HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 4 (2023)

Understanding Dog Behavior Through Visual and Auditory Sensing Using Machine Learning

Amy Lin[1] and Mark Eastburn[#]

[1]Princeton High School, USA
[#]Advisor

ABSTRACT

This work aims to understand a dog's behavior towards environmental stimuli. Different from previous works, we collect multi-modality data including both video and audio data observed from the dog's egocentric perspective. We propose to model the association between a dog's reaction and the visual and auditory stimuli perceived by the dog using machine learning, in particular through an extended Convolutional Neural Network (eCNN). The eCNN model takes colored images, Short Time Fourier Transform (STFT) of audio, and motion fields extracted from image sequences as input, and outputs a prediction of the dog's reaction, classified as *Sit*, *Stand*, *Walk*, or *Smell*. Our proposed model achieves promising prediction results, with an average accuracy of 79.02% over all four classes. We also evaluate model performance by separately using one of image, audio, and motion information. Our results show that the dog responds strongly to low-frequency sounds and various color differences in its field of view. These research findings provide valuable insights to understanding animal behavior and intelligence as well as insights for building robotic companion dogs.

Introduction

Artificial intelligence (AI) has undergone rapid development in the past decade and revolutionized the way people live. Since current AI technologies rely on large amounts of data, they are prone to mistakes that may seem trivial to humans. For example, self-driving cars may misrecognize obstacles on the road, leading to fatal accidents. In contrast, animals such as dogs can easily avoid unfamiliar objects in their path. A better understanding of natural intelligence is critical to advancing the next generation of AI. In this work, we are interested in studying how a dog behaves and reacts to different stimuli in its surroundings. Understanding dog behavior could provide useful insights in understanding animal intelligence as well as human intelligence.

Like humans, dogs have five senses: sight, hearing, smell, touch, and taste. Their sense of hearing and sense of smell are much more powerful than humans. They have a wider angle of vision than humans, but cannot always see objects in focus. Studies have shown that domestic dogs perform similarly to human infants in cooperative communicative tasks (MacLean et al. 2017). They are capable of expressing emotions such as happiness, anger, fear, and jealousy, and experiencing depression and anxiety like humans (Marks et al. 2022). Understanding dog behavior can help maintain a dog's health and well-being. In addition, it can help create effective ways of training dogs for various services, especially when it is done from a dog's perspective rather than a human trainer's perspective.

Previous work on dog behavior has been reported in literature. Gregory Berns et al. (2012) studied dogs' brain activity in response to human hand signals using fMRI scans. They showed that dogs tend to pay close attention to human signals and display significant brain activity when seeing familiar hand signals for rewards. Several research groups used cameras and wearable devices to recognize a dog's posture and movements for health monitoring (Mealin et al. 2016, Kim & Moon 2022, Hussain et al. 2022, Atif et al. 2023). Robinson et al. (2015) studied the behavior of companion dogs in emergency situations for the purpose of

developing assistive technology. Other studies used videos and pictures of dogs' full body postures and facial expressions to predict their emotional states (Boneh-Shitrit et al. 2022, Ferres et al. 2022). Most existing work used a monitoring setup from a human perspective. In contrast, a recent study by Ehsani et al. (2018) used an egocentric setup. They analyzed how a dog acts and plans her movements in response to visual information perceived by the dog. Their machine learning model was able to predict walkable paths in the dog's environment, as a first step towards building robotic companion dogs.

In this work, we use an egocentric setup to study a dog's reaction to multi-modality stimuli, including visual and auditory stimuli, in its environment. Unlike most studies that capture images or videos of the dog from a human perspective, our setup captures video and audio stimuli in the environment from the dog's perspective. In addition, we expand the work of Ehsani et al. (2018) and introduce auditory stimuli into our study. We collect video and audio data from a dog's egocentric view and create a database of egocentric visual and audio stimuli that represents what a dog sees and hears. We propose an extended Convolutional Neural Network (eCNN) model to learn the association between the dog's reaction and the visual and auditory stimuli jointly perceived by the dog.

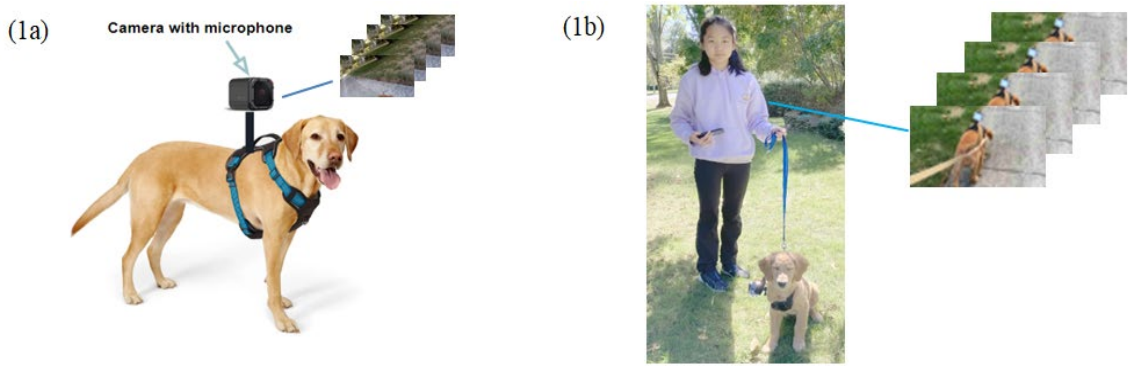

Figure 1. Data collection. In (1a), a camera with a microphone is mounted on the dog harness to capture video and audio from the dog's perspective. In (1b), a hand-held camera captures videos of dog movements from the human's view.

Methodology

We design a data collection system with one hand-held camera capturing the dog's movements, and a camera with a microphone mounted on the dog's harness capturing video and audio data the dog senses from the environment. Video and audio data are jointly analyzed by a machine learning model to learn the association between the dog's reaction and the visual and auditory stimuli sensed by the dog.

