**Review of Existing Research**

**Image Classification Models**

In (3) the authors describe a method for making an assessment of the presence, and then classification to the correct sign, for street signs that would be detected in the environment in real-time. This type of system would be required for an autonomous or assistive vehicle to operate in the environment with other vehicles. They built upon prior research, which focused around shape and color applied to specific areas of the image and sign. Their assessment asserts that when an image fails the first step, either color or shape recognition, it would not proceed to the second step (either shape or color). The authors in this paper propose an AdaBoost based detection system, paired with a Bayesian classification system. The AdaBoost part of the model includes both a color and shape. The model delivered low rates of missing signs (1.4%) and a very low rate of false positives (detecting signs when none was present, 0.03%). Signs were correctly classified in 94% of cases.

In (4) the authors are attempting to classify different images into classes that represent the overall composition of the image, instead of the items within the image. In this case, an image could be a label of rooms in a house (Kitchen, Bedroom, Dining Room), instead of the items in the room (Bed, Refrigerator, Table, Chair). The model described attempts to follow a similar process to a human labeling an image based on the items comprising the scene. In this case, the image is divided into a codebook of parts of the images. The codebook is a random selection of regions of the original image that provide insight into the reoccurring elements that comprise images in different classes. The codebook is further reduced by using k-means clustering to remove low frequency arrangements from the codebook. The advantage of the model shown by the authors is the low degree of supervision required, and the ability to learn classes that might be composites of other classes. The model achieved a 76% accuracy rate when tested with 100 images.

In (8) Knopp et al. attempted to recognize the location depicted in a geo-coded image from a street view type database. Their model uses the geocoding information to reduce the domain of potential classifications of the scene depicted in the image. The database of images consisted of 17,000 images harvested from Google Street View, combined with 8,000 images from the photosharing website Panoramio. Images were concentrated in Paris, with an objective of classifying the place shown in the image. The model used a “Visual bag of words” approach to describe features within the images, with a second step that considers the layout of those features within the image. A confusion score was developed to help show the amount of confusing features that, while recognizable, do not help discriminate the location of an image. The best result provided 47.9% accuracy of classifying locations within Paris.

In (9), Le et al seek to create a classification model without the pretense of a set of labeled objective data. The authors characterized this type of learning as “self-taught” learning that has previously been used to find basic features like blobs and edges. Data in this study was comprised of single frames of video from 10 million different YouTube videos. The best performing neuron achieved a 81% accuracy in facial recognition.

**Convolutional Neural Networks for Images**

In (5) Goodfellow and a team from Google used the DistBelief method of deep neural networks to train a model to recognize multi-digit numbers from Google Street View. The problem presented is very broad – that street address numbers are represented in a variety of different formats, locations, and with changing conditions from lighting and obstructions. The paper addressed first using the model to recognize numbers from the Google Street View House Number dataset, and then from a more complete Street View image dataset. The recognition of street numbers in images represents a special problem – in order to be useful, all digits in the sequence must be correctly recognized. Images were pre-processed for this paper by “subtracting the mean of each image.” Working in isolation for the Street View House Number dataset, the model was trained over 6 days and achieved 97.84% accuracy (just short of their benchmark of 98% for human recognition.) The Street View House Number data set does have additional pre-processing work to isolate the house numbers and bound the area that the numbers exist within. The more complete Street View data set carried additional challenges of a larger image field to locate the numbers, with more variation in the type and layout of the numbers as well as blurring of some images. The Google Street View data set yielded 91% accuracy. The full image data set used five convolutional layers. For Google, this method has provided scalability to the translation of Street View images to maps: more than 100 million addresses at time of the writing of the paper had been translated to images to mapping. (As a sidenote to this paper, DistBelief was a closed-source effort that was a predecessor to the well-known and open source “TensorFlow” library.)

In (6) Simonyan and Zisserman used convolutional neural networks to classify images for the ImageNet Large-Scale Visual Recognition Challenge. Their neural networks were configured to use a small 3x3 receptive field, versus prior years where fields as large as 11x11 were used. Images were also randomly scaled for training, so that object identification became independent of the size of the object within the image. Training was carried out using logistic regression and gradient descent. Initially, the learning rate was set at 0.01 and was reduced by a factor of 10 as the error rate was not changing further. Their final model used the combined outputs of two neural networks, labeled as multi-crop and dense evaluation and achieved an error rate of 23.7%, better than the comparison set of models from prior years.

