Visual Attention Project - FIGRIM Dataset

Valerijan Matvejev & Hana Ujcic

December 10, 2023

Chapter 1

FIGRIM dataset

1.1 Description

We are working with .jpg images from the FIGRIM (FIne GRained Image Memorability) dataset. It was published in the "Intrinsic and extrinsic effects on image memorability" article by Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, Aude Oliva from 2015, and contains 636 images grouped in 21 scene categories. The images span 21 scene categories from the SUN database. Each scene category was chosen to contain images of size 700x700 or greater (1000x1000 in our case).

1.2 Limitations and challenges

The dataset's focus on specific scene categories introduces limitations in terms of generalizability to images outside of these categories. Models trained on such biased data may not perform well on diverse scenes.
The fixed image size of 700x700 pixels limits the dataset's applicability to scenarios where images have different resolutions or aspect ratios. This may be a constraint in real-world applications with varying image sizes.

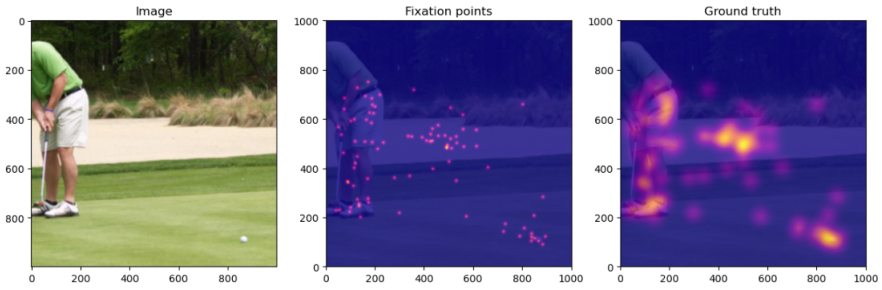
Chapter 2

Analysis of dataset

2.1 Ground truth data

The observation of high clustering in gaze tracks suggests that there is a tendency for observers to concentrate their visual attention on specific regions of the images. This behavior may indicate that certain areas attract more attention or are perceived as more salient, leading to a clustered distribution of gaze points.
We can observe centered-bias, and it refers to the tendency of observers to concentrate their gaze more towards the center of an image, where the main interest of the image is located. The center bias can be seen in images in which the main interests are on the edges of the image, but there is still a high clustering of gaze tracks in the center, such as in Figure 2.1.

Figure 2.1: Center bias

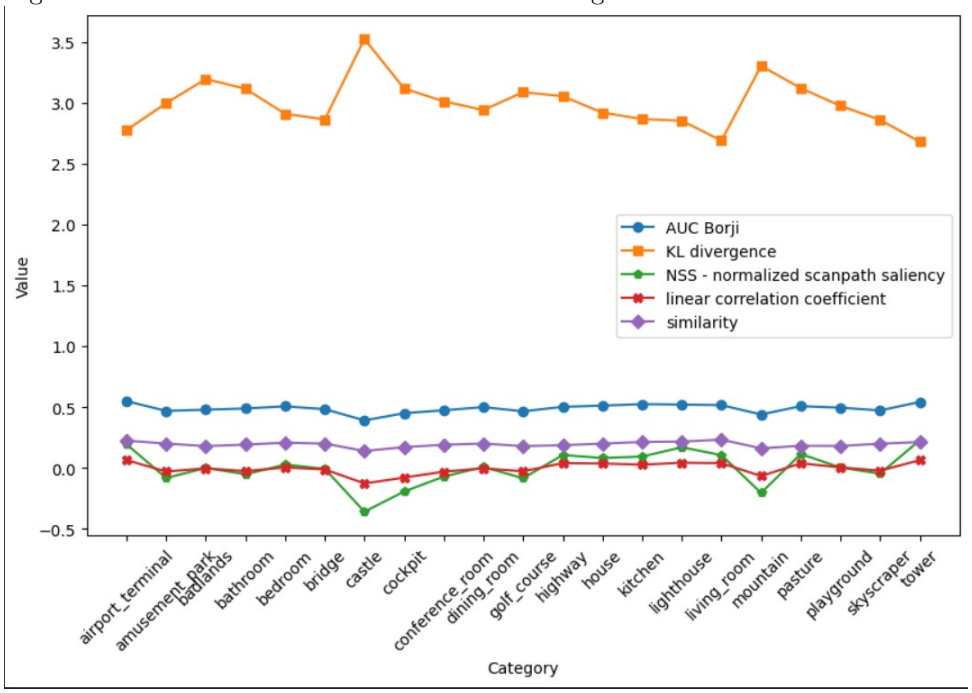


2.2 Non deep model - Achanta

For the non-deep model we chose the Achanta model. It is based on low-level image features and has been widely used in computer vision for saliency detection

tasks. The model focuses on extracting spectral residual and spatial intensity contrast features to identify salient regions in images. It gets the spectral features by transforming the image into the frequency domain and emphasizes spectral residuals. Achanta model achieves computational efficiency and simplicity by relying on handcrafted features, so in situations where deep-learning models are expensive, it is the go-to option.
The Achanta model gave suboptimal results. Non-deep models such as Achanta, are really sensitive to the content and characteristics of the images. These models, relying on handcrafted low-level features, might excel in certain scenarios, such as natural scenes, but struggle when faced with more complex or abstract images, as in the FIGRIM dataset. The FIGRIM dataset contains a mix of outdoor and indoor images, which all are high-resolution and rather complex images. Its performance therefore falls short in scenarios such as this one, with intricate relationships and complex scenes.
The performance of the Achanta model is shown in the figure below.