Data Collection

Data collection focuses on collecting the visual and audio stimuli in the dog's environment. A family pet, a Golden Retriever, is brought to different environments including parks and neighborhood streets. As Figure 1 shows, a GoPro camera is attached to the harness that the dog wears to capture what the dog sees and hears (Fig. 1(a)). The video stream captured by this GoPro camera is referred to as the dog-view video. At the same time, a hand-held camera is used by a person to record the dog's movements and its surroundings. The video stream captured by the hand-held camera is referred to as the human-view video. When recording the videos, both cameras are turned on at roughly the same time. An audible start signal, e.g. "take one", is used to synchronize the recordings made by the two cameras. A total of 22 pairs of videos have been collected.

Data Preparation

Three types of data, image frames, audio signals, and dog actions, are extracted from the dog-view and human-view videos. Image information is extracted from the dog-view videos. When the dog is walking or running, the dog-view camera may record extra noise from the leash. Therefore, we use audio signals extracted from the hand-held camera for its better audio quality to replace audio signals from the dog-view camera. Dog actions are manually labeled using the images from the human-view videos.

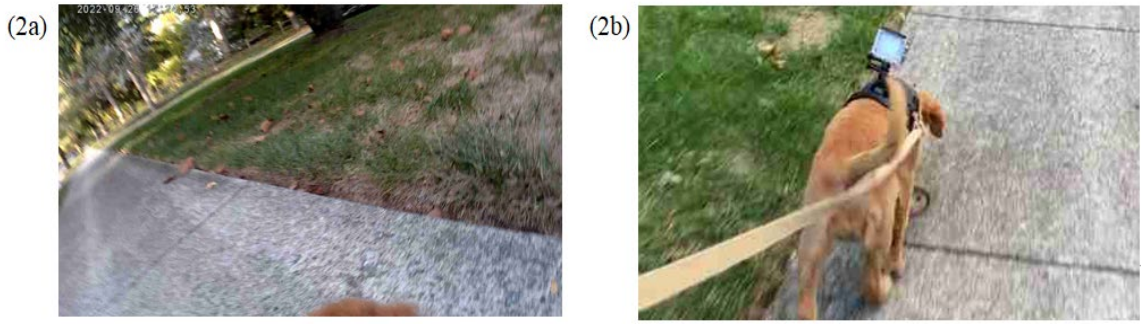

Figure 2. Audio signals, dog view images, and human view images are aligned by time.

To establish the correspondence between the dog's actions and its visual and audio stimuli, we establish the time correspondence between the image frames from the dog's view and the image frames from the human's view. Assume FPS_D and FPS_H are the frame rates (i.e. frames per second) of the dog-view camera and the human-view camera respectively. Assume T_{D0} and T_{H0} are the *start times* when the start signals of the dog-view camera and the human-view camera respectively. Given the dog-view image frame fr_D, the corresponding human-view image frame fr_H is

$$fr_H = \left(\frac{fr_D}{FPS_D} - T_{D0} + T_{H0}\right) \cdot FPS_H$$

Once the correspondence between the dog-view and human-view image frames are established, the correspondence between the dog's actions and the visual and audio stimuli is known.

As Figure 3 shows, there are 4 types of actions defined: *Sit, Stand, Walk,* and *Smell*. When the dog starts an action, the dog tends to continue the action for a period of time. Therefore, the image frames are only labeled when the dog changes its actions, and the same action is assigned to the subsequent image frames until a new action is presented.

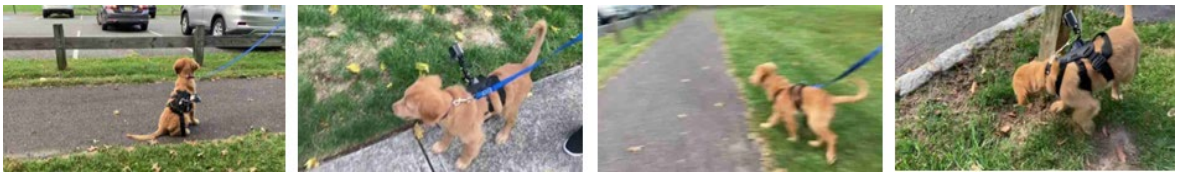

Figure 3. Images showing 4 actions of the dog: *Sit, Stand, Walk,* and *Smell*.

Data Analysis

Given the visual and audio stimuli a dog senses and its corresponding actions, we want to understand how the dog reacts to what it sees and hears. For example, what makes a sitting dog start walking? We propose an extended Convolutional Neural Network (eCNN) to learn the association between visual and audio stimuli and the corresponding dog actions. Compared to the CNN model, which only takes images as input variables, the eCNN model is able to explore data from multiple modalities, including images, motion, and audio. Our problem is formulated as a multi-class classification problem, where we use image, audio, and motion information to classify the dog's action into one of 4 classes: *Sit, Stand, Walk,* and *Smell*, shown in Figure 3.

Input Features

Image frames of size 672x378 pixels are extracted from the dog-view videos to capture what the dog sees. Each image is composed of three color channels: red, green, and blue. For faster computation, all image frames are resized to 151x85 pixels without losing useful visual information. The frame rate of the dog-view camera is 30.05 frames per second (FPS) and consecutive image frames are highly similar. Therefore, we perform down sampling over time and select one in every 5 frames to include in the dataset.

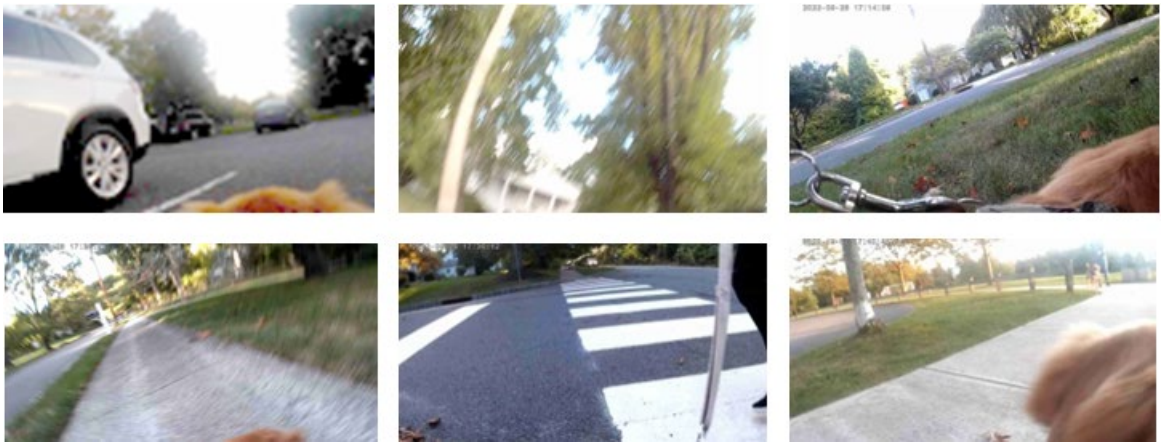

Figure 4. Sample images captured by the dog-view camera in the dataset.

