QUEST FOR THE BEST CAT PHOTO

Kyra Ballard, Andrew Fu, Shravan Nageswaran, Emily Xie
IACS Capstone '20 in Partnership with Adoptimize and Austin Pets Alive

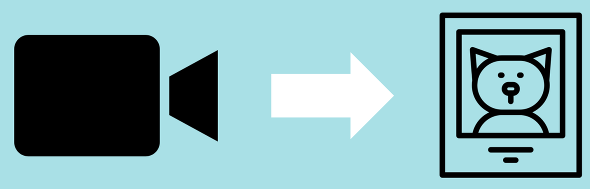
Harvard John A. Paulson School of Engineering and Applied Sciences

OVERVIEW

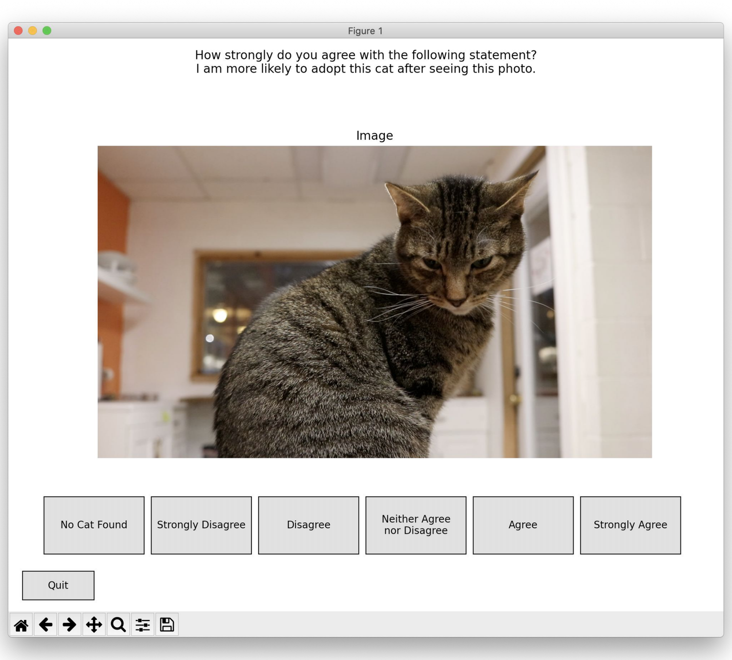
Our aim is to help increase the adoption rates of cats at shelters in order to prevent their euthanasia. While high quality photos are crucial in helping pets find new owners, it's difficult to take good shots of cats given their skittish nature. Hence, we have partnered with Adoptimize and Austin Pets Alive in researching a model that, given a video of a cat, finds the optimal frame.

GOAL

Given a video of a cat, find the best photo.



DATASET



Our partners supplied us with an initial data set of cat videos.

Missing Dependent Variables
One hurdle in our dataset was lack of annotation. As we needed a response variable that measured the quality of the frames, we developed a bare-bones application and used it to score 497 randomly selected frames across our dataset, using a 5 point Likert scale, where 1 represents poor photo quality, and 5 signifies excellent.

Our initial dataset at a glance:

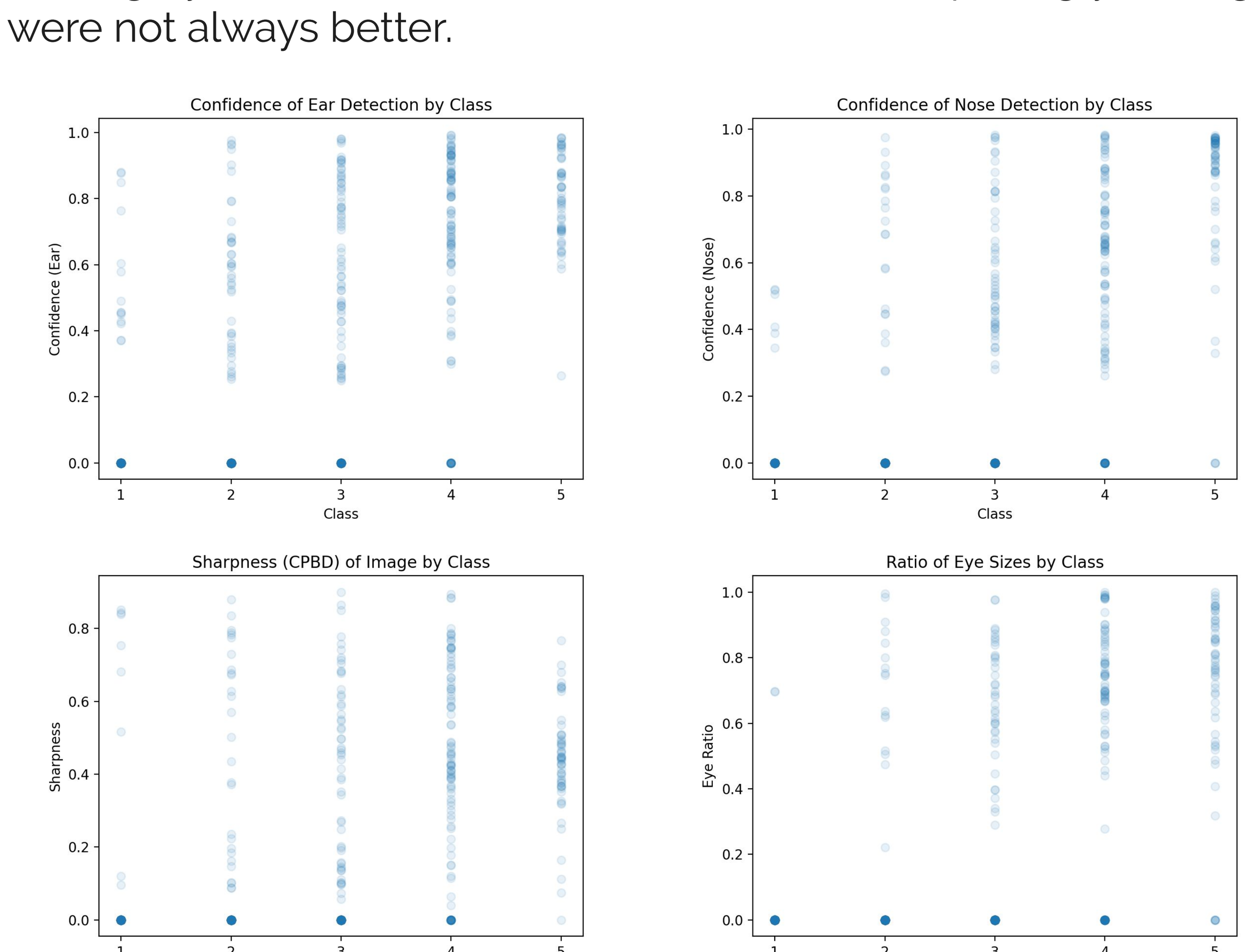
79	23	55	60	2
total videos	average duration in seconds	max duration in seconds	median frames per second	median resolution in megapixels

