10-2022

Making Robotic Dogs Detect Objects That Real Dogs Recognize Naturally: A Pilot Study

Dingtao Hu

Zhizun Wang

Benjamin M. Fung

David Meger

Rupendra Raavi

See next page for additional authors

Authors

Dingtao Hu, Zhizun Wang, Benjamin M. Fung, David Meger, Rupendra Raavi, Patrick C. Hung, Hidenori Mimura, and Kamen Kanev

18. Making Robotic Dogs Detect Objects That Real Dogs Recognize Naturally: A Pilot Study

Dingtao Hu, Zhizun Wang
Benjamin C. M. Fung, David Meger
McGill University, Canada
{dingtao.hu, zhizun.wang}@mail.mcgill.ca
{ben.fung, david.meger}@mcgill.ca

Rupendra Raavi, Patrick C. K. Hung
Ontario Tech University, Canada
rupendra.raavi@ontariotechu.net
patrick.hung@ontariotechu.ca

Hidenori Mimura, Kamen Kanev
Shizuoka University, Japan
mimura.hidenori@shizuoka.ac.jp
kanev@inf.shizuoka.ac.jp

Abstract

The recent advancements in artificial intelligence (AI) and deep learning have enabled smart products, such as smart toys and robotic dogs, to interact with humans more intelligently and express emotions. As a result, such products become intensively sensorized and integrate multi-modal interaction techniques to detect and infer emotions from spoken utterances, motions, pointing gestures and observed objects, and to plan their actions. However, even for the predictive purposes, a practical challenge for these smart products is that deep learning algorithms typically require high computing power, especially when applying a multimodal method. Moreover, the memory needs for deep learning models usually surpass the limit of many low-end mobile computing devices as their complexities boost up. In this study, we explore the application of lightweight deep neural networks, SqueezeDet model and Single Shot Multi-Box Detector (SSD) model with MobileNet as the backbone, to detect canine beloved objects. These lightweight models are expected to be integrated into a multi-modal emotional support robotics system designed for a smart robot dog. We also introduce our future research works in this direction.

Keywords: Robotic dogs, Smart toys, SqueezeDet, MobileNet, Object detection

1. Introduction

Demographic transition is regarded as one of the motivators of technological development in the twenty-first century. Millions of individuals are negatively affected by population ageing and the epidemic of loneliness, which have caused damage to their psychological and physiological health (Cacioppo et al., 2006; Luo et al., 2012; Anderson et al., 2018). While addressing these issues will take significant work, the industry has already developed smart robots as novel solutions. For example, a social robot, defined as an autonomous robot that communicates with humans by following the social rules attached to its role, could provide individuals with emotional support, make routine activities easier, link distant family members, and be employed in various professional jobs. The Japanese technology start-up Groove X developed a home robot named LOVOT (*GROOVE X* n.d.), which is advertised as a product that "stirs your instinct to love". Boosted by 50 surface sensors, the robot can remember the faces of its close partners, warm up when embraced, and move closer to the door when its owner arrives late. Sony also created a dog-shaped social robot named AIBO with a variety of sophisticated functions (Melson et al., 2005). The robot maps out the space ahead of it with a forward-facing camera and has Wi-Fi and Long Term Evolution (LTE) connectivity, allowing it to work inside

and outside the home. Four microphones detect speech instructions, while two Organic Light-emitting Diodes (OLED) panels act as its eyes.

Intelligent emotional-support robots involve multiple modalities as they need to react to the external environment and interact with individuals in collaborative activities. Huge volumes of sensory-motor data, such as raw RGB image frames, joint angles, and voice commands, are merged to generate higher-level multimodal representations based on deep learning algorithms. The method of integrating data from various input modalities into a compact multi-modal representation is referred to as multi-modal fusion. The efficient multi-modal fusion of data from different sensors helps the model learn important tasks and exhibit robustness against noise. For instance, a significant application of this fusion method lies in the area of Physical Human-robot Interaction (PHRI), where the modalities of force, torque and tactility are recognized as direct contact modalities (Xue et al., 2020). They need to be efficiently detected and integrated to conduct collision avoidance between humans and robots. Indirect contact modalities in PHRI include vision and natural language, which are crucial for modeling and inferring the space-time relationship between the perception and operation domains for interaction tasks.