The audio signals extracted from the human-view videos capture what the dog hears. We want to analyze the frequency content of the audio signal. Short-Time Fourier Transform (STFT) is used to decompose the audio signal into individual frequency components. STFT is defined as

$$S(m,\omega) = \sum_{n=-\infty}^{\infty} x(n)w(n-mH)e^{-i\omega n}$$

$x(n)$ represents the audio signal at time n, m is the index of the moving window, H is the hop length, $w(n)$ is the windowing function, and ω is the frequency. Since the STFT $S(m,\omega)$ is a complex function, we take the magnitude $|S(m,\omega)|$ as the input feature to the eCNN model. Figure 4 shows an example spectrogram of an audio signal produced by STFT.

In our experiment, we use the hop length of 1/90th second and a window size of 1 second. Figure 5 shows the resulting spectrogram (right) of a recorded audio signal (left) produced by STFT.

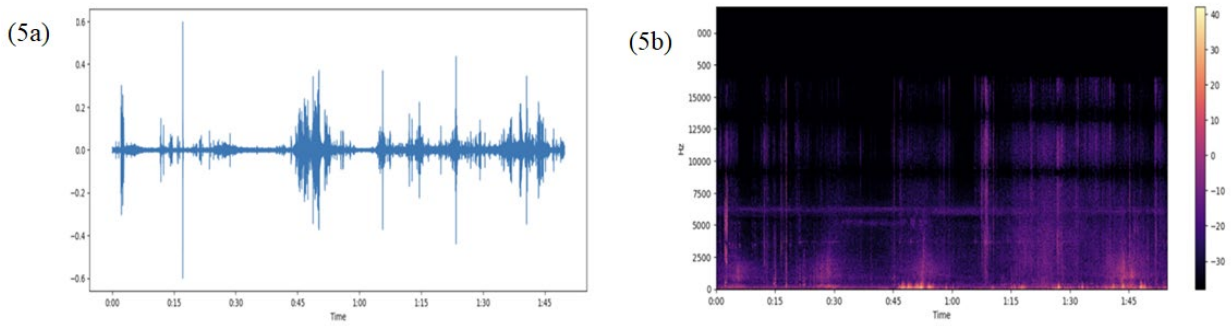

Figure 5. Spectrogram of audio signal produced by STFT. (Left: audio signal; Right: spectrogram)

To capture the sequential nature of the visual data, we included motion information as the third type of input feature. To compute image motion, we use the template matching method (Gonzalez & Woods 1992) to estimate the motion vectors of image blocks, i.e. how an image block moves from one frame to the next. Define I_t as the image frame at time t. As Figure 6 shows, for an image block centered around pixel (x_0, y_0), template matching finds the new block location $(x_0 + mx, y_0 + my)$ in the next image frame I_{t+1} such that the difference between the two image blocks measured by sum of squared error is minimized. A motion vector $mv(x_0, y_0) = (dx, dy)$ is found as

$$(dx, dy) = argmin_{\substack{kx=\{-D,\dots D\} \\ ky=(-D,\dots D\}}} \sum_{i,j} I_{t+1}(x_0 + i + kx, y_0 + j + ky) - I_t(x_0 + i, y_0 + j)$$

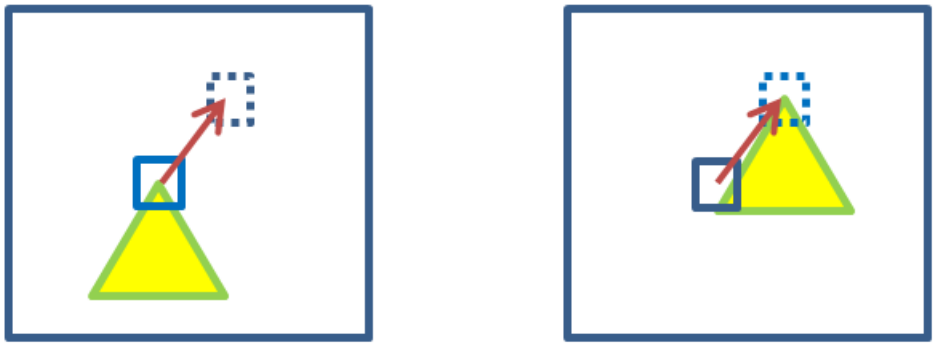

Figure 6. Motion vector of an image block.

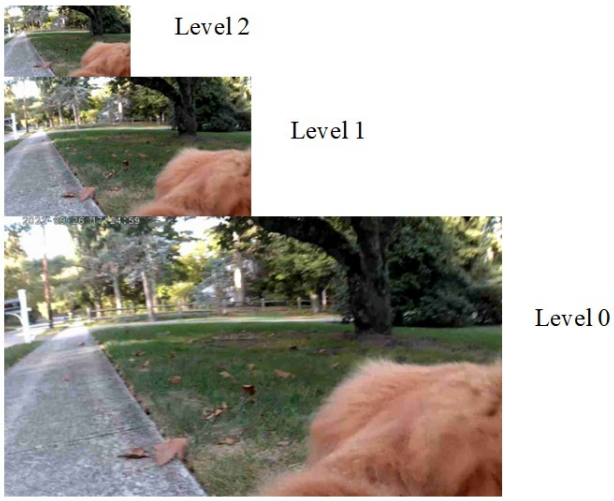

Figure 7. Image pyramid with 3 levels.

