RecVis16 Final Project
Using Visual Elements to Infer Human Perceptions of Places

Kimia Nadjahi, Rachid Riad

1. Introduction

Sense of Place is a feeling or perception held by people about a location: some characteristics of a place can be perceived at first sight, such as wealth or safety. Lately, there has been recent interest in predicting these human judgments with computer vision techniques [Ordonez and Berg 2014].

The CNN architecture with the NetVLAD layer from [Arandjelović et al. 2016] significantly outperforms non-learnt image representations as well as off-the-shelf CNN descriptors, and improves over the state-of-the-art on challenging image retrieval benchmarks. The goal of this project is to transfer the CNN representation learnt for Visual Place Recognition to predict human judgments of safety and wealth of locations.

2. Datasets and implementation details

We will use the VGG ConvNet from [Arandjelović et al. 2016] pre-trained on Pittsburgh, as well as two main datasets: Place Pulse V1.0[1] for New-York and Chicago, which is a collection of streetview images labeled with human judgement scores of perceived safety and wealth ; a larger unlabeled collection of streetview images for these two cities. All images will be sampled from Google Street View API.

The whole project will be implemented in Matlab in order to use the MatConvNet toolbox[2]. The distribution of work can be followed on GitHub [3].

3. Experiments

As in [Ordonez and Berg 2014], we will model and evaluate prediction of perceptual characteristics as a classification problem and a regression one. But first, we will need to decide how to represent images.

Image Representation. The features would be the output of the CNN layer that gives the best results.

Classification. The classification will be binary: "images with high perceptual scores" and "images with low perceptual scores" (e.g. "safe" vs "not safe"). The models to predict the labels from input image representations will be trained using an ℓ_2-regularized with a squared hinge-loss function linear SVM classifier. The performance will be evaluated with precision-recall plots and mAP values.

Regression. We want to predict scores using linear regression with a ℓ_2 regularization. The performance will be evaluated with r-correlation coefficients, and a visual comparison of the ground truth scores to the predictions on heat maps.

In both problems, we will consider two scenarios: training and testing on the same city, then training on one city and testing on another. Plus, we will also show some qualitative results by studying images with the lowest/highest scores.

Large Scale Experiments on Unlabeled Data. We will finally run our trained models on the extended dataset, and compare our predictions for each city to the ones found on the original samples from Place Pulse. This comparison would be qualitative, using heat maps.

References

[Arandjelović et al. 2016] Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[Ordonez and Berg 2014] Ordonez, V. and Berg, T. (2014). *Learning high-level judgments of urban perception*, volume 8694 LNCS of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 494–510. Springer Verlag, Germany, part 6 edition.

[1] http://pulse.media.mit.edu/data/

[2] http://www.vlfeat.org/matconvnet/

[3] https://github.com/Rachine/VisualPlaceRecognition