Multi-Stage Hybrid-CNN Transformer Model for Human Intent-Prediction

Cyrille Cervantes, Matthew De Mesa, Joshua Ramos, Stephen Singer, Dale Joshua del Carmen, Rhandley D. Cajote
Electrical and Electronics Engineering Institute
University of the Philippines Diliman
Quezon City, Philippines
{ cyrille.cervantes, matthew.de.mesa, joshua.berbigal.ramos, stephen.singer, dale.del.carmen, rhandley.cajote, } @eee.upd.edu.ph

***Abstract*—Human intention prediction (HIP) is one aspect of Human-Robot Interaction (HRI) that could facilitate understanding and improving how humans interact with robots and computers. However, current gaze-based intent prediction models that perform well often require invasive methods using specialized equipment. In this paper we present a non-invasive, contactless method for predicting human intentions using a multi-stage hybrid CNN-Transformer framework. The model consists of a depth estimator and two key components: a gazed object predictor and a human intent classifier. The gazed object predictor is a modified Detection-Transformer (DETR) and used a ResNet50 backbone for feature extraction and obtained an accuracy of 32.15% in the custom dataset. Meanwhile, the human intent classifier is a transformer-based classifier that achieved a 98% accuracy when predicting human intention based on a series of gazed objects. The resulting cascaded HIP system attained an accuracy of 54%.**

***Index Terms*—CNN, computer vision, gaze, HRI, intent prediction, transformer**

I. Introduction

As robots with Artificial Intelligence (AI) become increasingly prevalent, the study of how humans interact with robots, Human-Robot Interaction (HRI), becomes increasingly important [1, 2]. HRI examines various factors such as the robot's design and sociocultural contexts to ensure that robots function in ways acceptable to humans and are beneficial to society.

One method being pursued by researchers to develop HRI is utilizing Human-Intent Prediction (HIP) systems that use Machine Learning (ML) algorithms to analyze and predict human behavior. Several HIP systems have been designed with different human behavior as input data. These include human motion [3, 4], gaze behavior [5, 6], and even bioelectric signals [7, 8]. While most have high accuracy, they often require specialized equipment or invasive sensors (e.g., eye trackers, RGB-Depth cameras, and 3D cameras) to gather data from the subject, which can be impractical or inconvenient.

In this paper, the researchers present a contactless HIP model that accounts for temporal gaze behavior. A CNN-Transformer hybrid model architecture is used, with depth information to improve the gaze estimation accuracy [9]. The depth information is estimated via a monocular depth estimation model.

II. Methodology

The proposed framework consists of two main models: the gaze-object predictor and the human intent classifier. The gaze-object predictor is responsible for predicting the gazed object of the subjects per frame. Then, these gazed object predictions will be inputs to the intent classifier to determine the intention of the subject. This process is properly illustrated in Fig. 1.

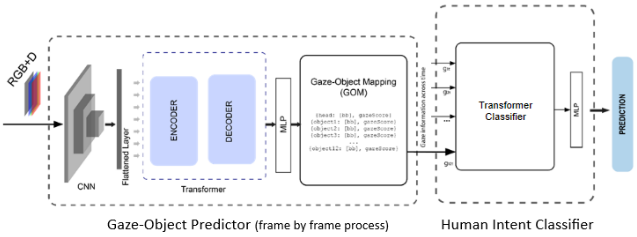
RGB+D
CNN
Flattened Layer
ENCODER
DECODER
Transformer
MLP
Gaze-Object Mapping (GOM)
{head: [bb], gazeScore}
{object1: [bb], gazeScore}
{object2: [bb], gazeScore}
{object3: [bb], gazeScore}
...
{object12: [bb], gazeScore}
Gaze information across time
Transformer Classifier
MLP
PREDICTION
Gaze-Object Predictor (frame by frame process)
Human Intent Classifier

Figure 1. Proposed Framework.

A. Gaze-Object Mapping Dataset

1) Dataset Specifications: To develop a computer vision model that can perform HIP, a custom dataset was developed consisting of video captures of participants looking at objects and having an intention in mind. Each frame has an annotation of the bounding boxes and classification of objects present and the human head. It also has a target label of what object is being gazed at. Each video sample has a target label of the participant's intention. A visualization is shown in Fig, 2.