To make the method more computationally efficient, we first construct image pyramids, sequences of resized images at different resolutions $\{I_t^l: l = 0, 1, \dots, L-1\}$ (Figure 7) with a scaling factor 2. Template matching is first performed on the lowest resolution image. Assume at level l a motion vector (dx^l, dy^l) is found for an image block centered around pixel (x_0, y_0). Then the motion vector at level $l-1$, (dx^{l-1}, dy^{l-1}), at pixel location $(2x_0, 2y_0)$ is found as

$$(dx^{l-1}, dy^{l-1}) = argmin_{\substack{kx=\{2dx^l-D,\dots,2dx^l+D\} \\ ky=\{2dy^l-D,\dots,2dy^l+D\}}} \sum_{i,j} I_{t+1}(2x_0+i+kx, 2y_0+j+ky) - I_t(2x_0+i, 2y_0+j)$$

In general, image motion can be caused by both camera movements and objects moving in the scene. Since we are interested in object motion, we want to remove the motion caused by camera movements. We use the global average of motion vectors to represent the motion caused by camera movements and subtract the global motion from the motion vectors:

$$\overline{mv}(x,y) = \left(dx(x,y) - dx_g, dx(x,y) - dy_g\right)$$
$$dx_g = \frac{1}{N}\sum_{x,y} dx(x,y), \quad dy_g = \frac{1}{N}\sum_{x,y} dy(x,y)$$

In our experiment, we construct image pyramids with 4 levels. Image blocks of 5x5 pixels are used in template matching. For computational efficiency, centers of image blocks are 32 pixels apart in the original resolution. A resulting motion field is shown in Figure 8.

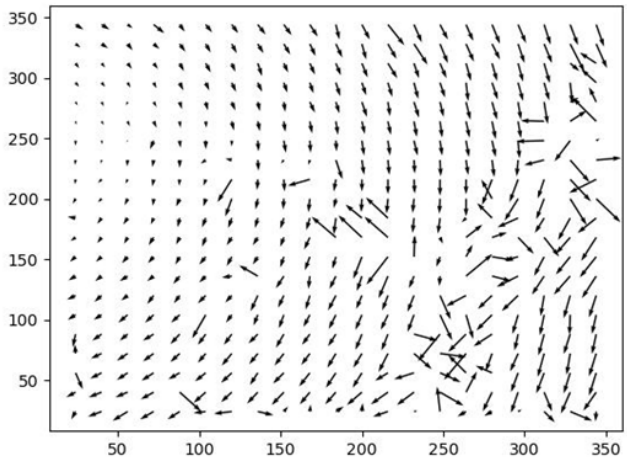

Figure 8. Motion field.

Extended Convolutional Neural Network (eCNN)

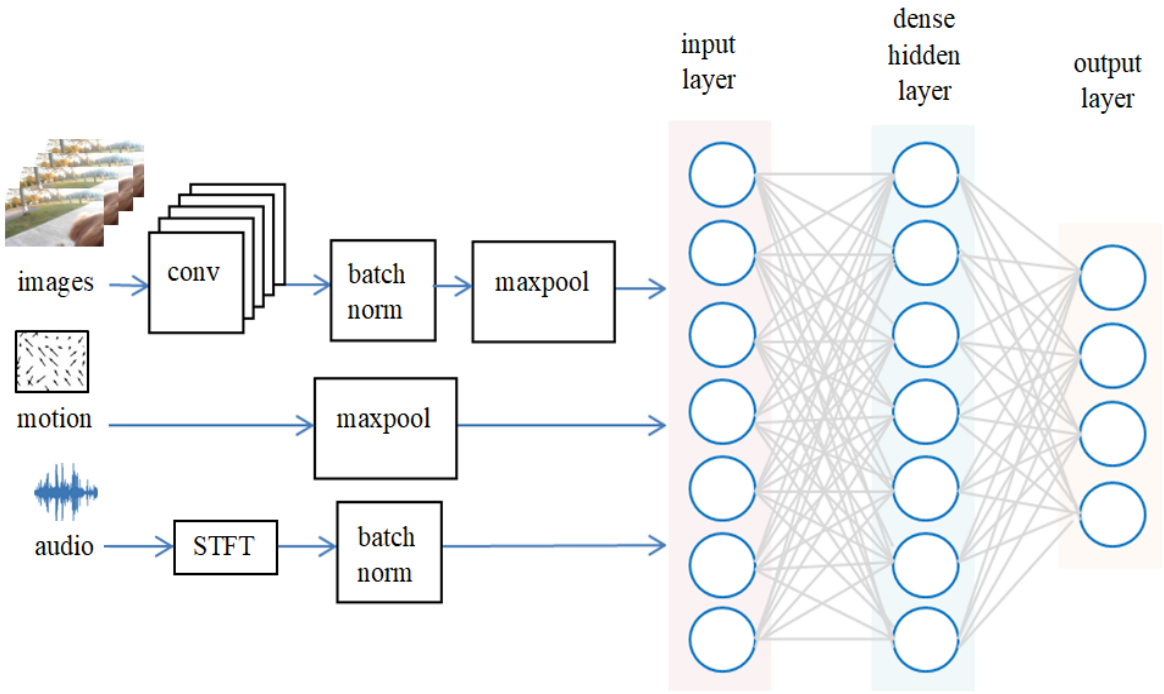

Figure 9. Extended Convolutional Neural Network (eCNN) model.

To utilize images, motion, and audio, we propose an extended CNN (eCNN) model to take multi-modality data as input, as shown in Figure 9. In the eCNN model, the first input type is an image. The image goes through a convolution layer, which is composed of 32 filters with kernel size 7x7 and stride 1. Batch normalization is then applied to the convolved image over the color channels, and a max pooling operation with a pool size of 3x3 is performed. The resulting output is flattened and fed into the input layer of the eCNN model. The second input type is a motion field. The magnitudes of the motion vectors go through a max pooling layer with a pool size of 9x3 before they are fed into the input layer. The third input type is an audio signal. First we apply STFT

to the audio signal to get a vector of the frequency domain representation. Batch normalization is performed on its magnitude spectrum and the resulting output is fed into the input layer. The input layer feeds the three types of input data into a dense hidden layer with 30 nodes and a sigmoid activation function. The dense layer is connected to an output layer of 4 nodes with a softmax activation function, corresponding to the one-hot encoding of each of the 4 classes: *Sit, Stand, Walk,* and *Smell*. TensorFlow Keras is used to implement the eCNN. The Adam algorithm is selected for optimization. We use a dropout of 20% on the input and output of the dense layer to reduce overfitting.

