Looking Beyond the Visible Scene

Aditya Khosla* Byoungkwon An* Joseph J. Lim* Antonio Torralba
Massachusetts Institute of Technology
{khosla, dran, lim, torralba}@csail.mit.edu (* - indicates equal contribution)

1. Extended Abstract

"Daddy, daddy, I want a Happy Meal!" says your son with a glimmer of hope in his eyes. Looking down at your phone, you realize its fresh out of batteries, "how am I going to find McDonald's now?" you wonder. Looking left you see mountains and on the right some buildings. Right seems like the right way. Still no McDonald's in sight, you end up at a junction; the street on the right looks shady, its probably best to avoid it. As you walk towards the left, you are at a junction again; a residental estate on the left and some shops on the right. Right it is. Shortly thereafter, you have found your destination, all without a map or GPS!

A common thread that ties previous works together is their focus on the aspects directly present in a scene. In this work, we propose to *look beyond* the visible elements of a scene; a scene is not just a collection of objects and their configuration or the labels assigned to the pixels, it is so much more. From a simple observation of a scene, we can tell a lot about the enviornment surrounding the scene such as the potential establishments near it, the potential crime rate in the area, or even the economic climate. See Fig. 1 for example. Can you rank the scenes based on their distance from the nearest McDonald's? What about ranking them by the crime rate in the are? You might be suprised by how well you did despite having none of this information readily available from the visual scene.

In our daily lives, we are constantly making decisions about our environment such as, is this location safe? Where can I find a parking spot? Where can I get a bite to eat? We do not need to observe a crime happening in real-time to guess that an area is unsafe. Even without a GPS, we can often navigate our environment to find the nearest restroom or a bench to sit on without performing a random walk. Essentially, we can look *beyond the visible scene* and infer properties about our environment using the visual cues present in the scene.

In this work, we explore the extent to which humans and computers are able to look beyond the immediately visible scene. To simulate the environemnt we observe around us, we propose to use Google Street View data that provides a panoramic view of the scene. Based on this, we show here that it is possible to predict the distance of surrounding establishments such as McDonald's or hospitals even by using a scenes located far from them. We go a step further to show that both humans and computers perform reasonably at navigating the environment based only on visual cues from scenes that contain no direct information about the target. Lastly, we show that it is possible to predict the crime rates in an area simply by looking at a scene without any real-time criminal activity.

Figure 1. Can you rank the images by their distance to the closest McDonald's? What about ranking them based on the crime rates in the area? Check your answers below[1]. While not directly visible i.e. we do not see any McDonald's or crime in action, we can predict the possible actions or the type of surrounding esstablishments from just a small glimpse of our surroundings.

We emphasize that the goal of this paper is not to propose complex mathematical equations; instead, using simple yet intuitive techniques, we demonstrate that humans and computers alike are able to understand their environment by just seeing a small glimpse of it in a single scene i.e. a scene is much more than just its constituent elements. Interestingly, we find that despite the relative simplicity of the proposed approaches, computers tend to outperform humans in a variety of tasks. We believe that inference beyond the visible scene based on visual cues is an exciting avenue for future research in computer vision and this paper is merely a first step in this direction.

Please visit our website, http://mcdonalds.csail.mit.edu, for the full paper/supplemental material.

[1] Answer key: crime rate (highest) B > E > C > F > D > A (lowest), distance to McDonald's: (farthest) A > F > D > E > C > B (closest)