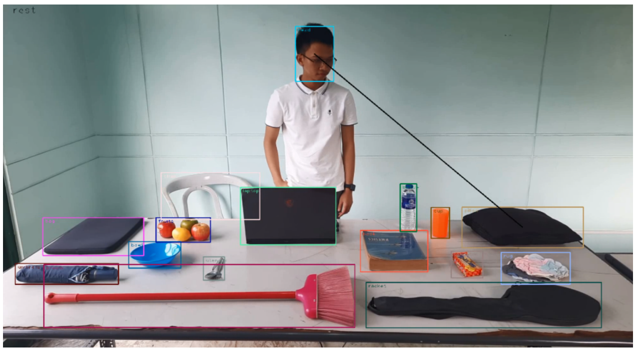

Figure 2. Example Frame Annotations

Participants were tasked to exhibit gaze behavior that implies their chosen intention. In each recording session, they were

This work was supported in part by the Republic of the Philippines, Department of Science and Technology - Science Education Institute

Table I
SUMMARY OF DATASET SPECIFICATIONS

Specification	**Value**
Number of Participants	20 (10 male, 10 female)
Number of Video Samples	1000 (50 per participant)
Video Sample Quality	1080p (Full HD), 30 FPS
Video Sample Duration	5-10 seconds
Intents	Go outside, Study, Eat, Drink, Clean the area, Rest, Indeterminate (Indiscernible), Ambiguous (No Specific Intention)
Objects	Umbrella, Bag, Racket, Laptop, Book, Utensils, Bowl, Sandwich, Fruits, Bottle, Cup, Broom, Rug, Pillow, Chair

recorded eight times, one for each intention. They decided the order of the intentions. In the last session, they chose two random intentions to complete fifty (50) video samples. The environment where the videos were recorded is controlled, except for the deliberate random location of the objects and the lighting conditions that vary across video samples.

2) Statistics: Participants tend to look at certain objects for a given intention. This correspondence is listed in Table II. In cases where a certain object can be tied to multiple intentions, e.g. Bag being under the intents of "Go Outside" and "Study", the intention would have to be evaluated based on the ratio of the number of frames per intention that can be attributed to that specific object. This means that the variability of the intent-distribution in the dataset would have a substantial effect on the performance of the system's intent classification.

Table II
INTENT-OBJECT CORRESPONDENCE

Intention	**Object**
Go Outside	Umbrella, Bag, Racket
Study	Laptop, Bag, Book
Eat	Utensils, Bowl, Sandwich, Fruits
Drink	Bottle, Cup
Clean the Area	Broom, Rug
Rest	Pillow, Chair
Indeterminate	None
Ambiguous	All

The dataset distribution is in Table III. To ensure low-count objects were selected, the random selection weight is the sum of the inverse count for each class in the video. The train-test split was the same for all model training and testing.

Intent Prediction is a classification task. Each intention or class is evaluated using the F1-Score since it emphasizes classes with smaller counts [10] to account for the dataset imbalance. This requires Precision and Recall metrics, which consider the True Positives (TP), False Positives (FP), and False Negatives (FN). If an element belongs to a class, it is TP if grouped to that class; Else, it's a FN. If the component is grouped into the wrong class, it's a FP. These metrics are averaged across classes k. Accuracy, the percentage of correctly classified elements over all predictions, is used for its intuition.