Train and Test

The dataset of images, motion, audio, and ground truth labels are randomly split into training, validation, and testing sets. 70% of the data which includes 3505 samples are used for training. 10% of the data which includes 458 samples are used for validation. The remaining 20% of the data which includes 954 samples are used for testing. When training the eCNN model, to prevent the optimization algorithm from getting stuck in local minima, we adopt mini-batches at two levels. First, the training samples are divided into a number of batches, referred to as "hyper-batches". Second, when each hyper-batch is used for training, mini-batches within the hyper-batch are used for optimization in TensorFlow. Similarly, training is run over epochs at two levels. First, when each hyper-batch is used for training, TensorFlow runs optimization over multiple epochs. Second, we run training on all hyper-batches over multiple hyper-epochs. Model performance over the number of hyper-batches and hyper-epochs is evaluated.

To evaluate the model's performance, we computed the overall prediction accuracy as well as a confusion matrix to assess the prediction accuracy in each class. We use $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ to represent the set of data samples, where x_i is the input feature and y_i is the ground truth class label, i.e. $y_i \in \{sit, stand, walk, smell\}$. We use $\hat{y}_i$ to represent the predicted class label for input x_i. The accuracy of class C ($C = sit, stand, walk, smell$) is computed as

$$accuracy(C) = \frac{|\{i: y_i = C, \hat{y}_i = C\}|}{|\{i: y_i = C\}|}$$

|S| represents the number of samples (cardinality) of set S. The overall accuracy is defined as the average accuracy across all classes.

$$accuracy = \frac{1}{4}\big(accuracy(sit) + accuracy(stand) + accuracy(walk) + accuracy(smell)\big)$$

A confusion matrix is defined as a two dimensional matrix$[M_{r,c}]$, where the rows r ($r \in \{sit, stand, walk, smell\}$) represent the ground truth class labels and the columns c ($c \in \{sit, stand, walk, smell\}$) represent the predicted class labels. A value in row r and column c represents the number of data samples that have a ground truth class r and were predicted as class c.

$$M_{r,c} = |\{i: y_i = r, \hat{y}_i = c\}|$$

Results and Discussion

We have trained the eCNN model on a training set of 3505 samples and a validation set of 458 samples. The model was tested on a testing set of 954 samples. The training set includes 338 samples labeled as *Sit*, 521 samples labeled as *Stand*, 1574 samples labeled as *Walk*, and 1072 samples labeled as *Smell*. The validation set includes 55 samples labeled as *Sit*, 65 samples labeled as *Stand*, 207 samples labeled as *Walk*, and 131 samples

labeled as *Smell*. The testing set includes 102 samples labeled as *Sit*, 149 samples labeled as *Stand*, 427 samples labeled as *Walk*, and 276 samples labeled as *Smell*. The experiments were run with 1 hyper-batch and 4 hyper-batches over 40 hyper-epochs. The best overall accuracy on the validation set was used to select the best number of hyper-epochs and number of hyper-batches. Training was run on a Windows machine with 24.0 GB RAM and 2.3 GHz AMD Ryzen 5 processor. One round of training took approximately 7 hours.

Table 1 shows the performance of the model over 40 hyper-epochs and 4 hyper-batches selected by validation. On the validation dataset, the model achieves an overall accuracy of 79.47%. It correctly predicts 88.00% of samples labeled as *Sit*, 72.73% of samples labeled as *Stand*, 78.95% of samples labeled as *Walk*, and 78.20% of samples labeled as *Smell*. On the testing dataset, the model achieves an overall accuracy of 79.02%. It correctly predicts 84.21% of samples labeled as *Sit*, 78.87% of samples labeled as *Stand*, 78.66% of samples labeled as *Walk*, and 74.33% of samples labeled as *Smell*. The performance of the eCNN model on the validation and testing sets is very similar, showing no obvious overfitting. The model achieves the highest accuracy for class *Sit*, but performance is relatively similar over all classes.

The confusion matrix on the testing set is as follows:

Prediction / Ground truth	sit	stand	walk	smell
sit	80	5	9	1
stand	3	112	23	4
walk	15	26	328	48
smell	4	6	67	223

Figure 10. Confusion matrix on testing set.

Table 1. Experimental results.

	Training	**Validation**	**Testing**
Number of Samples	3505	458	954
Number of Samples: class *Sit*	338	55	102
Number of Samples: class *Stand*	521	65	149
Number of Samples: class *Walk*	1574	207	427
Number of Samples: class *Smell*	1072	131	276
Overall Accuracy	94.34%	79.47%	79.02%
Accuracy of class *Sit*	99.11%	88.00%	84.21%
Accuracy of class *Stand*	95.59%	72.73%	78.87%
Accuracy of class *Walk*	96.19%	78.95%	78.66%
Accuracy of class *Smell*	86.47%	78.20%	74.33%

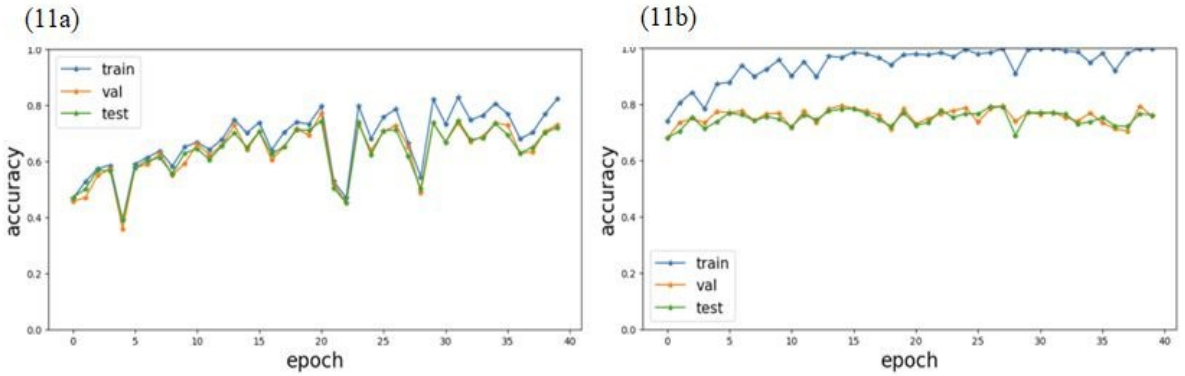

Figure 11. Overall performance of model over different hyper-epochs: (11a) 1 batch used; (11b) 4 batches used. Blue: training; Orange: validation; Green: test