Upon ranking individual frames, our resulting dataset was slightly biased, given that there was a low proportion of high quality frames (only 12% were scored at a 5). Thus, after creating our initial model, we decided to expand this dataset manually by acquiring 20 additional videos from the internet and ranking individual frames (luckily, the internet has no shortage of cat videos).

FEATURE ENGINEERING

Our most important features are the confidence values for the detection of the cats' heads, ears, eyes, and noses. We investigated both Haar Cascade models as well as YOLOv3 for object detection; our final model uses YOLOv3 (Redmon et al.) because it had a much higher recall rate.

From these features, we also investigated using the relative size of the head, the head's distance from the center of the image, and the ratio between the size of the eyes. Finally, we also used the sharpness of the image as calculated using the CPBD (Narvekar et al.) metric as a feature. The object detection confidence metrics were most highly correlated with the class labels. Surprisingly, though, sharper images were not always better.



MODELING

Baseline Model

The baseline model selected a random image from the set of frames in which a cat head was detected.

Initial Model

Our initial model was based on a subjective function of the features. Once we had sufficient data labeled with image quality classes between 1 and 5, we were able to use a more data-driven approach to constructing the model. From the set of frames in which a cat head was detected, this initial model selected the image with the highest combined scores of sharpness and cat head ratio.

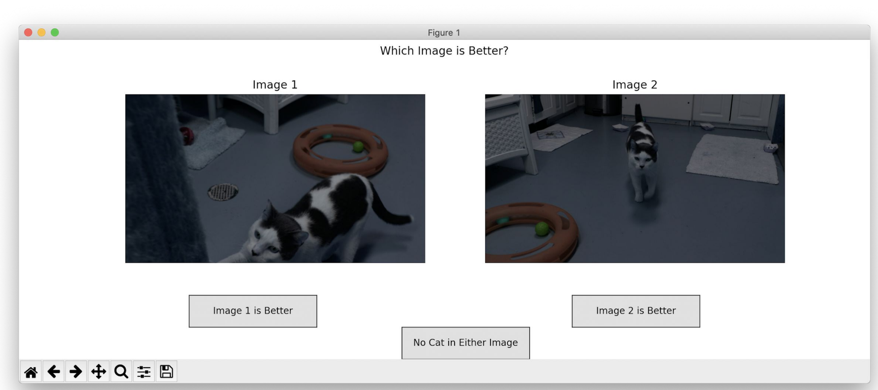
Final Model

We used **logistic regression** to build a classifier that predicts the image quality label given each of the features. To turn this into a model that selects a frame from a video: we select the frame that has maximum probability of being good or excellent (class 4 or 5). Note that the results of the regression revealed that the size of the head and its distance from the center are uninformative, so were removed.

Other Models Explored

We considered using deep learning models as well, including training a convolutional neural network from scratch as well as fine tuning the InceptionV3 network pretrained on ImageNet. However, the accuracy of these models was low because we had very little data. Linear regression also did not outperform logistic regression.

TESTING



Because of the subjective nature of the problem, to test the performance of our model, we built a basic blind A/B testing platform, and used it to rate the frames from our developed models against the baseline.

FINAL RESULTS

After refining our selected logistic regression model, we tested its performance, and found decent results. The following are the performance results we achieved on the initial and final models relative to the baseline:

64%	96%
Initial Model Performance	Final Model Performance

Below is a detailed breakdown of these results; 96% of the time, testers chose either the output of the model or could not differentiate between the model and the baseline.

Frame Selected	Percent of Time Selected
Frame from Model	74.6%
Frame from Baseline	3.8%
Neither	21.5%

As our ultimate deliverable, we created a simple web application which takes in a video and displays the output of our final model.

CITATIONS AND REFERENCES

J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," CoRR, vol. abs/1506.02640, 2015.

Narvekar, Niranjan & Karam, Lina. (2011). A No-Reference Image Blur Metric Based on the Cumulative Probability of Blur Detection (CPBD). Image Processing, IEEE Transactions on. 20. 2678 - 2683.

Icons used on this poster made by Freepik from flaticon.com.