$$Precision_k = \frac{TP_k}{TP_k + FP_k} \tag{1}$$

Table III
DATASET DISTRIBUTION SUMMARY

ID	**Object**	**Training Set**		**Testing Set**		**Total**
		Count	***Rate***	***Count***	***Rate***	***Rate***
0	***None***	**25545**	**12.42%**	**17863**	**8.68%**	**21.10%**
1	***Bag***	5863	2.85%	639	0.31%	3.16%
2	***Book***	14086	6.85%	2234	1.09%	7.93%
3	***Bottle***	13785	6.70%	2765	1.34%	8.04%
4	***Bowl***	5802	2.82%	303	0.15%	2.97%
5	***Broom***	12599	6.12%	3256	1.58%	7.71%
6	***Chair***	4414	2.15%	165	0.08%	2.23%
7	***Cup***	7780	3.78%	826	0.40%	4.18%
8	***Fruits***	12099	5.88%	977	0.47%	6.36%
9	***Laptop***	9863	4.79%	1195	0.58%	5.37%
10	***Pillow***	14205	6.90%	4964	2.41%	9.32%
11	***Racket***	11395	5.54%	1098	0.53%	6.07%
12	***Rug***	8892	4.32%	1363	0.66%	4.98%
13	***Sandwich***	6196	3.01%	252	0.12%	3.13%
14	***Umbrella***	9305	4.52%	2115	1.03%	5.55%
15	***Utensils***	3826	1.86%	78	0.04%	1.90%
	Total	165655	80.51%	40093	19.49%	100.00

$$Recall_k = \frac{TP_k}{TP_k + FN_k} \tag{2}$$

$$F1 - Score_k = 2 \cdot \frac{precision_k \cdot recall_k}{precision_k + recall_k} \tag{3}$$

Participants performed HIP on the video samples for a human-based reference. The participants have a Macro F1 of 61.86% and Accuracy of 61.90%. The highest and lowest accuracies were 96% and 32.00% respectively. The prediction and performance per class are summarized in Table IV. Values are normalized across target classes. Note that recorded participants labeled their intention per recording as the ground truth.

Table IV
PARTICIPANT CLASS PERFORMANCE

		Target Class							
		A	***B***	***C***	***D***	***E***	***F***	***G***	***H***
Output Class	***A***	**0.59**	0.13	0.03	0.07	0.02	0.13	0.07	0.08
	B	0.05	**0.50**	0.08	0.05	0.02	0.08	0.01	0.06
	C	0.02	0.03	**0.68**	0.04	0.09	0.00	0.02	0.04
	D	0.04	0.03	0.00	**0.61**	0.04	0.05	0.07	0.03
	E	0.03	0.06	0.09	0.05	**0.61**	0.04	0.01	0.03
	F	0.17	0.15	0.03	0.06	0.02	**0.60**	0.03	0.08
	G	0.08	0.03	0.05	0.02	0.10	0.08	**0.70**	0.04
	H	0.03	0.08	0.04	0.09	0.09	0.02	0.08	**0.64**
Recall		0.51	0.58	**0.73**	0.70	0.67	0.55	0.65	0.60
Precision		0.59	0.50	**0.68**	0.61	0.61	0.60	0.70	0.64
F1 Score		0.55	0.54	**0.70**	0.65	0.64	0.57	0.68	0.62

A. Go Outside B. Drink C. Rest D. Clean the Area
E. Study F. Eat G. Indeterminate H. Ambiguous

3) Monocular Depth Estimator: To add depth information to the video frames, the PixelFormer [11] was used as it is one of the top performing monocular depth estimators in the current rankings [12, 13] with publicly available pretrained weights. The output of the PixelFormer model is transformed into a single grayscale image, with each pixel intensity representing proximity or depth, for each video frame. Subsequently, this grayscale depth image is appended as a fourth channel to the input data for the Gazed Object Predictor.

B. *Gaze-Object Predictor*

The gazed-object predictor is a hybrid CNN-Transformer architecture that is based on two existing models with similar applications: the DETR [14] and the MGTR [15]. The DETR is built for object detection and recognition tasks while the MGTR is designed for detecting mutual gazes between two human subjects. The gaze-object predictor's functionality is a combination of these two models' key characteristics. Not only should the proposed model be able to predict the gazed object by considering the contextual information from the frame, but it also needs to recognize what kind of object it is. Given that the MGTR model is based on the DETR architecture, it is highly feasible to develop a model for this specific use case by using these two models as guides.

The resulting modification of the models can be divided into three parts: a CNN backbone for feature extraction, an encoder-decoder transformer for the extraction of gaze characteristics and object detection and recognition, and a simple feedforward network with 2048 dimensions for gazed object prediction. In the first part, an additional 2D convolution layer was added to the backbone to take into account the additional depth information of the input. Moreover, in order to use the pretrained weights for the backbone, the weight from one of the RGB layers was used for the added convolution layer to make the feature extraction of RGB-D images possible. In the case of RGB images as inputs to the model, they will be padded with zeros as an additional depth layer to make it compatible with the added convolution layer.