To make a robot dog more intelligent, it is natural to apply multi-modal fusion to it. By combining different modalities and taking advantage of the complementary information in multi-modal data, the robot dog is expected to show behaviours similar to a real dog. If objects that catch the attention of a real dog are placed inside the field of view of the robot dog, it should also react to those objects. In this paper, we focus on the object-detection task in the emotional-support robotic system, to identify items such as dog feeders, dog cushion beds, chew toys, treat balls and human faces in the frames extracted from real-time videos. We will utilize the AIY Vision Kit from Google (*Vision kit* n.d.) as the camera to be installed on the robotic dog. Due to the constraints on the memory of the device, the major goal of this study is to elaborate on the accuracy of SqueezeDet (Wu et al., 2016) and SSD-based detection with MobileNet (Howard et al., 2017) as the backbone on the customized item dataset.

2. Literature Review

Nowadays, we see a vast improvement in the creation and usage of robotic dogs, such as Sony AIBO. For example, Bruno et al. (2019) developed a robotic dog that guides visually impaired people. They used technologies such as ultrasound sensors, vision sensing, etc., to help robotic dogs navigate the visually impaired. The robotic dog they developed initially detects the obstacles with the help of ultrasound. Once an obstacle is detected, the images are taken, and the object-detection algorithm You Only Look Once (YOLO) is used to detect the objects and then guide without colliding. Schellin et al. (2020) surveyed the dog likeness of Sony's AIBO, where they performed a study by considering two factors: putting fur on the robot and keeping it as it is. Thirty-three participants were recruited with 12 of them being females, and all the participants had previously owned a pet. The final result was that, except for three participants, all other participant liked the dogs. Robotic dogs are also used in therapy for treating loneliness. Banks et al. (2008) conducted surveys to check the loneliness of older adults in three scenarios: one with a robotic dog. another one with a real dog, and the last one with no dog. The results showed that elderly residents living with either a real or a robotic dog felt substantially less lonely than residents without a dog. Interestingly, it was also found that there was no significant difference between using a robotic dog and a real dog. Next, Jones & Deeming (2007) discussed how an emotional interaction of a dog would affect humans, making them more alive. They created software capable of differentiating users' emotions based on their speech properties. With such a software we, a robotic dog would not be able to obey users based on their words. Instead, it follows the user based on the emotion used to speak those words. This is done on the top Sony AIBO robotic

dog, which was able to differentiate anger, sadness, happiness, boredom, surprise and perform actions based on emotions. After the survey, most of the participants were enthusiastic about this robot. Stanton et al. (2008) conducted a study to see whether children would prefer to interact with the Sony robot AIBO or with a simple mechanical toy dog called Kasha. Where authors noted the time, of each child spent in interacting with kasha and AIBO and based on the results, children spent an average of around 72 percent with AIBO and more minor of 52 percent with Kasha. The results also show that the number of words that children spoke to the robotic dog was higher compared to spoken to the simple mechanical toy dog. In an earlier study, Melson et al. (2005) surveyed 72 children to see whether they would interact more with the robotic dog AIBO or with a living Australian shepherd. It turned out that children were more interested in interacting with the real dog than with the robotic dog. Still, the survey also showed that many children tried to interact with the robotic dog AIBO similarly to the way they treat a real dog. Further, Weiss et al. (2009) conducted a survey in a shopping mall for three consecutive days as a voluntary free exploration case study where 147 participants were selected. According to their study, all 129 children among the participants showed a lot of enthusiasm, expresses by statements such as ’That’s cool', ’May I play with it?', and actions to run towards AIBO to play with it. But researchers would also report children saying that AIBO was not responding properly, which as mentioned in the paper was due to the multiple voice commands issued simultaneously. As a result, most of the children, on average, spent around 20 minutes and interaction was stopped just because their parents wanted them to go.

3. Project Pipeline

The semantic perception of canine beloved objects involves a number of development phases and a customized training dataset. A collection of images of dog supplies from various online pet stores (Amazon, JD, Chewy, etc.) and royalty free stock photography providers (iStock, Getty Images, etc.) was established in this work. The images were annotated by Computer Vision Annotation Tool (CVAT) (Sekachev et al., 2019) in the PASCAL VOC format with the assistance trained deep learning models provided by the open-source tool. Furthermore, our pipeline can automate multiple image augmentation operations and the process of labeling augmented images, which expands the volume of the original dataset and enhance the effectiveness of annotation. The labels and the coordinates of bounding box in the augmented images were automatically generated by transformation processing based on those in the original images. Given these materials, lightweight neural network models were trained utilising the TensorFlow framework (Abadi et al., 2015). We used the model with better performance on image prediction in the preliminary experiment to conduct real time inference on the video captured by laptop webcam and will integrate it with into the AIY Vision Kit from Google (*Vision kit* n.d.) and robotic dog in the further study. Figure 1 demonstrates the pipeline of canine beloved object detection object.

