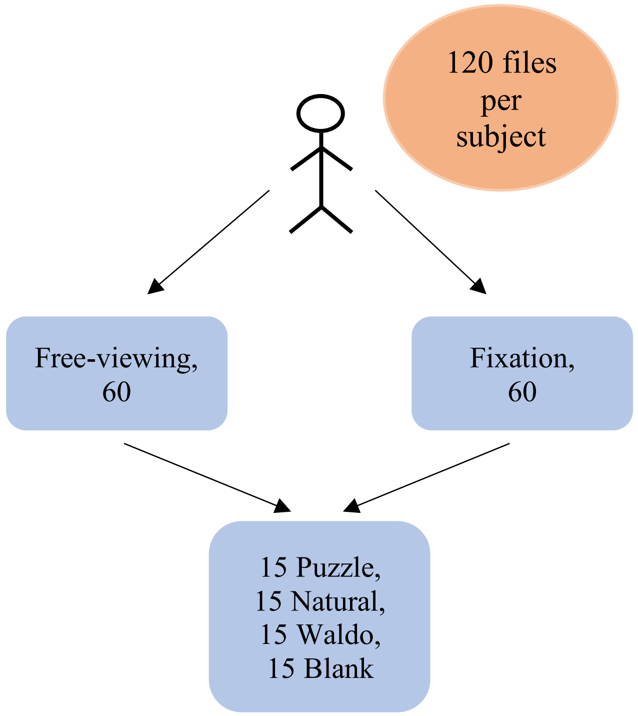


Figure 3 Input data files generated by a single subject

1. Data Pre-processing

One of the most important tasks of Machine Learning implementations is the quality of data that is being used. Missing data, unnormalized values, outliers, etc. can significantly affect model performance. We found that our raw data in the form of

CSVs contained a few rows with missing values. The missing values corresponded to the data points where the user looked outside of the viewing area. Using the data as it is affected model performance and gave a very poor accuracy; no better than guesswork. Eliminating such rows (observations at a time instant) drastically improved overall performance.

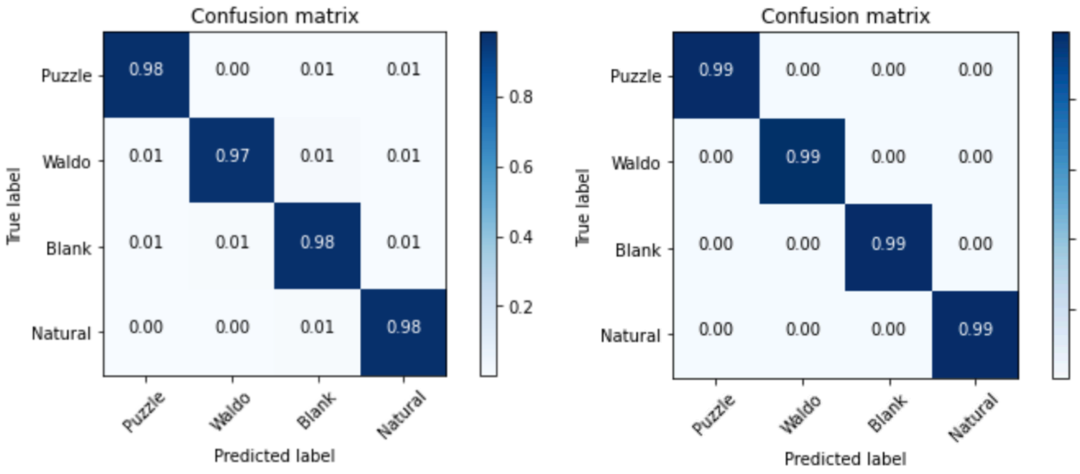
IV. Working on Raw Data

Kumar et. al's work in [3] showed 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 such 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 data.

We chose this particular subset of features which includes the gaze fixation points (LXpix, LYpix, RXpix, RYpix) and the pupil information (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.

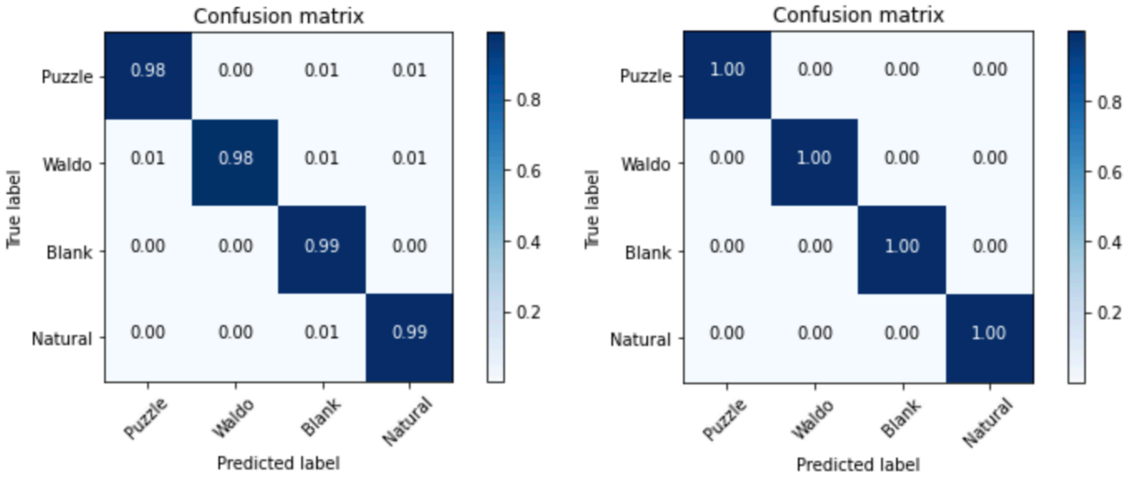
All features = { LXpix, LYpix, RXpix, RYpix, LXhref, LYhref, RXhref, RYhref, LP, RP }
Subset = { LXpix, LYpix, RXpix, RYpix, LP, RP }

The results in Figure 4 and Frigure 5 are of the random forest experiments with the number of trees set to 10 and the maximum depth to be until all the leaves are pure.



(a) (b)

Figure 4 Confusion matrixes for 'Free Viewing' condition using Random Forests when using (a) a subset of features (b) all features



(b) (b)

Figure 5 Confusion matrixes for 'Fixation' condition using Random Forests when using (a) a subset of features (b) all features

Table 1 shows the accuracies for both Free-viewing and Fixation conditions using Random Forests.

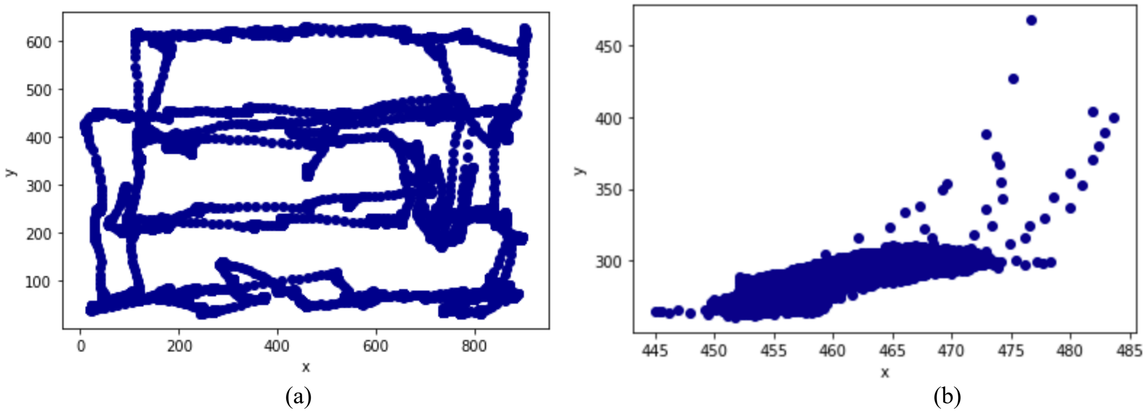
Table 1 Summary of Results using Random Forests

Random Forest	Subset of Features	All Features
Free Viewing	98%	99%
Fixation	99%	100%

Similar experiments were run using Random Forests seem to work fairly impressively with an almost perfect accuracy for both viewing conditions when all the features are considered. The issue we felt with this approach, and as discussed in [9], is that each timestep of the viewing observation of a user is considered as a separate training sample. In reality, each time step is just a part of the whole observation and represents a single gaze point. Data from a single timestep can possibly be useful to understand cognitive load at a particular task state but not automatically be useful for the entire scanning pattern across the entire task time period. The observation as a whole should ideally make up a single training sample to reflect the correct cognhitive state.

