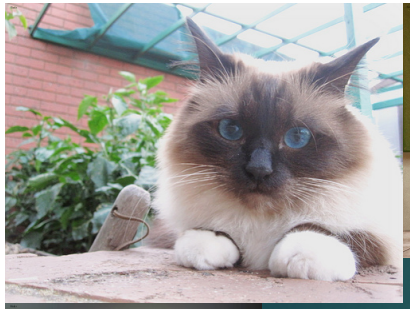
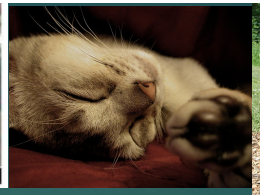
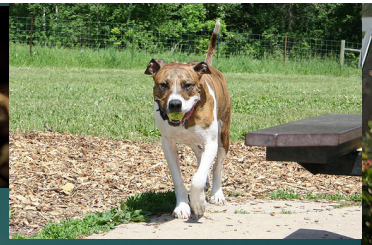
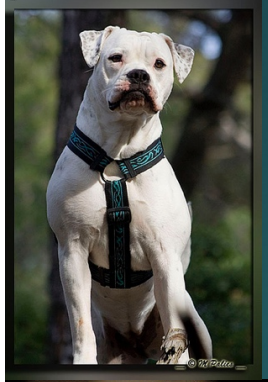
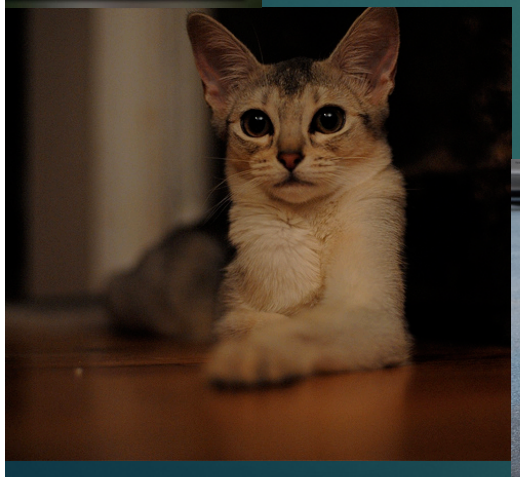
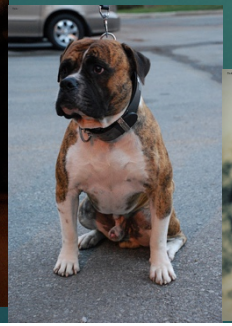
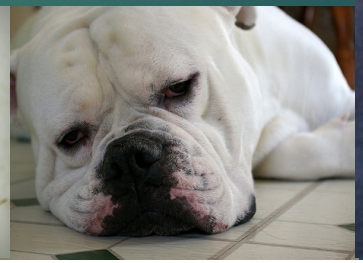







Classifying Cats vs. Dogs

STAT W4249 APPLIED DATA SCIENCE - GROUP 5

GU, XINGHAO

ISLAM, SCHINRIA REMA

LAU, ARNOLD CHUA

ZHOU, YI

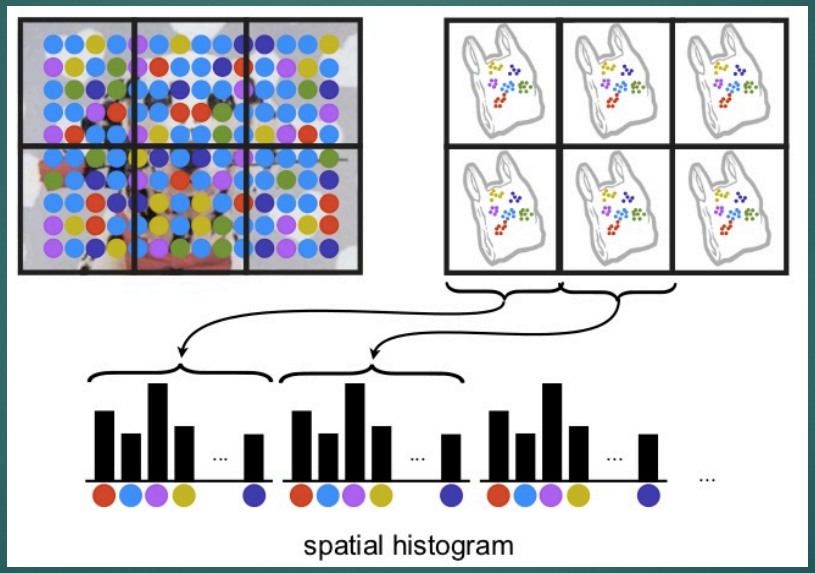
ZHU, YIBO

Contents

- I. Feature Selection and Extraction
- II. Model Selection
- III. Comparison with Baseline