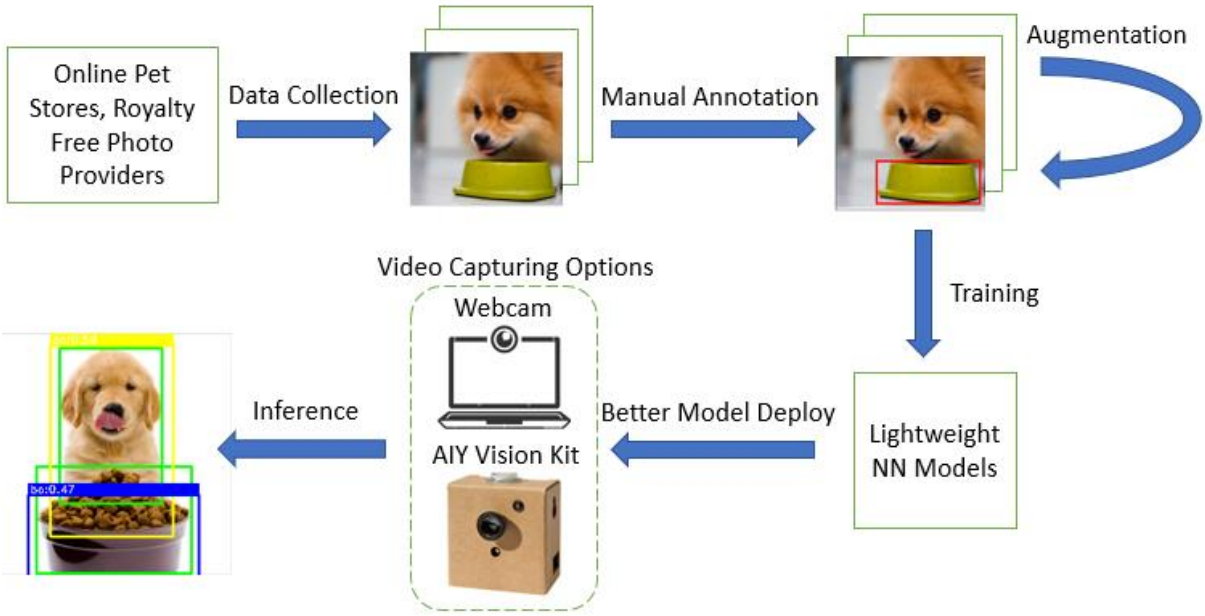


Figure 1: Canine beloved object detection framework. The process begins with data collection from online pet stores and royalty free photography providers, which is followed by the manual object annotation. To expand the dataset size, multiple data augmentation techniques are utilised. Furthermore, the lightweight neural network model with better performance in the preliminary experiment is integrated into the dedicated hardware devices.

4. Data Processing

We created our dataset from scratch. We collected 1400 images from online pet stores and galleries for three categories of canine beloved items: dog feeders, dog beds and dog toys. Eight labels were defined: dog, cat, human face, hand, treat ball, chew toy, dog bowl and dog bed. All images were resized to 256 × 256.

We applied the following five image data augmentation techniques to improve the performance and the generalization capabilities of the model:

- Flipping: each image is flipped horizontally;
- Gaussian Noise: the noise from the distribution $N(0, 0.1\times255)$ is added to images. Specifically, the noise value is different per pixel and channel (i.e. adding different noise values to red, green and blue channels of the same pixel) ;

Figure 2: An example of a training image to which we apply data augmentation or data annotation. Images (a)-(g) are the original image, the flipped image, the image with Gaussian Noise, the image with Sigmoid Contrast, the image with Linear Contrast, the image with Channel Shuffle, and the image with bounding boxes, respectively.

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

(e) **(f)** **(g)**

- Sigmoid Contrast: we adjusted the image contrast by scaling pixel values to the size of $255 \times \frac{1}{1+e^{gain \times (cutoff - \frac{v}{255})}}$, where *v* is the original single pixel value, the *gain* is uniformly sampled from the interval [5, 20], and *cutoff* is uniformly sampled from the interval [0.25, 0.75];
- Linear Contrast: we modified the image contrast by scaling pixel values to $127+\alpha \times (v-127)$, where *v* is the original single pixel value, and α is uniformly sampled from the interval [0.4, 1.6] for each image;
- Channel Shuffle: we rearranged the RGB channels of each of the images at random.