For the encoder-decoder transformer, the format of the queries in the decoder layer is changed accordingly to output the desired gaze-object mapping. The utilized format was derived from the DETR, which contains the object classification and bounding box, with an additional gazed flag to determine if the object detected by the query is being gazed at by the subject. An additional linear layer was also cascaded to the model for the added component to the query.

Gaze-object prediction is also a classification task. The model is generally assessed in this task using Accuracy. Object detection is evaluated using the mean Average Precision (mAP). It is a metric often used for assessing bounding box predictions based on their overlap with the annotated ground truths and the relationship between Precision and Recall [16].

Variations. To further improve the base model's performance, various model architectures were developed. First, the depth information is no longer used, and thus the input CNN layer of the ResNet backbone is reverted back to its original structure. For convenience, it is called the No Depth variant. Second, an architecture that now requires a cropped image of the head as input. Since the base model is also capable of detecting heads, with Head variant was used. With this design, two backbones were needed: one for the scene and another for the head. Their outputs are concatenated at the transformer stage.

C. *Human Intent Classifier*

The primary objective of the human intent classifier is to effectively categorize the gaze behavior data obtained from the gazed-object predictor into human intentions. This is achieved by leveraging the temporal information present in the video, wherein the gazed objects over time are used as a sequence input to the classifier. The gazed object per frame is obtained by applying softmax and arg max functions to the logits generated by the gaze estimator model. In situations where the previous model yields multiple gazed objects, the one with the highest confidence score is selected.

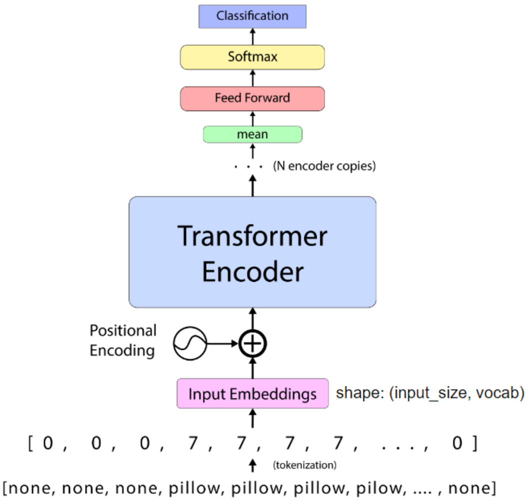

Figure 3. Human Intent Classifier Architecture.

Fig. 3 provides the visual diagram of the architecture of the human intent classifier. It adopts a transformer-based classifier design, composed only of the encoder component. In this framework, the sequence of gazed objects acquired from the preceding model is treated as a sequence of words. These words are subsequently tokenized according to its corresponding ID in Table III and passed through the input embedding layer, where a positional encoder is used before being fed into the transformer encoder. A sinusoidal positional encoding was used to integrate spatial information into the input embeddings of the encoder layers. Since the output of the transformer encoder has a shape of (input_size, vocabulary_size) a mean function is employed to flatten the dimensions to (input_size,), making it compatible with the Linear Feed Forward classifier module. Finally, a softmax function is applied to generate the classification results.

III. Results and Analysis

A. *Gaze-Object Predictor*

Various configurations and modifications were trained on a small scale by using only one participant's worth of dataset to find the best-performing setup before training with the whole dataset. Using the results from the small-scale training as the basis, the following large-scale training setups shown in Table V were configured and run. The skip in this table refers to the number of frames skipped on every video in the dataset for the purpose of avoiding over-fitting, since the frames near each other are similar, making it seem like a duplicate.