V. IMAGE REPRESENTATIONS OF SCANPATH

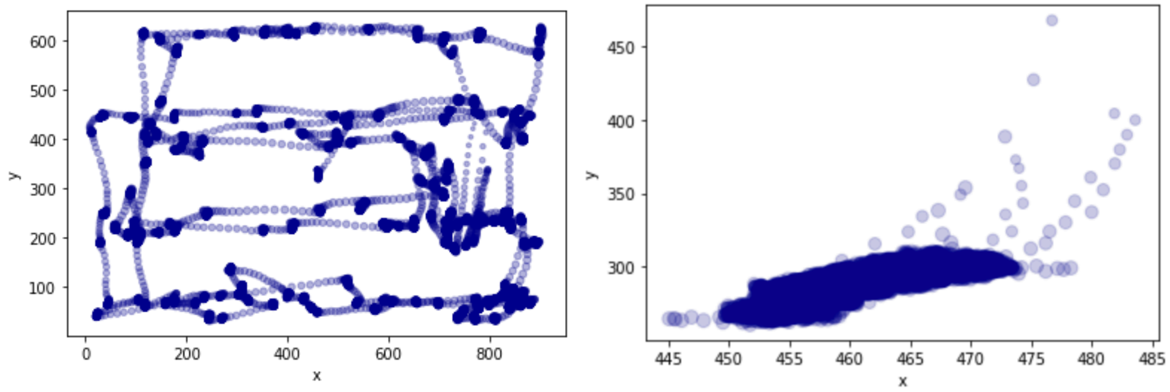
Thentu [9] presented a novel technique by using image representations of the scan path information for classification. The author took the average of the LXpix and RXpix values to get the resultant x co-ordinate, and LYpix and RYpix values to get the Y co-ordinate to plot one single gaze point, and such subsequent points gave rise to the entire scan path image representation.



(a) (b)

Figure 6 Scanpath of the "Waldo" image in Figure 2 for a random user in (a) Free-viewing and (b) Fixation conditions

But we can imagine that not all gaze points are distinct from each other and some amount of overlap is inevitable. Would we not look at the same point twice if we find something interesting on it? By adjusting the opacity of the pixels, we can easily see what points are looked at repeatedly i.e. where the user is focusing more. This can be visualized in Figure 7. The points where there are darker colored pixels probably are points of interest for the user than the ones which are lighter.

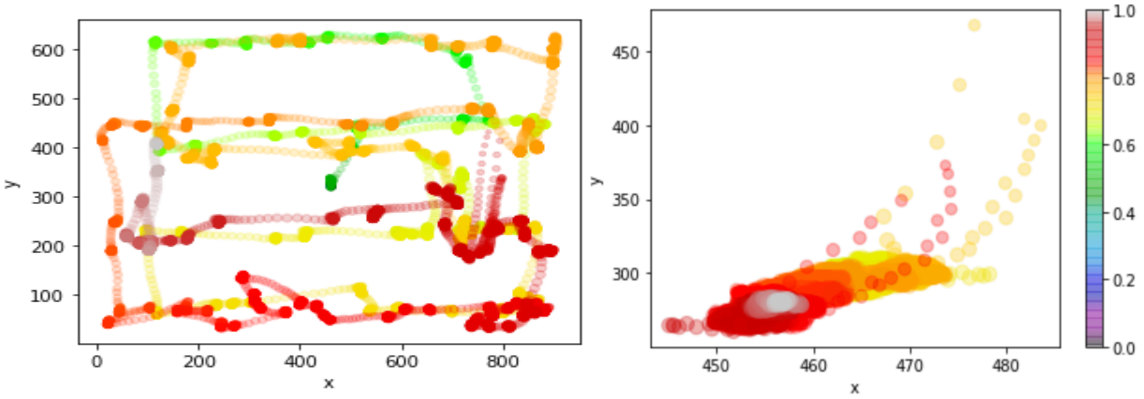


(c)

(b)

Figure 7 Scanpath of "Waldo" image in Figure 2 in (a) Free viewing and (b) Fixation conditions where darker points show points looked at repeatedly

Pupils are known to dilate in darker environments to let more light in while to contract in brighter environments. In addition to this relationship of pupil diameter with the illuminance of the environment, studies [15] have shown that pupil size can be affected by the degree of attention and the type of emotion that the target invokes. The “degree of attention” finding is especially interesting to us. To further help analyze the theory that pupil dilation plays a role 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 7, we adopted a color range to map the range of pupil values found across the dataset. Meaning, 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 said to be emitting information on a uniform level. Thus, in our experiment, we can consider *opacity* to denote “interest” and *color* to denote “pupil dilation”. We generated such 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.



(a) (b)

Figure 8 Scanpath of "Waldo" image in Figure 2 with colors reflecting pupil dilation in (a) Free-viewing and (b) Fixation conditions

A. Choice of color scheme

Would it matter if we chose only shades of red to represent this data? Or just red and orange? Would it be better if there was a contrasting element? Contrasting values, as opposed to monochromatic values, would be able to better distinguish different pupil values, and for computer vision tasks, this distinction can be important [16]. We assessed the idea of how important a role, color could play for the task of classification. We ran 1 layered CNNs on scan paths represented by (a) shades of blue as shown in Figure 7 and (b) contrasting red and green shades (option 'nipy_spectral' in the Python package Matplotlib) as represented in Figure 8. The first approach gave us accuracies of 62%, 71%, 77% while the second approach gave us accuracies of 66%, 75%, 79% for classification on 4 classes, 3 classes, and 2 classes respectively keeping the rest of the parameters same. Even this slight improvement can be useful which led us to choose the latter approach for the rest of our experiments.

VI. DEEP LEARNING

Convolutional Neural Networks are known to be successful in identifying patterns well in images that are not necessarily evident to the human eye. Unlike most other machine learning algorithms, the process of feature extraction and selection is not on the shoulders of the experimenter but is taken care of by the algorithm itself. The image to be classified is the only input required. Scan path images that were generated for our experiments had attributes that matched the needs of this algorithm. We alone as humans cannot easily identify all task types from the images. But, a CNN might find some useful patterns that make a “Waldo” scan path different from a “Natural” scan path.

CNNs usually have an input layer, one or more convolutional layers and max-pooling layers, and finally an output layer. The convolutional layers are responsible for extracting features with the help of some “filters”. The number of such layers, the number of filters used, and the size of the filters are parameters that can affect the performance of the network.

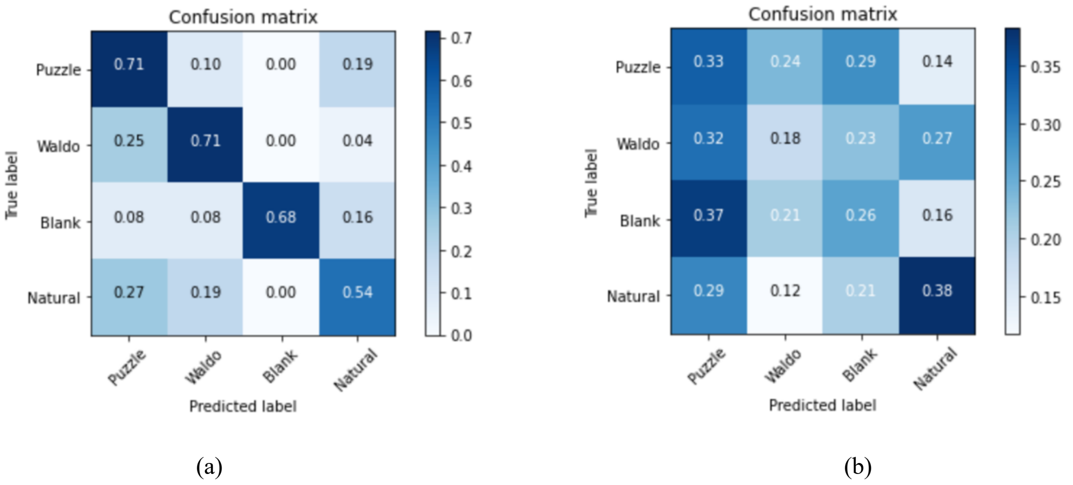
In the following sections, we have run our models on both Free-viewing and Fixation conditions while considering all 4 classes for classification as well as two additional subsets of classes. As mentioned earlier, of the four task types, two are visual search type tasks, one is a visual exploration type and the other a blank scene type. We wanted to see if the models could differentiate between the visual exploration tasks versus the blank task, hence one of the subsets includes only the Puzzle, Waldo and Blank task types and eliminates the Natural type for classification. As a third experiment, we now also eliminated the Blank task type and assessed how the models could differentiate between the two Visual Search tasks of Puzzle and Waldo.