We annotated the objects in the original images using Computer Vision Annotation Tool (CVAT), an open-source interactive video and image annotation tool for computer vision research developed by Intel (Sekachev et al., 2019). CVAT supports a list of shapes with which we can annotate the images, such as bounding boxes and polygons. We conducted data annotation by looking at an image in our dataset, finding the objects that belong to one of the 8 labels we defined above, and manually annotating them with bounding boxes. CVAT allows us to export the annotated data in different formats, including PASCAL VOC, TFRecord and KITTI, which we can directly feed into our models. CVAT also provides trained deep neural network models, such as Mask R-CNN and YOLOv3, to assist the users to improve the annotation efficiency.

5. Lightweight Neural Networks

5.1 SqueezeDet

SqueezeDet is one of the smallest lightweight, fully Convolutional Neural Networks (CNNs) for object detection, which integrates SqueezeNet as the backbone network for feature extraction into the YOLO framework. It has been verified that SqueezeDet could achieve good performance in single-shot object detection tasks on many benchmark datasets (including KITTI, VOC 2007, COCO) with significantly smaller model size than other networks. In particular, it is valid to employ SqueezeDet in low-end edge devices or embedded systems within a short running time.

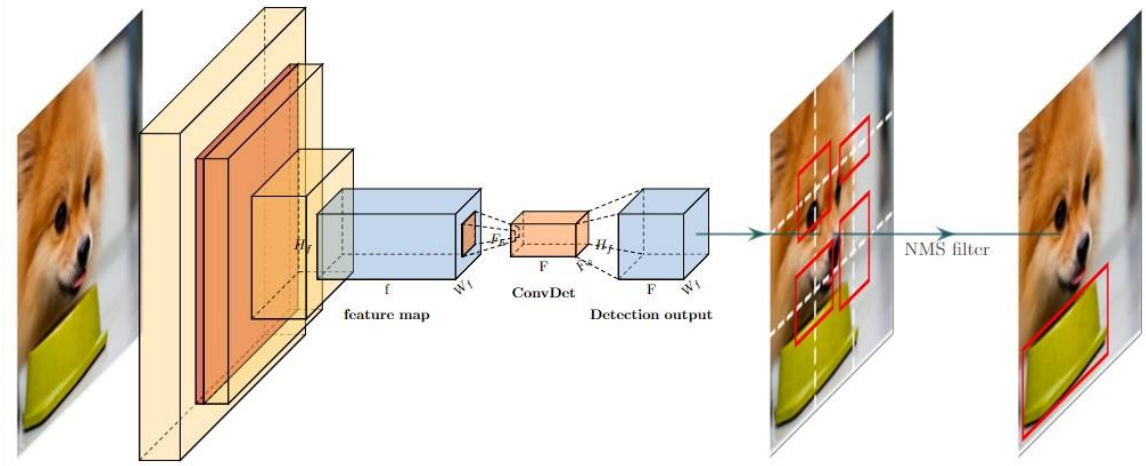


Figure 3: SqueezeDet pipeline. The feature map is of size W_f,H_f,f, where W_f and H_f refer to the width and height of the feature map, respectively, and f represents the number of channels. The output of the ConvDet, a $F_w \times F_h$ convolution, is of size $F := K \times (5 + C)$. It is then trained to compute the class probability and the values associated with the bounding box at each grid center.

Inspired by the Region Proposal Network in Faster Region-based Convolutional Neural Network (RCNN) and YOLO, SqueezeDet employs a single-shot object detection pipeline to achieve region proposition and classification by one network stream concomitantly. First, the convolutional neural network takes the input image and extracts feature maps, which act as the backbone of the object localization network. Then, the ConvDet layer takes the feature maps as inputs, overlays a uniformly distributed $W \times H$ spatial grid on top of them, where W and H represent, respectively, the number of grid centres horizontally and vertically, goes across each spatial position as a sliding window and computes K reference bounding boxes (termed as anchors) with preset shapes at each grid center. Each anchor is represented as 4 scalars, (x_i,y_j,w_k,h_k), where (x_i,y_j) is the spatial coordinate of the grid center (i,j), and w_k and h_k refer to the width and height of the k-th reference bounding box. Besides, each anchor is also associated with C conditional classes and a confidence score Pr(Object)*IOU, where IOU denotes Intersection of Union. A high confidence score indicates that the probability of the existence of a targeted object is high and the intersection between the candidate anchor and the corresponding ground truth is large. As a result, SqueezeDet's fixed output is $W \times H \times K(4+1+C)$. Finally, the most appropriate bounding box for the object is obtained by filtering out redundant bounding boxes by Non-Maximum Suppression (NMS).