To evaluate the effects of multiple hyper-epochs and hyper-batches on the performance of the model, we plot the overall accuracy over different numbers of hyper-epochs and hyper-batches in Figure 10. With 1 batch, the training reaches optimal performance at around 20 hyper-epochs. With 4 batches, the training reaches optimal performance much earlier at around 5 hyper-epochs. In both tests, the model's performance on validation and testing sets are very similar, suggesting no obvious overfitting. However, with 4 batches, there is a relatively large difference between the training performance and the validation performance. This is likely caused by the smaller number of data samples in each batch fed into the training algorithm. The performance of each class over different hyper-epochs is shown in Figure 11. Although each class has a different number of training samples, the model's performance is consistent across all classes.

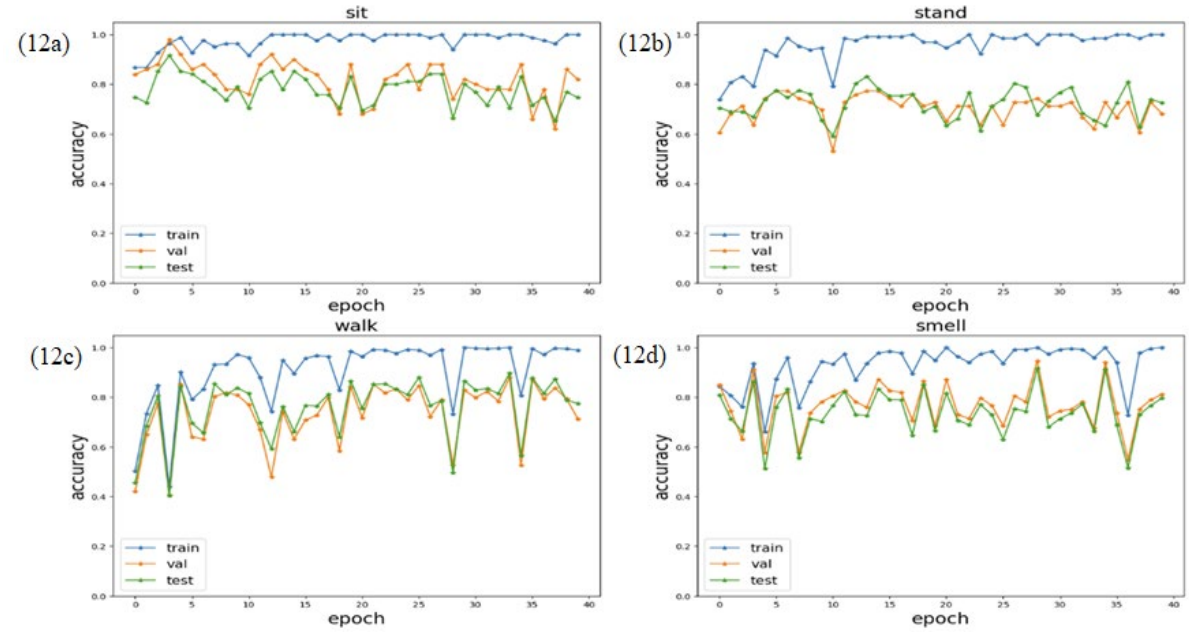

Figure 12. Performance on each class over different hyper-epochs. (12a) *Sit*; (12b) Stand; (12c) *Walk*; (12d) *Smell*

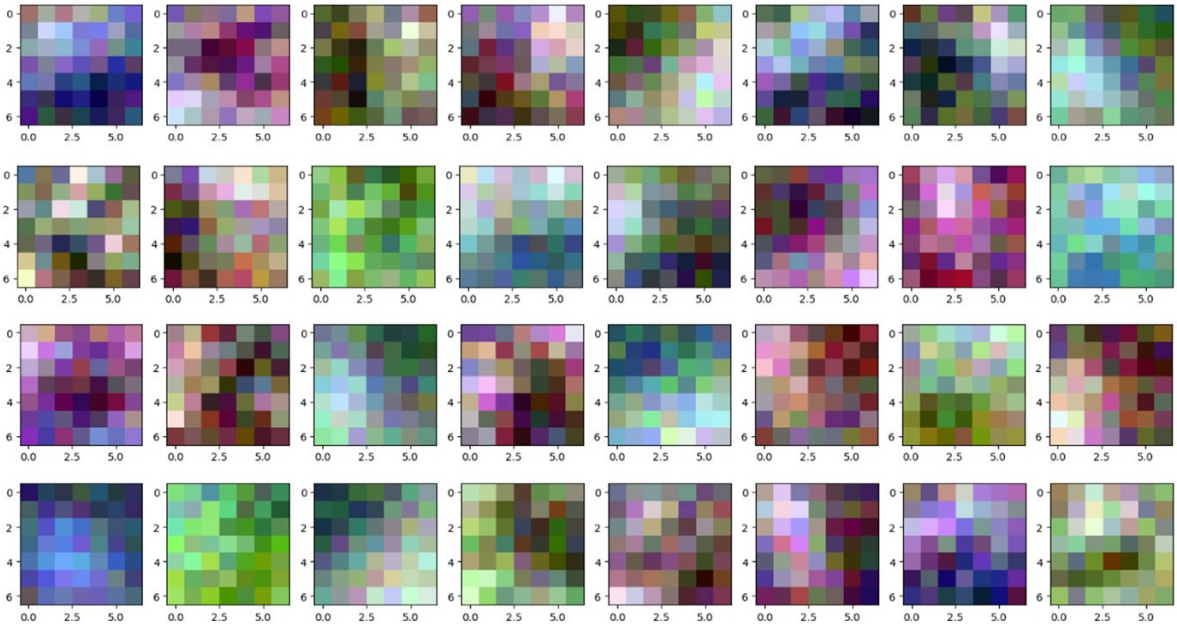

Figure 13. 32 convolutional filters with size 7x7 learned by the eCNN model.