A. 1 Layered CNN

Before we moved ahead to employing deeper CNN models, we first ran 1 layered CNNs to establish a baseline. This shallow architecture included a single convolutional layer with 16 3*3 filters.

1. Results – 4 classes

The confusion matrices in Figure 9 show us that the classification for the "Natural" condition is the least accurate for the Free-Viewing condition. The Natural task type suffers the most probably because of the similarity in the viewing patterns in Blank and Natural task types. Another consideration to note is that the task of exploration likely depends significantly on the user's cognitive functions. The "idea" of exploration can vary from user to user. On the contrary, searching tasks of either finding an object (Waldo) or finding the difference (Puzzle) demand some specific information from the user. Visual Exploration tasks do not ask the user to perform a specific activity, hence classifying these tasks can be challenging due to the lack of existence of evident and uniform patterns.



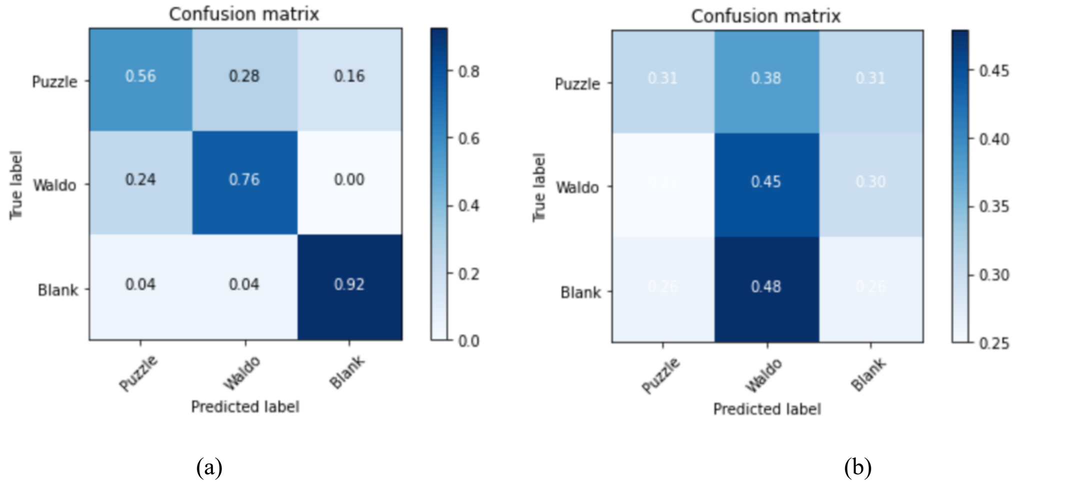
(a)

(b)

Figure 9 Confusion Matrices (4 classes) using 1 layered CNN in (a) Free-viewing and (b) Fixation conditions

2. *Results – 3 classes (Puzzle, Waldo, Blank)*

Upon eliminating the “Natural” class, we now tested if running the algorithm on only the other three classes gives us any better performance. We see that in Figure 10(a) the “Blank” class now seems to perform quite well with a sensitivity of 92%. The model can differentiate between Visual Search task types vs Blank screen task types in the Free-Viewing condition.



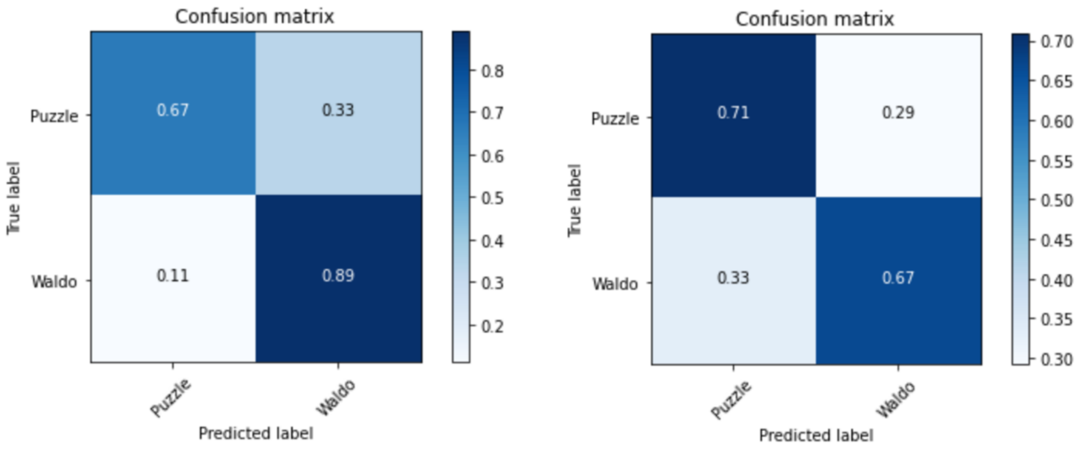
(a)

(b)

Figure 10 Confusion Matrices (3 classes) - 1 layered CNN on (a) Free-viewing and (b) Fixation conditions

3. *Results – 2 classes (Puzzle, Waldo)*

Figure 11 shows the results in the case of the experiments on only two classes. Results show only a slight improvement for the Free-viewing condition, but a drastic jump for the Fixation condition. An explanation for it could possibly be the fact that even if the users are asked to fixate at the center of the image, some amount of distraction is bound to happen. The distraction provided by the Puzzle and Waldo tasks can be a little different and specific since in the Puzzle images, users might get distracted sideways in a restrained attempt to compare while in the Waldo image the distraction could take place in all the directions in a restrained attempt to locate Waldo. The Blank image adds no specific distraction and the wavering gaze patterns found in this case can be random and possibly act as noise for the classifiers.



(a)

(b)

Figure 11 Confusion Matrices (2 classes) - 1 layered CNN on (a) Free-vieing and (b) Fixation conditions

Figure 12 summarizes the results for the shallow CNN.

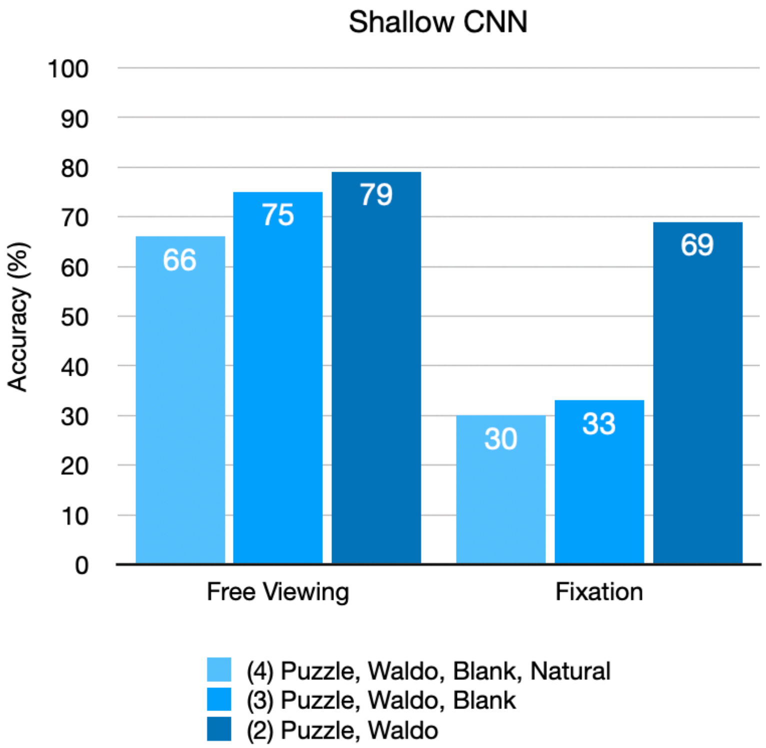


Figure 12 Summary of results – I layered CNN

The experiments helped us establish a starting point for our study. We did find promising results and also some possible ways to extract the most out of the data at hand.