5.2 MobileNet-SSD

MobileNet, also referred to as MobileNetV1, is a type of CNN specifically designed for embedded vision applications and mobile devices (Howard et al., 2017). Using depthwise separable convolutions, it builds lightweight deep neural networks on top of a streamlined architecture. Depthwise separable convolution is a form of factorized convolution. In other words, it factorizes a standard convolution into two separable convolutions: a depthwise convolution, which applies a single filter per input channel, and a pointwise convolution, which is a 1× 1 convolution combining the outputs of the depthwise convolution and changing the dimension. Two non-linearities, Batch Normalization (BN) and Rectified Linear Unit (ReLU), are applied to the depthwise separable convolutions. In the complete network structure of MobileNet, there are 28 layers in total, if the depthwise convolutions and the pointwise convolutions are counted as separate layers. Each layer is followed by both BN and ReLU, with the exception of the final fully connected layer which feeds into a softmax layer.

Because MobileNet splits the standard convolution operations into two types of steps, namely filtering steps and combination steps, via the use of depthwise separable convolutions, it can achieve substantial reduction in computational cost. Single Shot Multi-Box Detector (SSD) is a novel architecture with a

single deep neural network (Liu et al., 2015). In the object detection task, this feed-forward convolutional network detects the presence of instances that belong to the specified object classes. It produces a fixed-size collection of bounding boxes and scores for these instances. Then it uses a method of non-maximum suppression to yield the final detection results. The structure of the SSD model is based on a truncated version of the VGG-16 network; a CNN commonly used for high quality image classification (Simonyan & Zisserman, 2014). A set of convolutional feature layers is added to the end of the truncated network, allowing predictions of object detection at multiple scales. The SSD model is able to complete the tasks of object localization and classification in a single forward pass of the network, which is exactly the meaning of "single shot". The SSD training objective is inspired by the objective of MultiBox (Erhan et al., 2014), an approach for the bounding box coordinate proposal generation.

MobileNet-SSD is essentially an SSD model that uses MobileNetV1 as a base network, or a backbone, in its structure. The original backbone, the truncated VGG-16 network, is replaced by a truncated version of MobileNetV1. The models reduce the computation cost while exhibiting a similar object detection accuracy. It is suitable for situations where we have only low computing power devices to perform object detection tasks in real time. For example, if MobileNet uses 3×3 depth-wise separable convolutions, its computational cost is between 8 to 9 times less than that of the full standard convolutions, while the accuracy of MobileNet is only about 1% less than that of the standard convolutions (Howard et al., 2017).

6. Experiments

6.1 Implementation Details

Our preliminary experiment focused on training the lightweight models to recognize dog bowls mainly. The 301 images containing at least one dog feeder were expanded to 1806 images by the five data augmentation techniques listed earlier. The input size of the original SqueezeDet model is 1242 × 375, which is too large to run on Google Vision Kit. So, we changed the input size to 256 × 256 and reduced both the number of vertical anchors and the number of horizontal anchors to 16 to make SqueezeDet model fit the supported configuration of the device. We implemented the network in Keras and did the training with the Adam optimizer, employing an initial learning rate of 0.001, learning rate decay factor of 0.5, decay step size of 10000 and a batch size of 4. A Tesla T4 GPU was adopted to train SqueezeDet in the experiment.

The input to the original MobileNet-SSD model was in the size of 300 × 300. We resized the input image to 256 × 256 and set depth multiplier to 0.125. When training MobileNet-SSD, we set the batch size to 24, the initial learning rate to 0.004 (with exponential decay), the decay factor to 0.95, and the number of decay steps to 800720. A single NVIDIA GeForce RTX 3090 GPU was employed.

The model with better performance in the preliminary experiment would be trained by larger dataset with more objects and deployed to the hardware devices. The training dataset was boosted by 500 images of dog beds, 600 images of dog toys (treat ball and chew toy) and their corresponding augmented images. The annotations of augmented images were generated automatically based on the labels on the original images.