Figure 12 shows the 32 convolution filters of size 7x7 that were learned by the model. They represent various color patterns. Many filters have a color difference along a diagonal, which suggests that the dog pays more attention and reacts to color differences in its field of view.

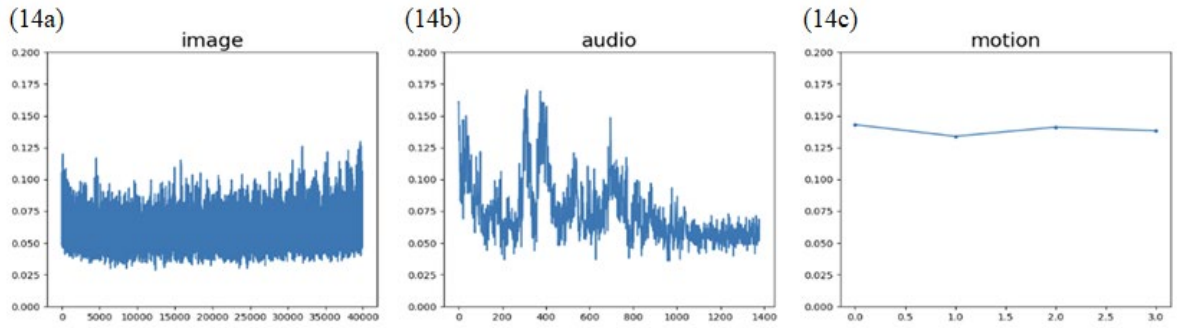

Figure 14. Average weights on image (14a), audio (14b), and motion (14c) features in dense layer.

To understand the role each sensing modality plays in predicting the dog's actions, we first display the weights on the input nodes learned by the model which correspond to image, audio, and motion features respectively. For each input node corresponding to image features, we average the magnitude of the weights of the dense layer nodes connected to that input node. The average weight is calculated for all input notes corresponding to image features and is shown in Figure 13. Similarly, the average weights for audio and motion features are calculated and shown in Figure 13 as well. We observe that for the audio features, some low frequency features are weighted more, which suggests that the dog likely reacts more to the low frequency components in what it hears.

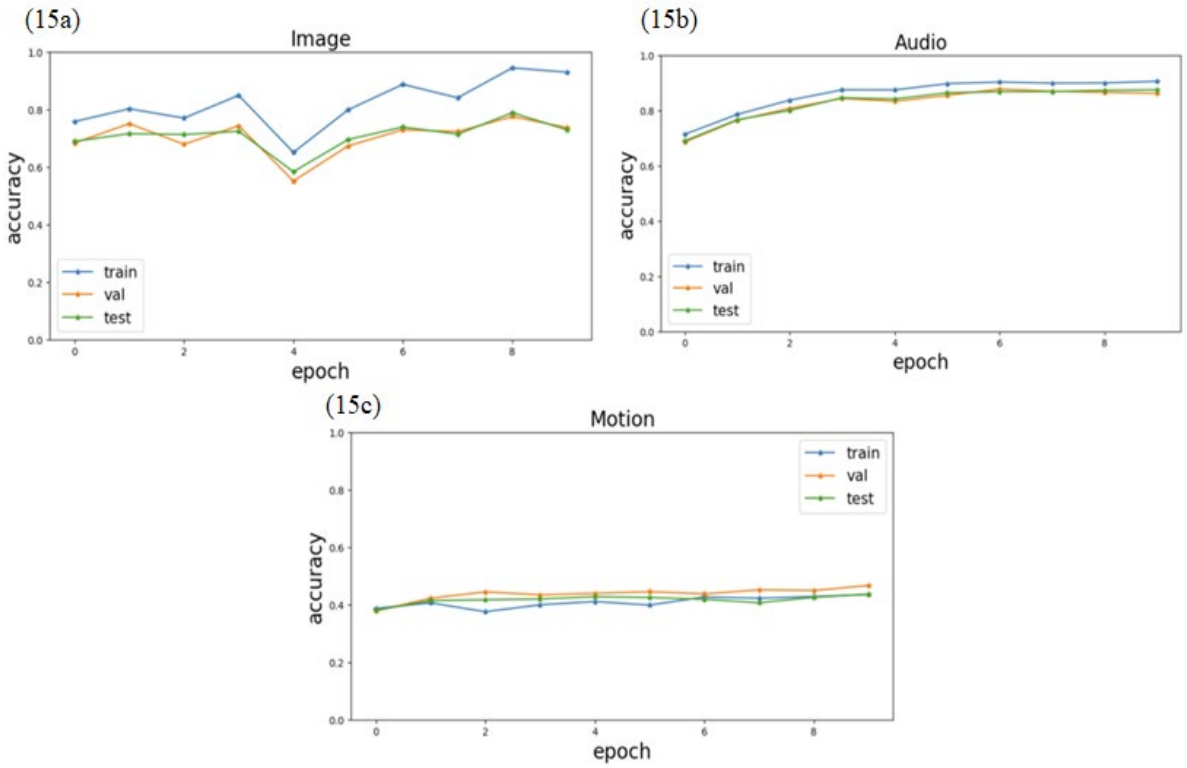

Figure 15. Overall performance of model using single-modal data. (15a): used only images; (15b): used only audio; (15c): used only motion

In addition, we have trained the model separately using only image, only audio, and only motion information. 10 hyper-epochs and 4 hyper-batches are used in the tests, shown in Figure 14. The optimal performance of the model trained with each type of single-modal input is shown in Table 2. The model achieves the highest overall accuracy when using only audio as input, suggesting that audio plays a significant role in the dog's behavior. Training using only motion information has a much lower performance than the multi-modality input, which is likely due to inadequate background motion correction.

Table 2. Optimal performance of model trained with only image, only audio, and only motion information.