B. Deep CNNs

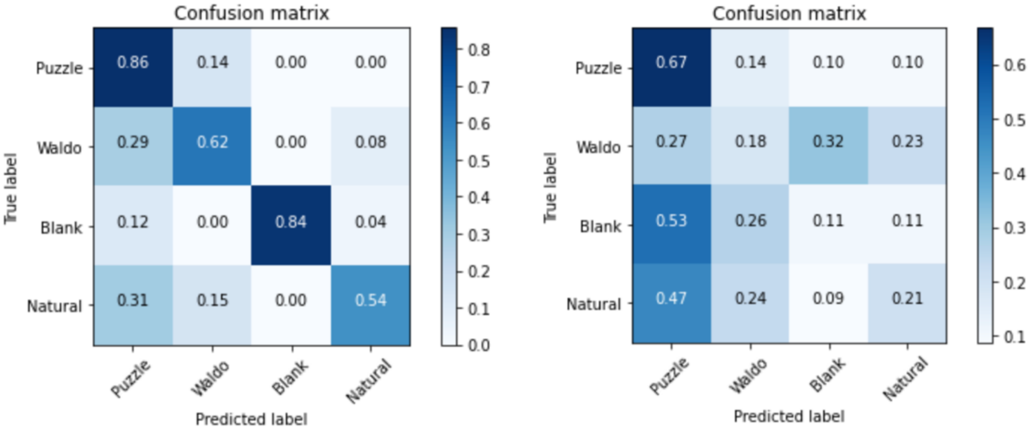
For the 1 layered CNN, we ran our models on images of size 64*64. Experiments on Deeper CNNs were conducted on different image sizes from 64*64, 128*128, and 256*256 to see how the performance of the CNN was affected as the image size changed. Our model had a depth of 3 and we tested the number of filters to be either 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 3, 4, and 5 summarize results by using CNNs on different image sizes. The discussion on the choice of the subset of classes and the probable reasons for the effect on accuracy upon the elimination of classes is carried forward from the previous sections.

1. *Results – 4 classes*

Results in Table 3 show that higher image size is helpful, but going any further is not beneficial. We experimented with image sizes of 256*256 and the results did not add any improvement, rather dropped in the accuracy.

Table 2 Summary of Results using CNN (4 classes)

Deep CNN, 4 classes	**64*64**	**128*128**	**256*256**
Free Viewing	60%	70%	54%
Fixation	26%	28%	25%



(a) (b)

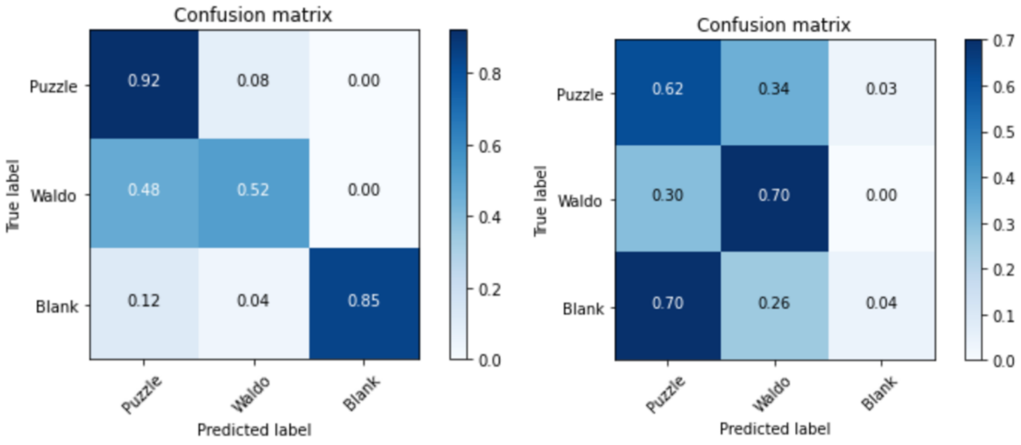
Figure 13 Confusion Matrices (4 classes) - Deep CNN on (a) Free-viewing and (b) Fixation conditions

2. *Results – 3 classes (Puzzle, Waldo, Blank)*

Results in Table 4 tell a different story about the image sizes. Here, it is seen that a smaller image size fared better than the 128 that performed well for 4 classes. Again, in this case, experiments on 256*256 yielded insignificant accuracies.

Table 3 Summary of Results using CNN (3 classes – Puzzle, Waldo, Blank)

Deep CNN, 3 classes	**64*64**	**128*128**	**256*256**
Free Viewing	78%	74%	62%
Fixation	46%	31%	33%



(b)

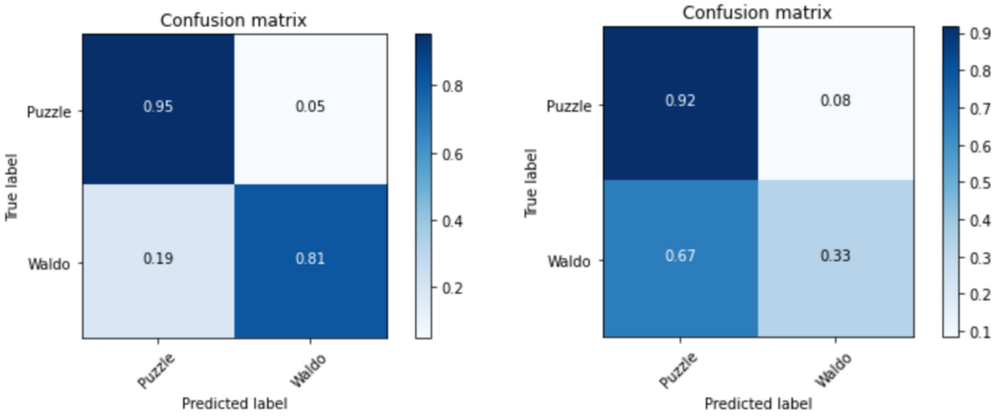
(b)

Figure 14 Confusion Matrices (3 classes) - Deep CNN on (a) Free-viewing and (b) Fixation conditions

3. Results – 2 classes (Puzzle, Waldo)

Table 4 Summary of Results using CNN (2 classes – Puzzle, Waldo)

Deep CNN, 2 classes	**64*64**	**128*128**	**256*256**
Free Viewing	88%	81%	78%
Fixation	62%	52%	52%



(c)

(b)

Figure 15 Confusion Matrices (2 classes) - Deep CNN on (a) Free-vieing and (b) Fixation conditions

4. Summarizing Deep CNN results

The accuracy does not come close to what Random Forests could achieve. One of the reasons could be the size of the dataset. Our dataset has 480 observations, giving us only 480 images in total for both train and test. CNNs are known to suffer from smaller datasets and perform their best when presented with huge training samples. To acknowledge this issue, we employed the techniques of transfer learning and data augmentation.