Figure 2.2: Different metrics evaluated over all categories for the Achanta model

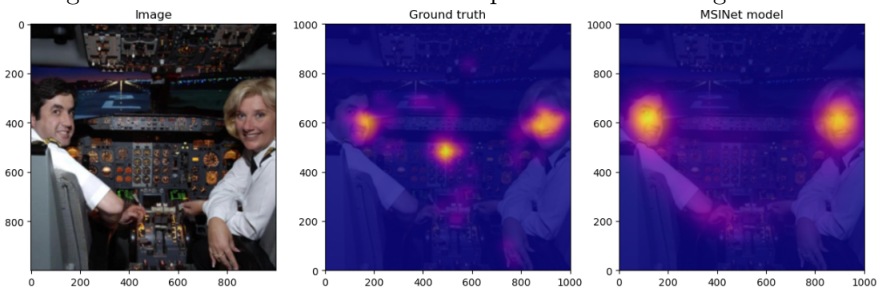


2.3 Deep model - MSINet

For the deep model we chose the MSINet trained on the SALICON dataset. It is based on a convolutional neural network (CNN) architecture and is trained on

large-scale eye-tracking datasets, and it learns hierarchical features from input images to predict saliency maps. Additionally, the MSINet model is trained on large-scale eye-tracking datasets where images are annotated with fixations (gaze points). We used the implementation available on the provided links.
The model is good at predicting where humans look, and recognizes faces, text and other important features that humans focus on pretty well. There are still images where the model struggles to predict all high clustering areas, particularly small text, occluded faces and center bias (Figure 2.3).

Figure 2.3: The MSINet model didn't predict the central gaze track



The scores of different metrics are shown in Figure 2.4. There is no significant difference between different categories, and no pattern as to which categories have a higher or lower metric.

2.4 Comparison of deep and non-deep model

When comparing the two models, we can very easily see that the deep model is better at predicting gaze tracks. An example is shown in Figure 2.5. In terms of computational time, the Achanta model is very fast, while the MSINet model takes about 1-2 seconds to generate a saliency map for an image.

Figure 2.4: Different metrics evaluated over all categories for the MSINet model

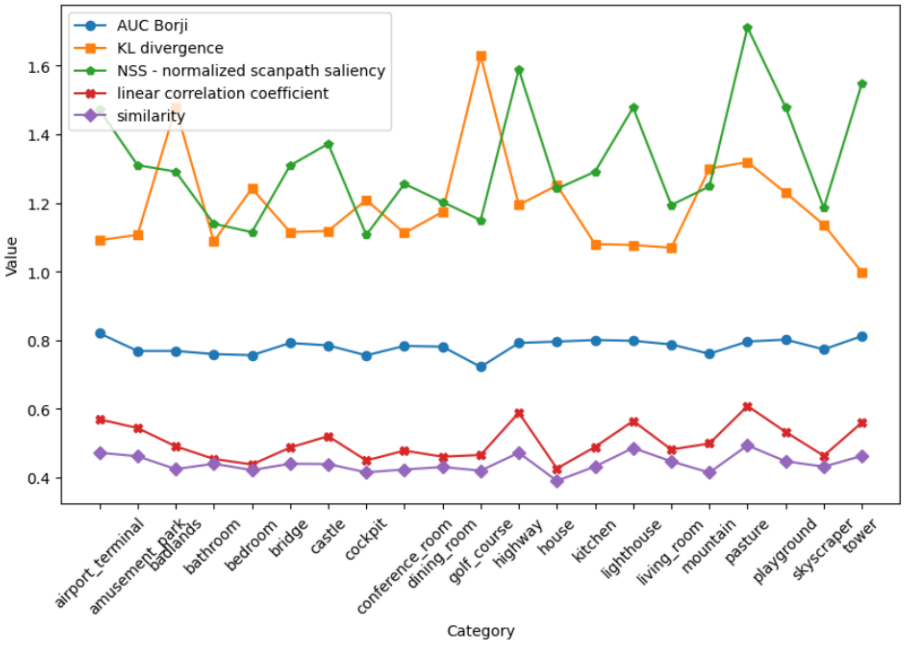


Figure 2.5: Comparison of Achanta and MSINet models