The first setup is one of the earliest runs with a configuration of one (1) batch size and five (5) skipped frames every other six (6) frames. The result is an accuracy of 17% on the

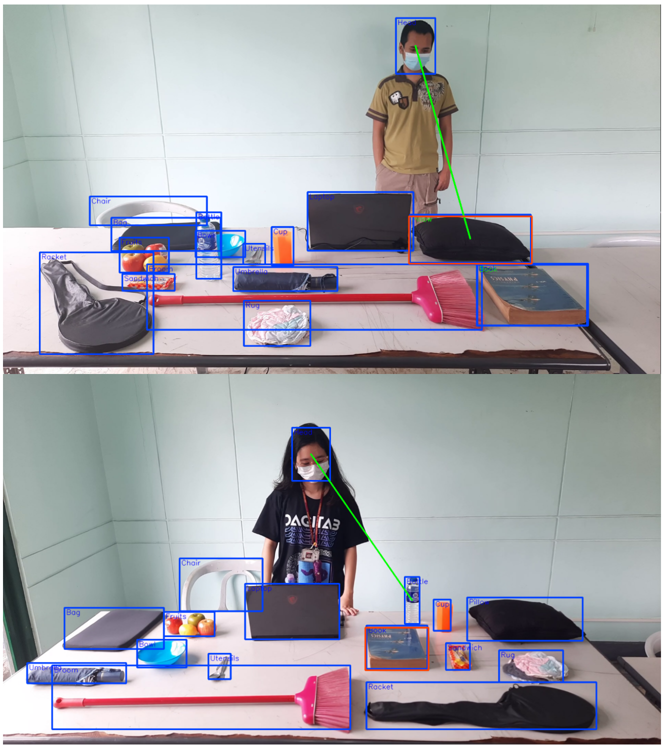

Figure 4. Visualization of the Gazed-object Predictor: Correctly classified instance (top) and Incorrectly classified instance (bottom).

test set. On the second setup, only the skipped frames were decreased to 4, which then resulted in a 29.72% accuracy. The third setup is similar to the small-scale training, where all the modified parts of the model to accommodate the depth layer were removed. This setup resulted in a 26.39% accuracy. The fourth setup is where the batch size is decreased to two (2), but the resulting accuracy is still low with an accuracy of 26.96%. Finally the last setup requires a separate head detector as a new input alongside the original scene input. This setup resulted in an accuracy of 32.15%.

In terms of mAP on the test set, the second setup had the lowest at 57.21 %, followed by the first at 61.48%, and then bested by the fourth setup at 67.90 %. The last setup had a mAP of 71.82% while the third setup that did not have depth performed the best with a mAP of 77.03%.

Table V
LARGE SCALE TRAINING RESULTS ON TEST DATASET

Setup	Accuracy	mAP
w/ Depth (Batch Size: 1, Skip: 5)	20.35%	61.48%
w/ Depth (Batch Size: 1, Skip: 4)	29.28%	57.21%
No Depth (Batch Size: 4, Skip: 4)	26.47%	**77.03%**
w/ Depth (Batch Size: 2, Skip: 4)	26.52%	67.90 %
w/ Head (Batch Size: 2, Skip: 4)	**32.15%**	71.82

B. Human Intent Classifier

The training and testing were conducted using the custom dataset developed by the researchers. Two variations of the model were developed: a model with six (6) transformer encoders and another with three (3) encoders. The models were trained in the augmented train dataset for 50 epochs using SGD (Stochastic Gradient Descent) with a learning rate of 0.001 and batch size of 10. To augment the training set, time-shift, time-stretch, and noise functions were applied. Moreover, the loss function utilized was categorical cross-entropy, which is well-suited for multi-class classification tasks. The same results were obtained from the two models. The experimental results show that both the two human intent classifiers achieved an accuracy of 98%, precision of 97.5%, recall of 96.74%, and F1 score of 97.07%. These results demonstrated the effectiveness of the proposed model in classifying human intentions based on gaze behavior data.

Table VI
INTENT CLASSIFIER TEST DATASET RESULTS

Model	Accuracy	Precision	Recall	F1 Score
6-Encoder Model	98.00 %	97.50%	96.74%	97.07%
3-Encoder Model	98.00 %	97.50%	96.74%	97.07%

C. Cascading the Models

After building and evaluating the gaze-object predictor and the human intent classifier individually, the two models are integrated to form the proposed model framework. The gaze-object predictor accepts RGB/RGB-D frames as input, which will be transformed into a sequence of predicted gazed objects per frame in the output. This gaze sequence will then be used as input to the human intent classifier. It will use this gaze sequence to classify the subject's intention, and output an intent prediction.