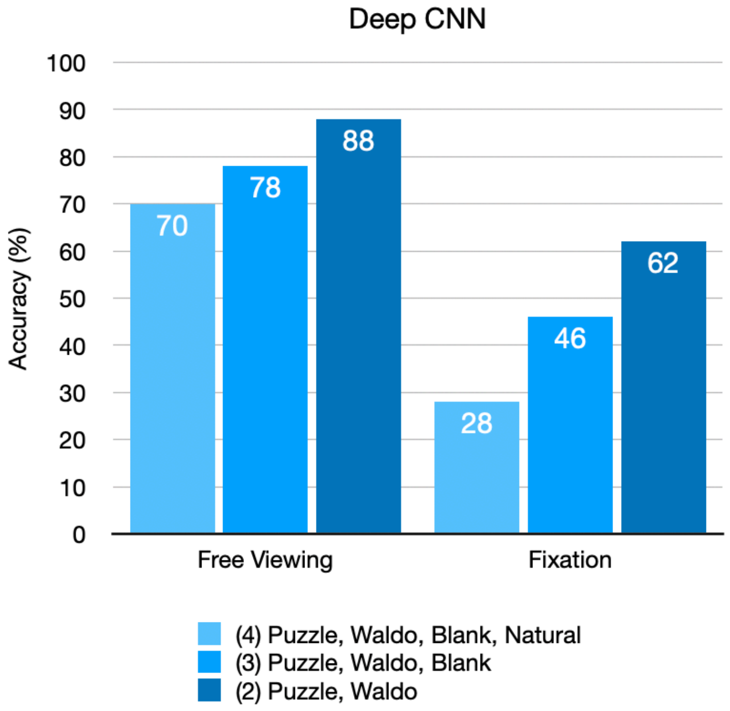


Figure 16 Summary of results - Deep CNN

C. Transfer Learning

Transfer Learning is the process in which the "learning" achieved in one problem is used to solve another problem [17, 18]. It can be a useful technique to apply when the data available is not enough for the problem and some pre-learned "knowledge" can be helpful. By using the MobileNetV2 pre-trained model and then adding a fully connected layer ahead, we were able to achieve the following results shown in Table 7-9. MobileNetV2 uses depth-wise separable convolutions as building blocks [21].

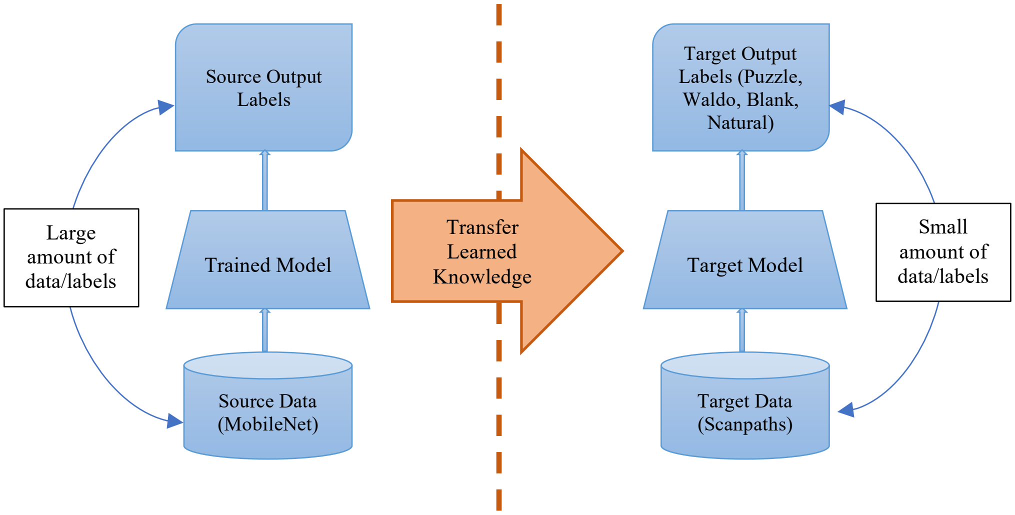

Source Output Labels
Trained Model
Source Data (MobileNet)
Large amount of data/labels
Transfer Learned Knowledge
Target Output Labels (Puzzle, Waldo, Blank, Natural)
Target Model
Target Data (Scanpaths)
Small amount of data/labels

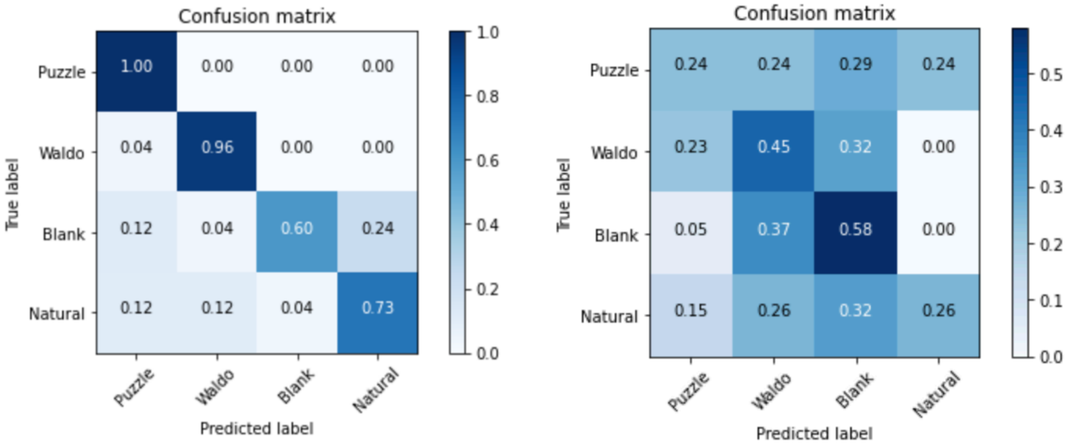
Figure 17 The process of Transfer Learning

1. Results – 4 classes

With Transfer Learning used, we see promising improvements. Even with an image size of 256*256, we now see comparable results. Fixation condition still seems to struggle.

Table 5 Summary of Results using CNN with Transfer Learning (4 classes)

CNN + Transfer Learning, 4 classes	**64*64**	**128*128**	**256*256**
Free Viewing	65%	81%	80%
Fixation	36%	29%	23%



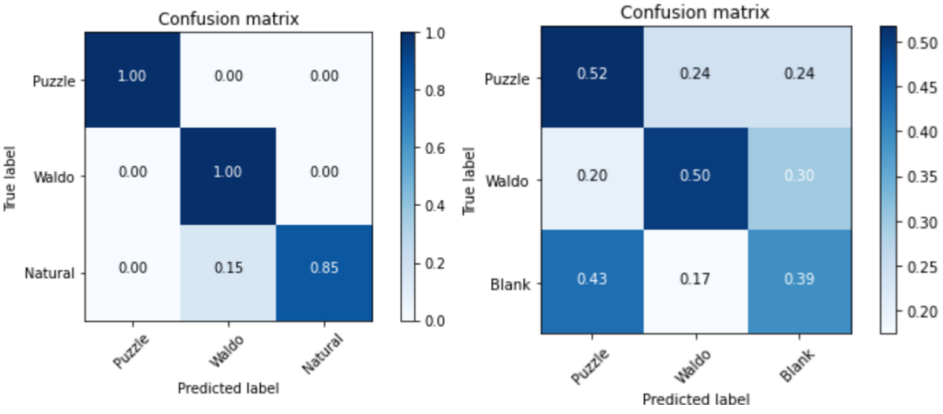
(a) (b)

Figure 18 Confusion Matrices (4 classes) - Transfer Learning on (a) Free-viewing and (b) Fixation conditions

2. *Results – 3 classes (Puzzle, Waldo, Blank)*

Table 6 Summary of Results using CNN with Transfer Learning (3 classes – Puzzle, Waldo, Blank)

CNN + Transfer Learning, 3 classes	**64*64**	**128*128**	**256*256**
Free Viewing	72%	90%	95%
Fixation	40%	47%	46%