Feature Extraction

- Spatial color histogram of images
- Harmonic coefficients from elliptic Fourier outline analysis



Reference: Lazebnik, Svetlana, Cordelia Schmid, and Jean Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories." Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Vol. 2. IEEE, 2006.
Ma, Yuting. ADS Image Analysis Powerpoint Presentation.

Spatial Color Histogram

- The baseline color histogram features do not take into account the possibility that more relevant information will be in the center of the image
  - "Images were dropped if any of the following conditions applied… (i) the image was gray scale, … (iv) *the pet was not centered in the image*" (Parkhi et al. 2012)
- We decided to divide the picture into a 3x3 grid and create a color histogram with 5 red, 5 green and 5 blue bins for each cell, for a total of 1,125 features

Harmonic Coefficients from Elliptic Fourier Outline Analysis

- We obtained image outlines by:
  - Converting to grayscale, smoothing to remove background noise, and then applying a global threshold determined for each image via Otsu's method
  - Pixel intensities beyond the threshold are made black and those under the threshold are made white

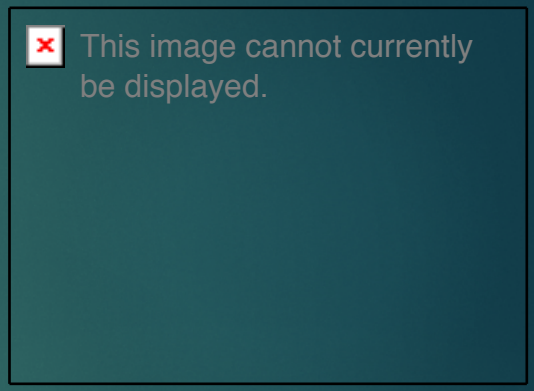
This image cannot currently be displayed.

This image cannot currently be displayed.

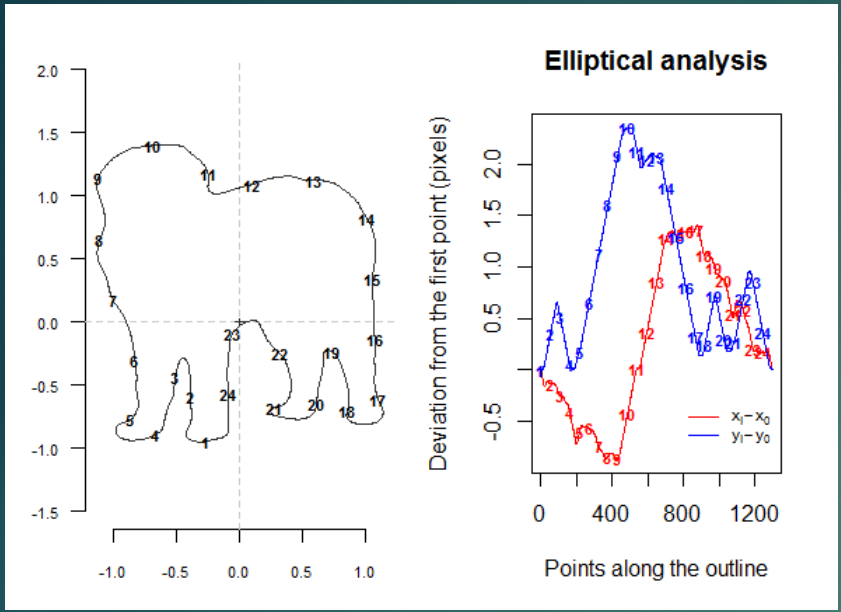
This image cannot currently be displayed.