The best-performing model from both the gazed object predictor and the intent classifier were cascaded, producing the overall results shown in Table VII. For the gazed object predictor, the *with Head* variation with 32.15% accuracy was used. On the other hand, for the intent classifier, the three-encoder (3-encoder) setup was used as it is more robust to noise. This cascaded model attained an accuracy of 54%, which is 17.5% lower than the human accuracy.

Table VII
OVERALL RESULTS ON THE TEST DATASET

Model	Accuracy	Precision	Recall	F1 Score
Cascaded Model	54.00%	46.19%	50.01%	45.53%
Human	*71.5%*	*67.96%*	*63.65%*	*64.83%*

The normalized confusion matrix of this cascaded models setup is shown in Fig. 5, where it can be observed that the Ambiguous, Go Outside, and Indeterminate intentions possessed the highest chance of being correctly classified with normalized values at 0.75, 0.72, and 0.70 respectively. On the other hand, the rest of the intentions didn't even reach the halfway point of 0.5, with the Drink intention being the lowest at 0.24. This implies that the model is finding it difficult to correctly classify the rest of the intentions, especially for the Drink intention. One possible reason for this is that the gazed object detector performs poorly in correctly classifying the gazed objects associated with those intentions, particularly since some objects have smaller dimensions. As a consequence of this, the sporadic and incorrect output of the gazed object detector leads to wrong classifications by the intent classifier. Ultimately, the effectiveness of the whole system heavily relies on the performance of the gazed object detector.

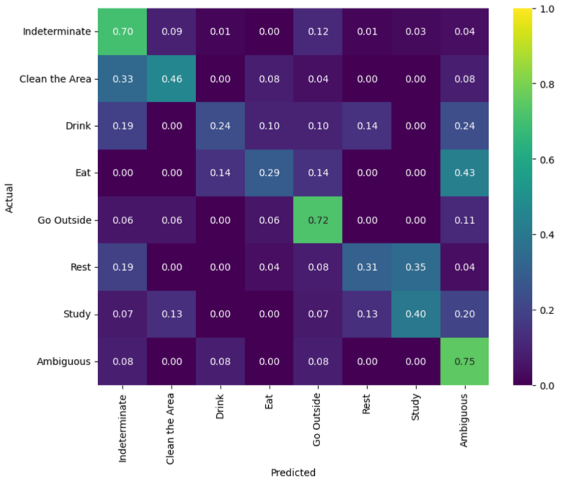

Figure 5. Confusion Matrix.

D. Comparison with Previous Works

1) Gaze-Object Prediction: The closest model that predicts the gazed object of an observer in a frame from a second-person reference is the GaTector [17]. It is trained with the GOO [18], a gaze estimation dataset in a retail setting, to determine an observer's gaze heat map. The model uses this to predict the bounding box and classification of the gazed object in a scene. However, unlike the Gaze-Object Predictor which focuses on classification, the GaTector evaluates the overlap and distance of the predicted bounding box from the ground truth using its novel metric, the wUOC (weighted union over closure). While a direct comparison can't be made due to the difference in metrics, the GaTector shares the same problem as the Gazed-Object Predictor of having difficulties in gaze estimation, achieving only a wUOC of 28.5% out of 100%.

2) Action Anticipation: Action Anticipation (AA) is similar to Human Intent Prediction, except it eventually shows the intention being performed while HIP omits any direct interaction. This allows models to perform better as more frames are shown. For instance, a Predictive Inverse Linear-Quadratic Regulation model [19] initially achieves an accuracy, macro precision, and macro recall of 66.7%, 50.1%, and 62.4% respectively, and reaches a perfect score for all metrics by the end for its predictions. It is trained on the Cornell Activity, a second-person-based RGB-Depth Video Dataset.