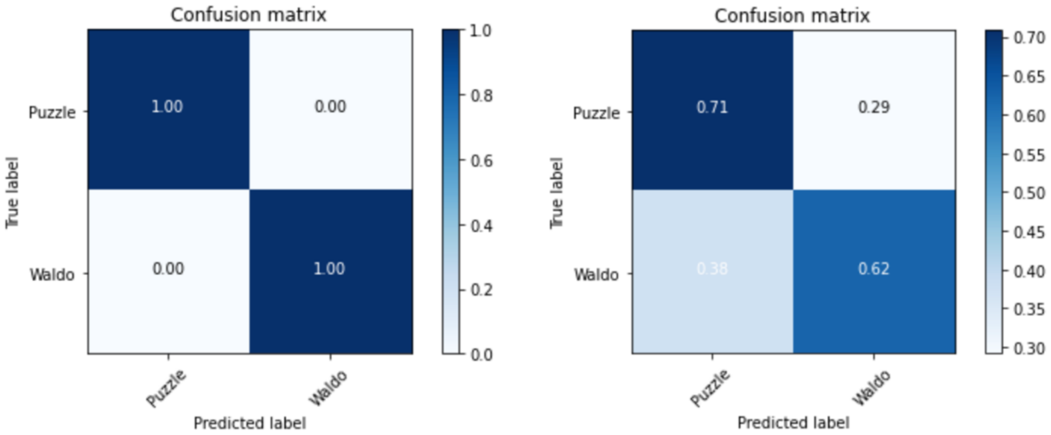
(a) (b)

Figure 19 Confusion Matrices (3 classes) - Transfer Learning on (a) Free-viewing and (b) Fixation conditions

3. Results – 2 classes (Puzzle, Waldo)

Table 7 Summary of Results using CNN with Transfer Learning (2 classes – Puzzle, Waldo)

CNN + Transfer Learning, 2 classes	**64*64**	**128*128**	**256*256**
Free Viewing	90%	100%	100%
Fixation	60%	68%	65%



(a) (b)

Figure 20 Confusion Matrices (2 classes) - Transfer Learning on (a) Free-viewing and (b) Fixation conditions

4. Summarizing Transfer Learning Results

Our efforts for transfer learning do yield significantly improved accuracies than without employing the technique. The model can perfectly distinguish between “Puzzle” and “Waldo” tasks in the free-viewing condition while it can classify about 97% correctly if “Blank” is added to the data. Classification on all 4 classes still doesn’t match the results of Random Forests.

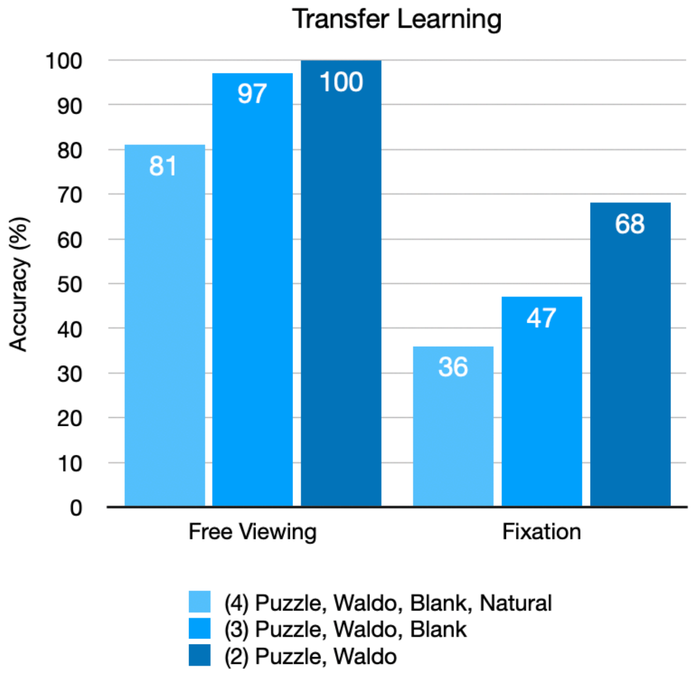


Figure 21 Summary of results – CNN with Transfer Learning

For the Fixation condition, however, the results don't compare with the Free-viewing counterpart. This can be understood because, even for the naked human eye, distinguishing the task types is difficult. Since the user is asked to fixate on the task, the scan paths do not differ much and hence the struggling classification accuracies.

D. Data Augmentation

Many machine learning tasks, especially computer vision tasks work the best with huge amounts of data [19]. Not always is the data available in such amounts and such tasks call for the need to synthetically produce data. In the case of our research too, we had a dataset of 480 images for both Fixation and Free-viewing. To augment it 3-fold, for each image, we generated 2 new images by randomly shifting each pixel by some small value in either the upward, downward, leftward or rightward directions giving us 2 synthetic images from every image and increasing our dataset to 1440 images from 480 images. We took this approach of pixel-wise shifting because our image data is highly positional and making drastic changes to the original images to generate new ones is not the best way of creating artificial data. We instead simulated different eye positions and generated new images accordingly. Figure 23 shows a sample of how a synthetically produced scan path image looks from its original scan path in Figure 22.

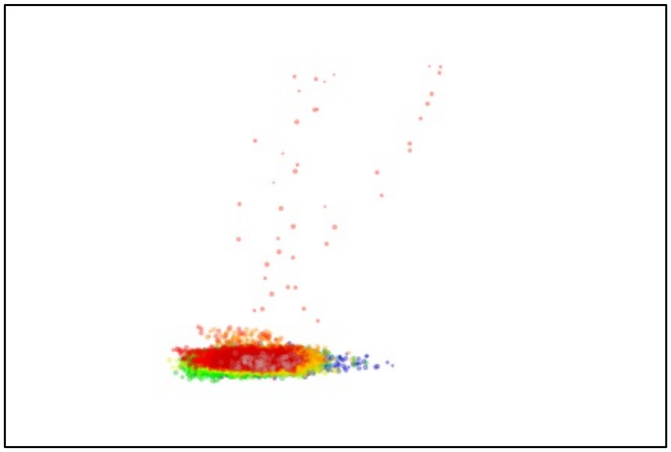


Figure 22 Original Scanpath of a random user looking at a Puzzle image for Fixation condition

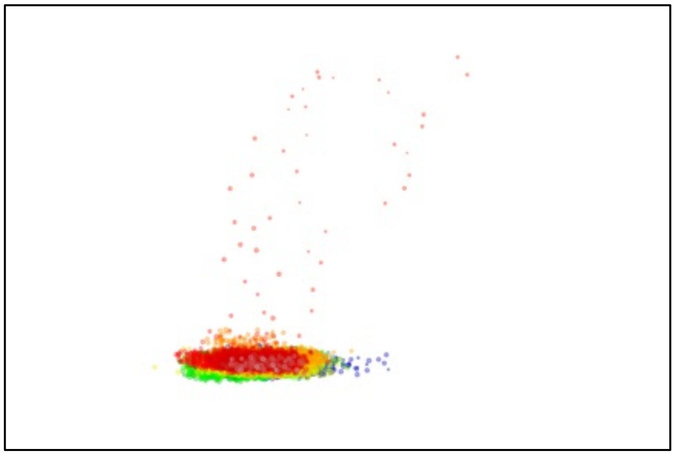
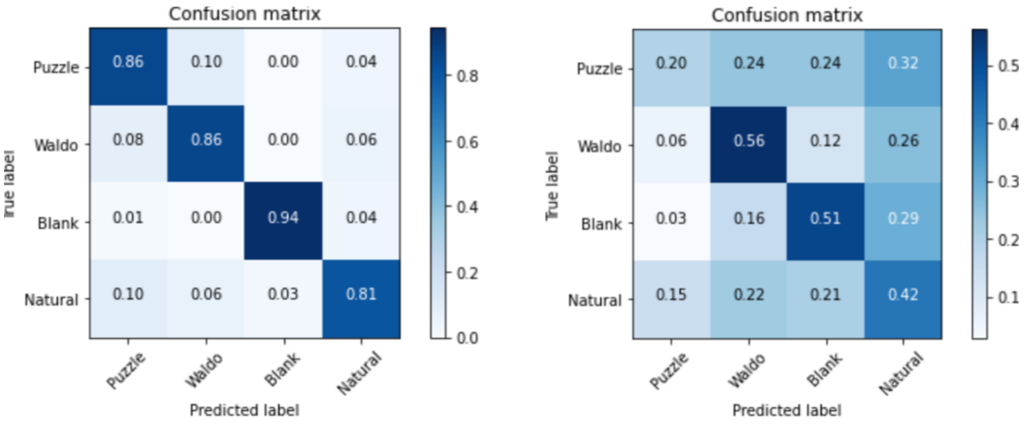