Image processing was not always perfect

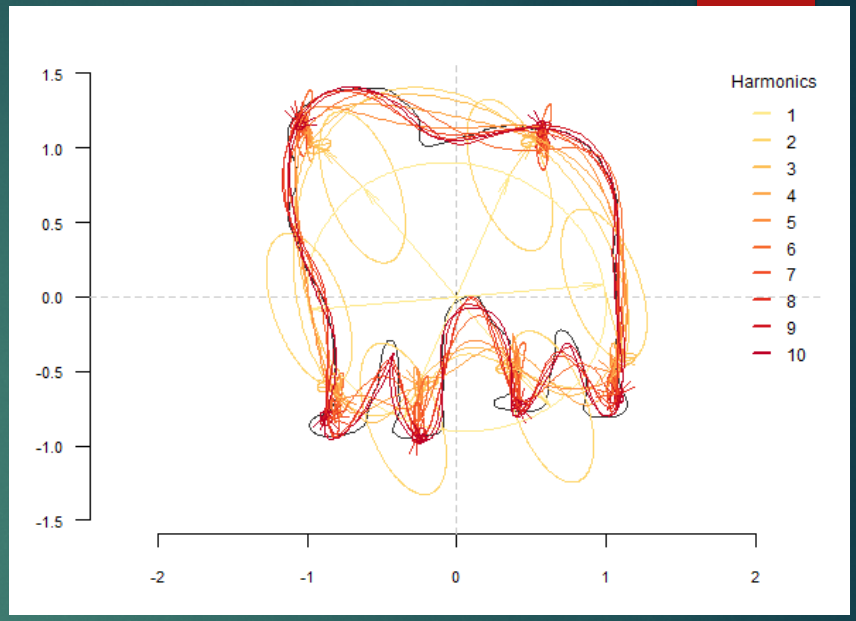
Harmonic Coefficients from Elliptic Fourier Outline Analysis



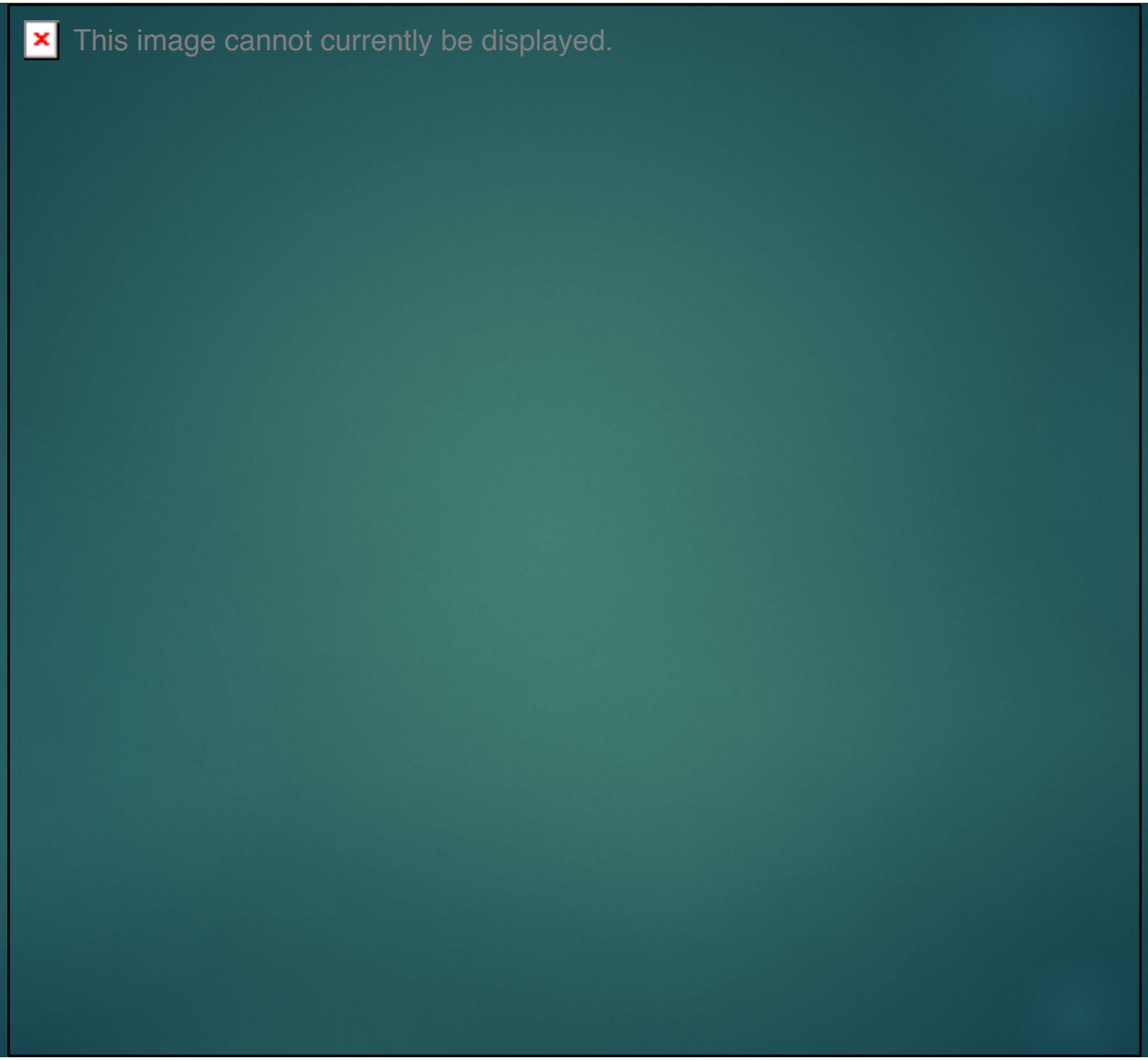
- We obtained image outlines by:
  - Running the Conte algorithm from the Momocs library
  - Centering and scaling the outline
- We extracted features by:
  - Applying an elliptical Fourier transform, where we break up the outline into separate x and y curves and then approximate it with 10 ellipses, which almost always approximated 99% of the outline. The coefficients through which each ellipse is defined become the features.
  - Since only the outline is used, these features are **scale-invariant**.