3) Human Object Interaction: Alternatively, predictions or anticipations may consider Human-Object Interactions (HOI), which provide more information and allows for hierarchical considerations. As an example, a Stochastic Context Sensitive Grammar-based event parser [20] can infer the current action being performed by a participant recorded in the second person with an accuracy of 90 %, and predict their next action with an accuracy of 87 %. It does this by considering the previous actions and inferring the remaining possible actions via an And-Or Graph (AoG). Intentions may be HOI, but HOI prediction may consider previous HOI which HIP omits.

4) Intent Prediction: To the best of our knowledge, no published paper was found that considers Human Intent Prediction using temporal gaze information, inferred from a second-person RGB video record of a subject, either with or without depth information. Current works focus on egocentric gaze information such as [21, 22].

IV. Conclusion and Recommendation

A. Conclusion

In this paper, the researchers presented a custom image dataset created for the purpose of training gaze-based intent-prediction deep learning models. The dataset consists of twenty (20) participants, fifteen (15) objects, and eight (8) intentions with a total of 205,748 individual RGB frames extracted and labeled from 100 videos and depth-generated using Monocular Depth Estimation techniques. While not entirely balanced in terms of object and intent distribution, the final dataset consists of over 200,000 frames that can serve as a solid foundation for further developments in AI learning when it comes to human intent-prediction using only monocular computer vision techniques and other similar topics. The researchers will make efforts to make this dataset available to the public in the future.

The researchers also presented a gaze-based intent-prediction system. This system uses two (2) sub-models trained using the created custom dataset, a gaze-object predictor model with an individual accuracy of 32.15% and a human-intent classifier model with an individual accuracy of 98%. Using the output of the gaze-object predictor model as the input to the human-intent classifier model results in the cascaded gaze-based intent-prediction system with a final over-all accuracy of 54%. The current training and testing results of the gaze-based intent-prediction system show that the concept of a Human Intent-Prediction Transformer is still in its very early stages and lacks the necessary accuracy to be used reliably in the wild. One of the main reasons why the overall accuracy is relatively low can be attributed to the imbalances in the dataset with regards to the intent and object distributions. An example of this would be the abundance of frames that contain no gazed objects, such as when a subject is looking outside the scene, transitioning between objects, or even staring straight at the camera. With the highest accuracy achieved in the cascaded system being less than that of an actual human guesser, it can be concluded that the dataset and the system still needs more fine-tuning in order to make both of them fit each other better as well as increase the current accuracy in each of the two sub-models for a better overall performance during individual and cascaded testings.

B. Recommendations

The researchers recommend the addition of more samples in order to increase the amount of variability in the dataset that will be used for training the model. It is also recommended to modify the train-test split of the dataset in order to ensure that there would be equally proportional amounts of training and testing for each object as well as for each intent. Another recommendation is to change the amount of the None objects in the dataset while training as it is currently the most numerous amongst all object types in the training dataset for the gaze-object predictor model. Furthermore, the current model in this paper uses pre-trained weights from the original gaze estimator that was modified to fit the intentions of the researchers. Therefore, it is recommended that future researchers re-train the model weights from scratch and compare the differences in gaze estimation, object classification, and overall system accuracy from the previous models.

Acknowledgment

The researchers acknowledge the contribution of the participants in the development of the dataset, and the graduate students and faculty of the University of the Philippines Diliman Electrical and Electronics Engineering Institute for making available the facilities and computing resources needed.

References

[1] K. Belhassein *et al.*, "Addressing joint action challenges in HRI: Insights from psychology and philosophy", *Acta Psychologica*, vol. 222, pp. 103–476, 2022. DOI: doi.org/10.1016/j.actpsy.2021.103476.

[2] A. Hayashi, L. K. Rincon-Ardila, and G. Venture, "Improving HRI with Force Sensing", *Machines*, vol. 10, no. 1, 2022. DOI: 10.3390/machines10010015.

[3] H. Sun, "Human Behavior Understanding and Intention Prediction", 2020. [Online]. Available: mospace.umsystem.edu/xmlui/handle/10355/88906.

[4] T. Zhou, Y. Wang, and J. Du, "Human Intent Prediction in Human-Robot Collaboration;A Pipe Maintenance Example", in *Construction Research Congress 2022*, pp. 581–590. DOI: 10.1061/9780784483961.061.