Figure 23 Synthetically produced scan path for a Puzzle image for Fixation condition

1. Results – 4 classes

Table 8 Summary of Results using CNN with Data Augmentation (4 classes)

Data Aug+CNN, 4 classes	**64*64**	**128*128**
Free Viewing	86%	25%
Fixation	43%	23%



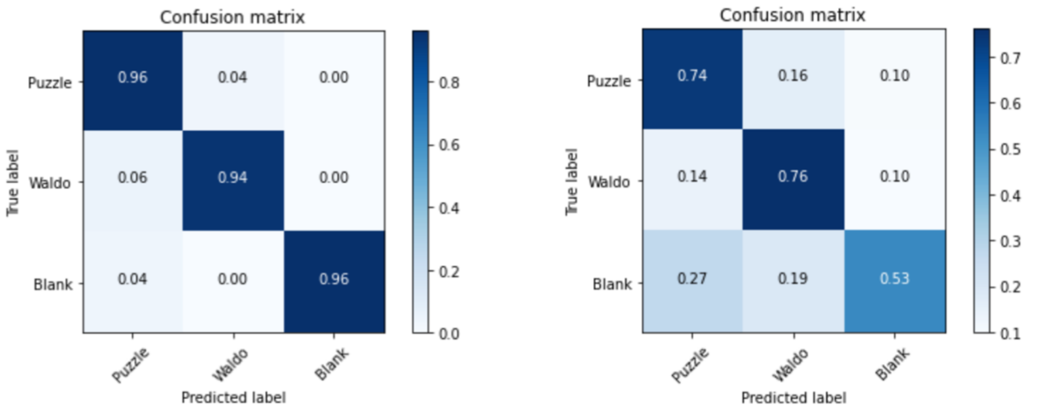
(a) (b)

Figure 24 Confusion Matrices (4 classes) - CNN on Augmented Data on (a) Free-viewing and (b) Fixation conditions

2. Results – 3 classes (Puzzle, Waldo, Blank)

Table 9 Summary of Results using CNN with Data Augmentation (3 classes)

Data Aug+CNN, 3 classes	**64*64**	**128*128**
Free Viewing	93%	95%
Fixation	62%	67%



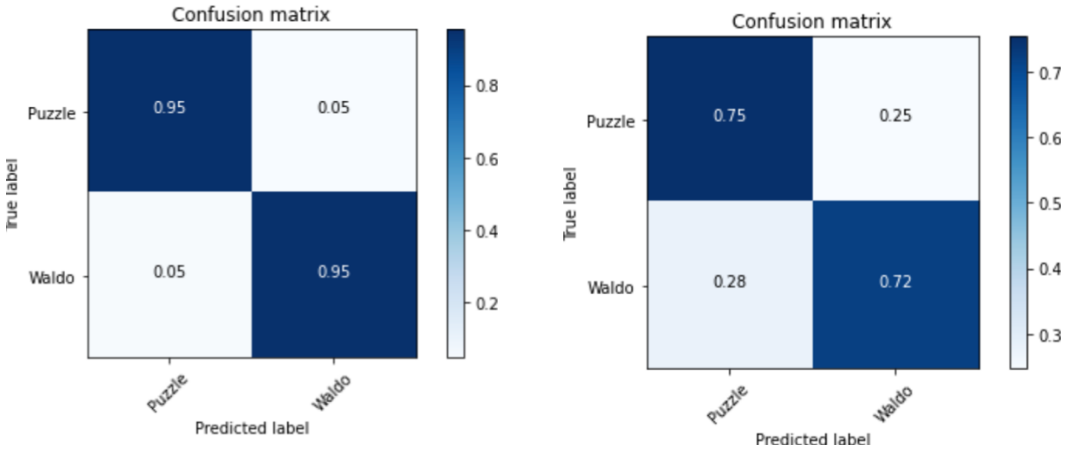
(b) (b)

Figure 25 Confusion Matrices (3 classes) - CNN on Augmented Data on (a) Free-viewing and (b) Fixation conditions

3. *Results – 2 classes (Puzzle, Waldo)*

Table 10 Summary of Results using CNN with Data Augmentation (2 classes)

Data Aug+CNN, 2 classes	**64*64**	**128*128**
Free Viewing	95%	90%
Fixation	73%	74%



(c) (b)

Figure 26 Confusion Matrices (2 classes) - CNN on Augmented Data on (a) Free-viewing and (b) Fixation conditions

4. Summarizing Data Augmentation results

Data Augmentation seems to have helped both viewing conditions pretty well, but especially the Fixation condition. Transfer Learning could only improve so much, but with this approach accuracy in classifying all 4 classes jumped from 28% to 43%, 38% to 67% for 3 classes, and 56% to 74% for 2 classes. Again, these results are not the best and are still not comparable to Random Forests.

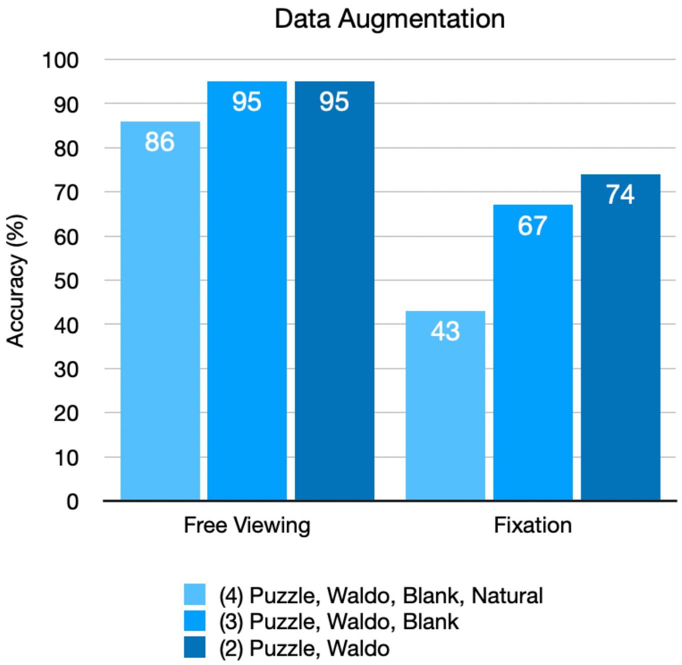


Figure 27 Summary of results - CNN with Data Augmentation

VII. Results and Discussions

This section consolidates and summarizes all the results and findings. The aim is to find the best model that fits this task and in the process also find the answer to the question – “Does pupil play an important role in distinguishing tasks?” The label ‘ConvNet+SVM’ just picks up results from the work in [9] for comparison where the author combined CNNs with SVMs to classify the image representations (without the pupil information).