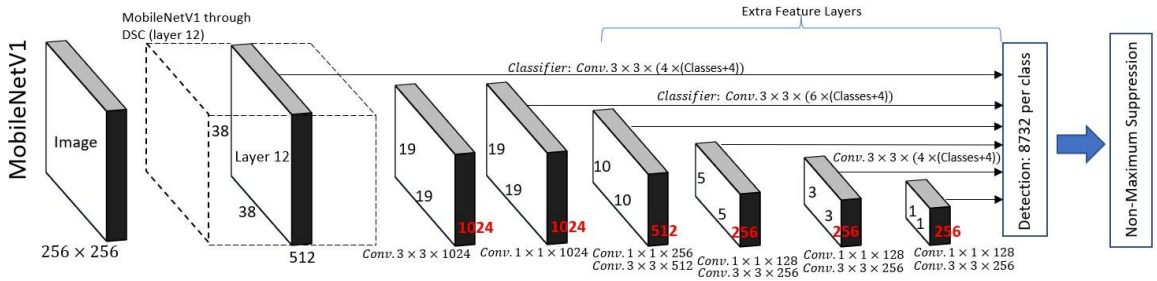
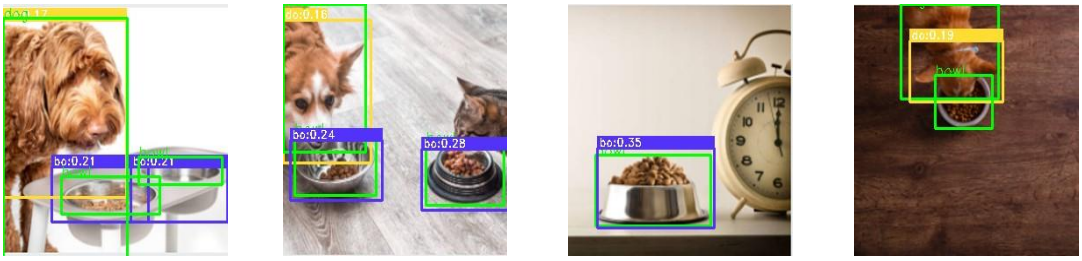


Figure 4: MobileNet-SSD architecture. To integrate MobileNetV1 within the original SSD framework, we truncate the last three layers of MobileNetV1 because the part of MobileNetV1 is not responsible for classification. It only needs to extract the features describing the contents of the images from the input images and pass them along to the other layers of the framework.

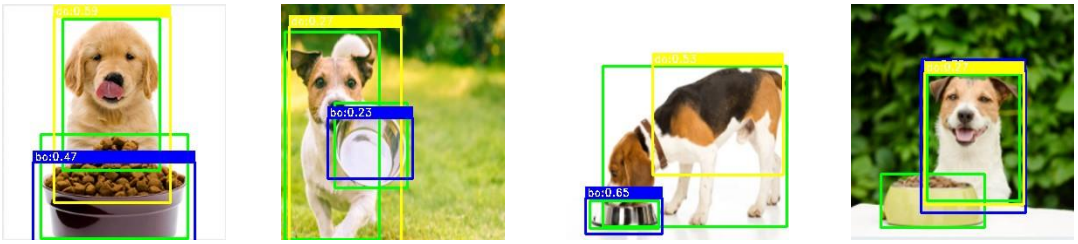
6.2 Experiment Results

In this subsection we demonstrate qualitative results of the bounding box detections produced by SqueezeDet and MobileNet-SSD on the categories *dog* and *dog bowl*. First, we analyze the performance of SqueezeDet. In Figure 5, we show the predictions of SqueezeDet. The ground truth bounding boxes are in green, while the prediction boxes of the *dog* and *dog bowl* are in yellow and blue, respectively. The threshold for the probability of detection is set to 0.13. According to Figures 5(a), 5(b), and 5(c), the trained multi-class object detector correctly localizes and labels the objects in the testing images with high (>0.7) IOU, but relatively low (approximately 0.3) probability scores. And some of the views, such as the top view of the partially occluded dog bowl, are challenging for the model to detect as Figure 5(d) displays. Besides, the detector tends to localize the main part with distinctive features of dog, for instance, in some cases the prediction boxes include the dog head but ignore the feet or trunk of the dog. While the prediction boxes of the dog bowl usually exceed the boundaries of the corresponding ground truth boxes. Furthermore, we examined the performance of MobileNet-SSD. The examples of MobileNet-SSD prediction are presented in Figure 6. In general, the probability scores are higher than those obtained by SqueezeDet. In particular, while the majority of the probability scores of MobileNet-SSD are between 0.4 and 0.6, the probability scores predicted by SqueezeDet are mostly between 0.2 and 0.4.