[5] T. Isomoto, S. Yamanaka, and B. Shizuki, "Dwell Selection with ML-Based Intent Prediction Using Only Gaze Data", *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 3, 2022. DOI: 10.1145/3550301.

[6] B. David-John, C. Peacock, T. Zhang, T. S. Murdison, H. Benko, and T. R. Jonker, "Towards Gaze-Based Prediction of the Intent to Interact in Virtual Reality", in *ACM Symposium on Eye Tracking Research and Applications*, ser. ETRA '21 Short Papers, New York, NY, USA: Association for Computing Machinery, 2021. DOI: 10.1145/3448018.3458008.

[7] B. Chen, E. Zheng, and Q. Wang, "A Locomotion Intent Prediction System Based on Multi-Sensor Fusion", *Sensors*, vol. 14, no. 7, pp. 12 349–12 369, 2014. DOI: 10.3390/s140712349.

[8] F. Ragni, L. Archetti, A. Roby-Brami, C. Amici, and L. Saint-Bauzel, "Intention Prediction and Human Health Condition Detection in Reaching Tasks with Machine Learning Techniques", *Sensors*, vol. 21, no. 16, 2021. DOI: 10.3390/s21165253.

[9] J. E. Jeong and Y. S. Choi, "Depth-enhanced gaze following method", in *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, ser. SAC '21, New York, NY, USA: Association for Computing Machinery, 2021, pp. 1090–1093. DOI: 10.1145/3412841.3442107.

[10] M. Grandini, E. Bagli, and G. Visani, *Metrics for multi-class classification: An overview*, 2020. DOI: doi.org/10.48550/arXiv.2008.05756.

[11] A. Agarwal and C. Arora, "Attention attention everywhere: Monocular depth prediction with skip attention", 2022. DOI: 10.48550/arXiv.2210.09071.

[12] *Papers with Code - KITTI Eigen split Benchmark (Monocular Depth Estimation).* [Online]. Available: paperswithcode.com/sota/monocular-depth-estimation-on-kitti-eigen.

[13] *Papers with Code - NYU-Depth V2 Benchmark (Monocular Depth Estimation).* [Online]. Available: paperswithcode.com/sota/monocular-depth-estimation-on-nyu-depth-v2.

[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, *End-to-end object detection with transformers*, May 2020. DOI: doi.org/10.48550/arXiv.2005.12872.

[15] H. Guo, Z. Hu, and J. Liu, *Mgtr: End-to-end mutual gaze detection with transformer*, Oct. 2022. [Online]. Available: arxiv.org/abs/2209.10930.

[16] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A survey on performance metrics for object-detection algorithms", in *2020 International Conference on Systems, Signals and Image Processing, IWSSIP 2020, Niterói, Brazil, July 1-3, 2020*, IEEE, 2020, pp. 237–242. DOI: 10.1109/IWSSIP48289.2020.9145130.

[17] B. Wang, T. Hu, B. Li, X. Chen, and Z. Zhang, "Gatector: A unified framework for gaze object prediction", *CoRR*, vol. abs/2112.03549, 2021. [Online]. Available: https://arxiv.org/abs/2112.03549.

[18] H. Tomas *et al.*, "Goo: A dataset for gaze object prediction in retail environments", in *CVPR Workshops (CVPRW)*, 2021, published.

[19] M. Monfort, A. Liu, and B. Ziebart, "Intent prediction and trajectory forecasting via predictive inverse linear-quadratic regulation", in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, 2015.

[20] M. Pei, Y. Jia, and S.-C. Zhu, "Parsing video events with goal inference and intent prediction", in *2011 International Conference on Computer Vision*, IEEE, 2011, pp. 487–494.

[21] F. Koochaki and L. Najafizadeh, "Predicting Intention Through Eye Gaze Patterns", in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Oct. 2018, pp. 1–4. DOI: 10.1109/BIOCAS.2018.8584665.

[22] R. Singh, T. Miller, J. Newn, E. Velloso, F. Vetere, and L. Sonenberg, "Combining gaze and ai planning for online human intention recognition", *Artificial Intelligence*, vol. 284, p. 103 275, 2020.