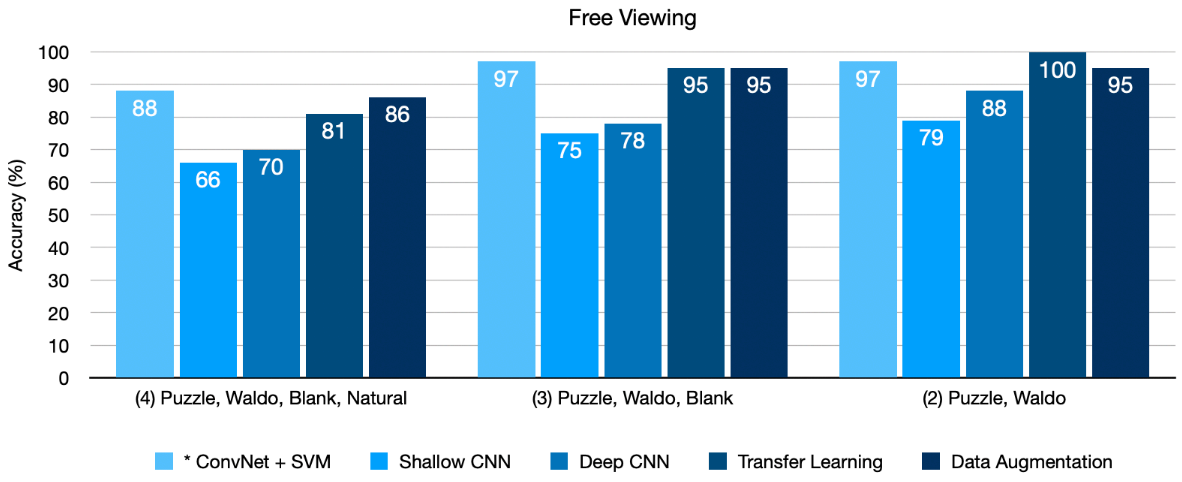


Figure 28 Summary of results for the Free Viewing condition

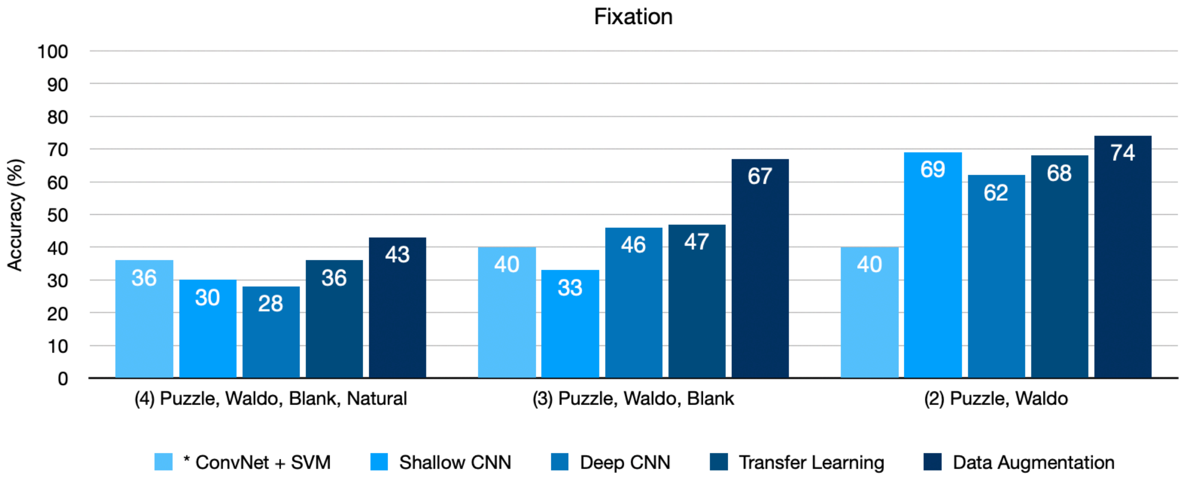


Figure 29 Summary of results for the Fixation condition

Figure 28 and Figure 29 summarize the results for the Free Viewing and Fixation conditions respectively. Classification on all 4 classes perhaps performs the worst. Of the models experimented as part of this work, CNN on augmented data gives the best results for both viewing conditions. Results with the use of Transfer Learning come a close second. But these results are in no way comparable to results from Random Forests. For both the classification between 3 classes and 2 classes, we observe that for the Free-Viewing condition, out of the CNNs that were experimented, Transfer Learning gives the best results whereas, for Fixation, Data Augmentation helps. We can say that for the Free-Viewing conditions, accuracy shows a promising increase when deeper networks are used.

Generally speaking, the free-viewing condition fares better than the fixation condition. In none of the experiments does the fixation condition come close to the accuracy achieved in raw data using Random Forests. Does adding pupil information help? Looking at the classification results for 4 classes, we cannot see a significant improvement in accuracy by including the pupil data. By comparing the results of our "Deep CNN" implementations with those from [9] using ConvNet + SVM, the results are comparable but do not exceed for the Free Vewing condition. On the contrary, adding pupil information shows improved results for the Fixation condition.

Although computer vision tasks are usually greatly solved with the help of CNNs, we observed, with our problem, that it is not always true. Machine Learning models like Random Forests seem to be working best in both the cases of Fixation and Free-Viewing conditions, but CNNs show comparable results with the help of techniques like Transfer

Learning and Data Augmentation for the Free-Viewing condition. For the Fixation condition, the models do improve in accuracy than previous CNN implementations.

VIII. CONCLUSION AND FUTURE WORK

In this study, we viewed the problem of task classification as a computer vision problem and tried to incorporate pupil information in our representation. Results showed that simple 1 layered CNNs were not equipped for this problem. As a solution, we tried Deeper CNNs that improved the accuracy as the added layers were able to extract more information. In addition to that, to combat the problem of a small dataset, we implemented the technique of Transfer Learning due to which 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, especially for the Fixation condition.

Although these results were promising, they do not yet compare to the Random Forest results. Machine Learning is still the winner. But an important point that can be observed by comparing the performance gain from CNNs in both viewing conditions is that CNNs techniques do better in situations where visual attention is maximum. Free-viewing conditions require the user for active participation. Whereas, the CNNs still struggle when user attention is minimal and when the visual representation fails to deliver all the information. The human eye as well as the models struggle to visually distinguish between different task type scan paths with low attention. We focused on the computer vision aspects of this problem as part of this study, but other studies can focus on leveraging the sequential information in the data and employing algorithms like HMMs, LSTMs and ensemble of classic machine learning algorithms.

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 behavior and user psychology combined with computational techniques open the paths to numerous researches.

References

[1] D. Ripley, T. Politzer. 2010. “Vision Disturbance after TBI”, NeuroRehabilitation vol. 27 pp215–216 doi 10.3233/NRE-2010-0599

[2] A. Yarbus, 2013, Eye movements and vision. Springer.

[3] A. Kumar, A. Tyagi, M. Burch, D. Weiskopf, and K. Mueller. 2019. “Task classification model for visual fixation, exploration, and search” *Proc. of the 11th ACM Symposium on Eye Tracking Research & Applications (ETRA '19).* Association for Computing Machinery, NewYork, NY, USA, Article 65, 1–4

[4] S. Hutt, J. Hardey, R. Bixler, A. Stewart, E. Risko, and S. D'Mello. 2017. “Gaze-Based Detection of Mind Wandering During Lecture Viewing”. *International Educational Data Mining Society.*

[5] 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. DOI: 10.3758/s13428-017-0857-y

[6] Z. Wang, T. Oates (2015). “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*

[7] J. Otero-Millan, X. Troncoso, S. Macknik, I. Serrano-Pedraza, and S. Martinez-Conde. 2008. “Saccades and microsaccades during visual fixation, exploration, and search: foundations for a common saccadic generator” Journal of vision 8, 14 (2008), 21–21.

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

[9] S. Thentu, "Task Classification during Visual Search with Deep Learning Neural Networks and Machine Learning Methods" (2021). *Master's Projects*. 1028. https://scholarworks.sjsu.edu/etd_projects/1028

[10] N. Attar, M. H. Schneps and M. 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

[11] O'Connell T. Walther D. “Fixation patterns predict scene category.” Journal of Vision, 12 (9): 801, 2012

[12] 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

[13] 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

[14] N. Attar, M. H. Schneps and M. Pomplun, "Working memory load predicts visual search efficiency: Evidence from a novel pupillary response paradigm." Memory & Cognition, 44(7):1038–1049, 2016.

[15] Kang OE, Huffer KE, Wheatley TP (2014) "Pupil Dilation Dynamics Track Attention to High-Level Information." PLoS ONE 9(8): e102463. https://doi.org/10.1371/journal.pone.0102463

[16] Gomez-Villa et al., "Color illusions also deceive CNNs for low-level vision tasks: Analysis and Implications", Vision Research Oxford, 2020-11, Vol.176, p.156-174

[17] 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.

[18] Li. Xuhong et al., "Transfer learning in computer vision tasks: Remember where you come from", Image and vision computing, 2020-01, Vol.93 (103853), p. 103853

[19] 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

[20] L. Chenchuang et al. " Review of Image Data Augmentation in Computer Vision", Journal of Computer Engineering and Applications Beijing, 2021-01 Vol.15(5), p.583-611

[21] [Website] https://ai.googleblog.com/2018/04/mobilenetv2-next-generation-of-on.html