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

Figure 5: Examples of SqueezeDet prediction. In images (a)-(c), the detector successfully localizes and labels the objects. In image (d), the model fails to detect the occluded bowl from the top-down view.



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

Figure 6: Examples of MobileNet-SSD prediction. Generally, the probability scores are higher than those obtained by SqueezeDet. Image (c) shows a localization error, while image (d) shows a classification error and a missed bowl object.

Therefore, MobileNet-SSD was selected to be trained by the boosted dataset with labeled dog toy and bed objects. Figures 7(a), 7(b) and 7(c) exhibit the detections of treat ball, chew rope and dog cushion bed with relatively high confidence scores (above 50%) compared to the results in the preliminary experiment. Both Figures (c) and (d) show a detection error in which the missed bed and dog are highly truncated and over-lapped with other objects.

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

Figure 7: Prediction results of the MobileNet-SSD trained by boosted dataset. Images (a), (b), (c) show the detected treat ball, chew toy and dog bed respectively, while image (d) displays the missed objects.

7. Discussion & Future Work

In this research project, we explored the possibility of implementing lightweight neural networks to help robotic dogs recognize objects that real dogs recognize naturally. Trained by small-scaled low-resolution images from scratch, the low power networks could localize and label multiple items with an acceptable accuracy of probability scores and a relatively high IOU. The experiment results show that the small-sized models under the configuration constraints have the potential to detect the objects in the real time video stream even if the input frames are of low quality. Further improvements in object detection may be achieved by leveraging supplementary sharpened images during the training process to encapsulate more details about objects of interest and improve the accuracy of the confidence and probability scores. Additionally, the trained model will be integrated into the Vision Bonnet by exporting the generated checkpoint as a frozen graph, using the compiler to convert the frozen graph into binary format, and copying it onto the Vision Kit. The feasibility of the models will be tested on the intelligent camera.

Our continuing work on the robotic dog system will explore a machine learning model aiming at specific hand sign language features based on the YOLO object detection algorithm (Redmon et al., 2015). With respect to this we have collected over 1,500 images from different angles to see the hand signs from a robotic dog perspective. Currently, those images are being augmented using noise and blur techniques that increase the size of the training dataset. While some training experiments with the model have been carried out, detailed evaluations and comparisons to other algorithms such as Single Shot Detector and Efficientnet (Tan & Le, 2019) are still to come. For this, we are designing further experiments with different distances and angles from a dog perspective.

An alternative approach that is also being explored is the employment of direct h4444and and finger motion tracking devices such as data gloves. While the camera-based tracking and recognition of hand signs require clear views, proper lighting, and suitable backgrounds, and is thus highly dependent on the environmental conditions, the direct data glove-based tracking is fundamentally more robust. With respect to this, we have conducted experiments with a deterministic low computing power model for data glove-based tracking and recognition of hand signs employing an extension of the Malossi alphabet (Gelsomini et al, 2022). The advanced Data Glove incorporating highly stretchable Carbon Nanotube sensors developed at the Research Institute of Electronics that was commercialized by Yamaha was used in the experimental work (Gelsomini et al, 2021). Further experiments are planned with newer Data Glove models incorporating haptic feedback and employing machine learning models evolving from the camera-based approach discussed in this paper.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kud-lur, M., & Zheng, X. "Tensorflow: Large-scale machine learning on heterogeneous systems," Software available from tensorflow.org.

Anderson, G. O., & Thayer, C, E. (2018). "Loneliness and social connections: A national survey of adults 45 and older," *AARP Research*.

Banks, D., Marian R., Willoughby, P., Lisa M., & Banks, M., William A. (2008). "Animal-assisted therapy and loneliness in nursing homes: Use of robotic versus living dogs," *Journal of the American Medical Directors Association*, 9(3), 173–177.

Bruno, D. R., de Assis, M. H., & Osório, F. S. (2019). "Development of a mobile robot: Robotic guide dog for aid of visual disabilities in urban environments," *2019 Latin American Robotics Symposium (LARS)*, 104–108.