Key points along the outline are identified and then x and y curves are graphed separately



Outline is then approximated with ellipses


This image cannot currently be displayed.

Model selection

This image cannot currently be displayed.

This image cannot currently be displayed.

Random Forest

- Bagging with tree classifiers as weak learners
- Uses an additional step to remove dimensions that carry little information

Adaboost

- combines the outputs of many weak learners to produce a better prediction
- Each classier is trained to the training data using weights

Algorithms

- Random forest uses the bagging method, which selects subsets of observations randomly and builds the tree models based on every subset. Thus all the tree models in random forest actually share the same weight. While for Adaboost, it trains the weak learners in order, and the weight of every weak learner depends on its prediction accuracy.
- Also, Adaboost is better at reducing correlation between the weak learners than bagging is.
- Although adaboost costs a little more time, while consider its accuracy improvement, the running time is completely acceptable.

Select Optimal Parameters

To cut shorten the program's running time and increase the prediction accuracy, we try to select the proper parameters for the adaboost model by using 5-fold cross-validation.

- Number of weak learners: 40

  Too many weak learners increases the running time and leads to overfitting

- Shrinkage parameter: 0.1

  Trade-off between running time and the risk of overfitting

- Boosting type: discrete

  The accuracy of model

Comparison with Baseline

Testing error if we predict that all images are dogs: 67.13%

	Color histogram	Spatial color histogram + harmonic coefficients
Linear SVM	Running time: 79.18 secs Training accuracy: 69.06% Testing accuracy: 66.9%	Running time: 66.3 secs Training accuracy: 69.18% Testing accuracy: 67.7%
AdaBoost	Running time: 50.9 secs Training accuracy: 72.58% Testing accuracy: 69.09%	Running time: 416.6 secs Training accuracy: 78.98% Testing accuracy: 72.13%

Our final model, AdaBoost with spatial color histogram and harmonic coefficient features, using 40 weak learners, takes much longer to run than the baseline but also correctly classified around 120-130 additional images on testing data.

* 5000 images were used for training data, and the remaining images were used for testing.

** Running times obtained on MacBook Pro running OS X Yosemite 10.10.5, 2.2 GHz Intel Core i7, 16 GB RAM

Conclusion

- Our final model improves accuracy but at the cost of running time
- Possible further avenues for improvement: dimension reduction (though AdaBoost may overfit with too few dimensions), finding a more consistent method to extract outlines, incorporating other possible features such as bag-of-visual-words

THANK YOU!