	Image only	Audio only	Motion only
Overall Accuracy	78.81%	86.74%	43.66%

Conclusion

In this work, we proposed a research framework to understand dog behavior. In contrast with most studies which collect data from a human perspective, we collected video and audio data from a dog's egocentric view. To our knowledge, we are one of the first to incorporate audio data. Through machine learning, we learned the association between the dog's reaction and the visual and audio stimuli perceived by the dog. We proposed an extended Convolutional Neural Network (eCNN) to utilize multi-modality features of images, audio, and motion information. The model achieved promising results with an overall prediction accuracy of 79.02%. We

observed that the dog reacts strongly to various color patterns and color contrasts in its field of view. It also reacts more to some low frequency components in what it hears. These findings can offer useful information when designing effective ways to train dogs for various services, such as companionship and rescue work, as well as offering valuable insights in understanding animal intelligence.

Acknowledgments

I would like to thank Mr. Mark Eastburn and Dr. Lei Zhang for providing valuable feedback to this work.

References

Ehsani, K., Bagherinezhad H., Redmon, J., Mottaghi, R., & Farhadi, A. (2018). Who let the dogs out? Modeling dog behavior from visual data, *CVPR*, pp. 4051-4060. https://doi.org/10.48550/arXiv.1803.10827

Berns, G. S., Brooks, A. M., & Spivak, M. (2012). Functional MRI in Awake Unrestrained Dogs. *PLoS ONE* 7(5): e38027. https://doi.org/10.1371/journal.pone.0038027

Stromberg, J. (2016). Why scientists believe dogs are smarter than we give them credit for, *Vox*.

Agrawal, P., Carreira, J., & Malik, J. (2015). Learning to see by moving, *International Conference on Computer Vision*, pp. 37-45. https://doi.org/10.48550/arXiv.1505.01596

Fathi, A., Farhadi, A., & Rehg, J. M. (2011). Understanding egocentric activities, *International Conference on Computer Vision*, pp. 407-414. https://doi.org/10.1109/ICCV.2011.6126269

Lee, Y. J., Ghosh, J., & Grauman, K. (2012). Discovering important people and objects for egocentric video summarization, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1346-1353. https://doi.org/10.1109/CVPR.2012.6247820

Pintea, S. L., van Gemert, J. C., & Smeulders, A. W. M. (2014). Deja vu: Motion prediction in static images, *European Conference on Computer Vision*, pp 172–187. https://doi.org/10.48550/arXiv.1803.06951

Gonzalez, R., & Woods, R. (1992). Digital Image Processing, 3rd Edition, *Pearson Prentice Hall*, pp. 414-428.

Kim, J., & Moon, N. (2022). Dog Behavior Recognition Based on Multimodal Data from a Camera and Wearable Device, *Applied Sciences,* 12(6):3199. https://doi.org/10.3390/app12063199

Hussain, A., Ali, S., Abdullah, & Kim, H. -C. (2022). Activity Detection for the Wellbeing of Dogs Using Wearable Sensors Based on Deep Learning*, IEEE Access*, vol. 10, pp. 53153-53163. https://doi.org/10.1109/ACCESS.2022.3174813

Siwak, C. T., Murphey, H. L., Muggenburg, B. A., & Milgram, N. W. (2002). Age-dependent decline in locomotor activity in dogs is environment specific, *Physiology & Behav*ior, 75(1-2), pp. 65-70. https://doi.org/10.1016/s0031-9384(01)00632-1

Quaranta, A., Siniscalchi, M., & Vallortigara, G. (2007). Asymmetric tail-wagging response by dogs to different emotive stimuli, *Current Biology*, vol. 17, no. 6, pp. 199-201. https://doi.org/10.1016/j.cub.2007.02.008

Völter, C. J., Lonardo, L., Steinmann, M. G. G. M., Ramos, C. F., Gerwisch, K., Schranz, M. T., Dobernig, I., & Huber, L. (2023). Unwilling or unable? Using three-dimensional tracking to evaluate dogs' reactions to differing human intentions, *Proceedings of the Royal Society*, 290(1991). https://doi.org/10.1098/rspb.2022.1621

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324. https://doi.org/10.1109/5.726791

Martinez, J., Black, M. J., & Romero, J. (2017). On human motion prediction using recurrent neural networks, *IEEE Computer Vision and Pattern Recognition Conference*, pp. 2891-2900. https://doi.org/10.48550/arXiv.1705.02445

Liu, J., Shahroudy, A., Xu, D., & Wang, G. (2016). Spatio-temporal lstm with trust gates for 3d human action recognition, *European Conference on Computer Vision*, pp 816–833. https://doi.org/10.48550/arXiv.1607.07043

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to Sequence – Video to Text, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4534-4542. https://doi.org/10.48550/arXiv.1505.00487

Mealin, S., Domínguez, I. X., & Roberts, D. L. (2016). Semi-supervised classification of static canine postures using the microsoft kinect. Proceedings of the Third International Conference on Animal-Computer Interaction, pp. 1-4. https://doi.org/10.1145/2995257.3012024

Robinson, C., Mancini, C., van der Linden, J., Guest, C., & Swanson, L. (2015). Exploring assistive technology for assistance dog owners in emergency situations. Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments, pp. 1-2. https://doi.org/10.1145/2769493.2769576

Atif, O., Lee, J., Park, D., & Chung, Y. (2023). Behavior-Based Video Summarization System for Dog Health and Welfare Monitoring. Sensors, 23(6), 2892. https://doi.org/10.3390/s23062892

Boneh-Shitrit, T., Feighelstein, M., Bremhorst, A., Amir, S., Distelfeld, T., Dassa, Y., Yaroshetsky, S., Riemer, S., Shimshoni, I., Mills, D. S., & Zamansky, A. (2022). Explainable automated recognition of emotional states from canine facial expressions: the case of positive anticipation and frustration. Sci Rep 12, 22611. https://doi.org/10.1038/s41598-022-27079-w

Ferres, K., Schloesser, T., & Gloor, P.A. (2022). Predicting Dog Emotions Based on Posture Analysis Using DeepLabCut. Future Internet 14(4), 97. https://doi.org/10.3390/fi14040097

MacLean, E. L., Herrmann, E., Suchindran, S., & Hare, B. (2017). Individual differences in cooperative communicative skills are more similar between dogs and humans than chimpanzees. Animal Behaviour, 126, 41–51. https://doi.org/10.1016/j.anbehav.2017.01.005