Cacioppo, J. T., Hughes, M. E., Waite, L. J., Hawkley, L .C., & Thisted, R. A. (2006). "Loneliness as a specific risk factor for depressive symptoms: Cross-sectional and longitudinal analyses", *Psychology and Aging*, 21(1), 140–151. doi: 10.1037/0882-7974.21.1.140

Erhan, D., Szegedy, C., Toshev, A., & Anguelov, D. (2014). "Scalable object detection using deep neural networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gelsomini, F., Hung, P., Kapralos, B., Uribe-Quevedo, A. J., Jenkin, M., Tokuhiro, A., Kanev, K., Hosoda, M., & Mimura, H. (2021). "Specialized CNT-based Sensor Framework for Advanced Motion Tracking," *54th Hawaii International Conference on System Sciences*, 1–8.

Gelsomini, F., Tomasuolo, E., Roccaforte, M., Hung, P., Kapralos, B., Dubrowski, A., Quevedo, A. J. U., Kanev, K., Hosoda, M., & Mimura, H. (2022). "Communicating with Humans and Robots: A Motion Tracking Data Glove for Enhanced Support of Deafblind," *55th Hawaii International Conference on System Sciences*, 1–9.

"Groove x," (n.d.). (Available online at https://groove-x.com/en/).

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications.

Jones, C. M., & Deeming, A. (2007). "Investigating emotional interaction with a robotic dog," *Proceedings of the 19th Australasian Conference on ComputerHuman Interaction: Entertaining User Interfaces*, 183–186.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C., & Berg, A. C. (2015). "SSD: single shot multibox detector," *CoRR, abs/1512.02325*.

Luo Y, Hawkley, L. C., Waite, L. J., & Cacioppo, J. T. (2012). "Loneliness, health, and mortality in old age: A national longitudinal study," *Social Science & Medicine (1982)*, 74(6), 907–914. doi: 10.1016/j.socscimed.2011.11.028

Melson, G., Kahn, P., Jr, Beck, A., Friedman, B., Roberts, T., & Garrett, E. (2005). "Robots as dogs?: Children's interactions with the robotic dog aibo and a live australian shepherd," *CHI '05 Extended Abstracts on human factors in computing systems*, 1649–1652.

Redmon, J., Divvala, S. K., Girshick, R. B., & Farhadi, A. (2015). "You only look once:

Unified, real-time object detection," *CoRR, abs/1506.02640*. arXiv: 1506. 02640.

Schellin, H., Oberley, T., Patterson, K., Kim, B., Haring, K. S., Tossell, C. C., Phillips, E., & Visser, E. J. D. (2020). "Man's new best friend? strengthening human-robot dog bonding by enhancing the doglikeness of sony's aibo," *2020 Systems and Information Engineering Design Symposium (SIEDS)*, 1–6.

Sekachev, B., Zhavoronkov, A., & Manovich, N. (2019). "Computer Vision Annotation Tool: A Universal Ap-proach to Data Annotation," 2019. (Available at https://www.intel . com / content / www / us / en/developer / articles / technical / computer - vision - annotation - tool - a - universal - approach- to - data -annotation.html;).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition.

Stanton, C., Kahn Jr, P., Severson, R., Ruckert, J., & Gill, B. (2008). "Robotic animals might aid in the social development of children with autism," *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 271–278.

Tan, M., & Le, Q. V. (2019). "Efficientnet: Rethinking model scaling for convolutional neural networks," In K. Chaudhuri & R. Salakhutdinov (eds.), *Proceedings of the 36th international conference on machine learning, ICML 2019, 9-15 june 2019, long beach, california, USA*, (pp. 6105– 6114). PMLR.

"Vision kit," (n.d.). (Available online at https://aiyprojects.withgoogle.com/ vision).

Weiss, A., Wurhofer, D., & Tscheligi, M. (2009). ""i love this dog"—children's emotional attachment to the robotic dog aibo," *International journal of social robotics*, *1*(3), 243–248.

Wu, B., Iandola, F. N., Jin, P. H., & Keutzer, K. (2016). “Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving,” *CoRR, abs/1612.01051*. arXiv: 1612.01051.

Xue, T., Wang, W.-M., Ma, J., Liu, W., Pan, Z., & Han, M. (2020). “Progress and prospects of multimodal fusion methods in physical human–robot interaction: A review,” *IEEE Sensors Journal*, *PP*, 1–1.