**Neighborhood Quality/Infrastructure Surveys**

**Walkability/Public Health Concerns**

In (7) Deehr and Shumann considered walkability of five neighborhoods in Seattle with populations that are usually underrepresented in the city processes. Data employed in their selection method considered health factors, occurrence of motorists striking pedestrians, and modes of transportation used by the area’s residents. Policy changes that were recommended by the project include consideration of walking as a mode of transportation equal to car, bicycle or public transit. The team also recommended changes in design for accommodation of pedestrians. The project also helped start community-led projects to add walking trails to several neighborhoods in Seattle. Finally, the project also led to promotional efforts oriented towards citizens to identify walking trails, and tie together walking with other modes of transit. A master plan for making Seattle the most walkable city in the country was adapted in 2009.

**Project Sidewalk Predecessors**

Project Sidewalk, a broad scale research project at the University of Maryland, has established several lines of research that will provide a body of work and basis for this project’s extension to machine learning and image classification.

Broadly, Project Sidewalk has used crowd sourced labor to label features on Google Street View related to the accessibility issues. In (1), the researchers had an objective of labeling bus stop attributes by site, using crowd sourced labor. First the researchers first worked to define requirements for cataloging transit access points for persons with disabilities. After defining the requirements for cataloging site attributes, the team crowd sourced labored to help define all of the physical attributes of each transit access point. The visuals within Google Street View were compared in an audit of 179 sites to ensure that the images in Google Street View were an accurate representation of the actual physical environment. The auditors found that 29/179 surveyed sites were missing from Google Maps data but were found in the physical survey.

Crowd sourced labor for labeling was provisioned through the Amazon Mechanical Turk service using minimally trained users. The training was incorporated for the user as part of the introduction to the work. The interface allowed users to label attributes, specific to bus transit, included signage, shelters, seating and trash disposal bins. The crowd sourced labor was found to have an accuracy of 82.5% in labeling features. Errors tended more in the direction of omission of features. The sample size of sites was 150.

A similar Project Sidewalk team in (2) applied a similar process in another research work. Again, images were sourced from Google Street View to attempt to catalog physical attributes of accessibility to disabled individuals. Again, the images were presented to crowd sourced labor for purposes of cataloging the attributes. The crowd sourced labor this time achieved a high score on a Spearman Rank Correlation. The process was further expanded in this paper with a discussion of a machine learning extension to the project.

A four step process was described. First, images were scraped from Google. Then crowd sourced labeling provided wire framing of locations of curb cuts, and other crowd sourced labeling verified the correctness of the labels. Using the correctly labeled dataset, another process called svDetect uses a “Deformable Part Model” to attempt to identify the features of a training set in another set of images. Another step of a support vector machine was also used to help refine the performance of the model. The performance of the model was limited, reaching less than 80% in recall.

**Sources**

1. Kotaro Hara, Shiri Azenkot, Megan Campbell, Cynthia L. Bennett, Vicki Le, Sean Pannella, Robert Moore, Kelly Minckler, Rochelle H. Ng, and Jon E. Froehlich. 2015. Improving public transit accessibility for blindriders by crowdsourcing bus stop landmark locations with Google Street View: An extended analysis. ACM Trans. Access. Comput. 6, 2, Article 5 (March 2015), 23 pages.
2. Kotaro Hara, Jin Sun, Robert Moore, David Jacobs, Jon E. Froelich. 2014. Tohme: Detecting Curb Ramps in Google Stret View Using Crowdsourcing, Computer Vision, and Machine Learning. UIST ’14, October 5-8 2014. Honolulu, HI.
3. Claus Bahlmann, Ying Zhu, Visvanathan Ramesh, Martin Pelkofer, Thorsten Koehler. “A System for Traffic Sign Detection, Tracking, and Recognition Using Color, Shape, and Motion.” IEEE Xplore, July 2005.
4. Fei-Fei, Li, and Pietro Perona. "A bayesian hierarchical model for learning natural scene categories." *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 2. IEEE, 2005.
5. Goodfellow, Ian J., et al. "Multi-digit number recognition from street view imagery using deep convolutional neural networks." *arXiv preprint arXiv:1312.6082* (2013).
6. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).
7. Deehr, Rebecca C., and Amy Shumann. "Active Seattle: achieving walkability in diverse neighborhoods." *American journal of preventive medicine* 37.6 (2009): S403-S411.
8. Knopp, Jan, Josef Sivic, and Tomas Pajdla. "Avoiding confusing features in place recognition." *Computer Vision–ECCV 2010* (2010): 748-761.
9. Le, Quoc V. "Building high-level features using large scale unsupervised learning." *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013.