UCD Forensics and
Security Research Group

Geolocation of Human Trafficking Images: A Perceptual Color-based Image Retrieval Approach

Jessica Herrmann, MSc

Supervisor: Assoc. Prof. Mark Scanlon

Personal Background

- Police Officer since 2008
- Cybercrime Investigator at the LKA Baden-Württemberg since 2020
- Since 2021 mostly investigations in the field of CSA
- Actually working in an AI project

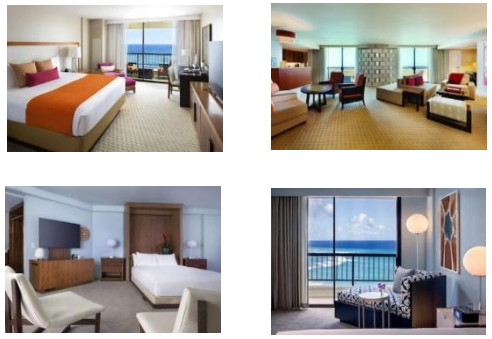
Research Background

Why? To evaluate the **effectiveness of colour-based descriptors** for **Content-Based Image Retrieval** (CBIR) in **human-trafficking investigations.**

How? Perceptual approach using the Hotels-50k dataset

Hotels-50K:

Sources: Travel Websites
from hotels worldwide



1.027.871 Images
50.000 Hotels

Source: TraffickCam App
(images submitted by travelers to help combat trafficking), which are more visually similar to images from trafficking investigations.

17.954 Images
5.000 Hotels

Colour Extraction

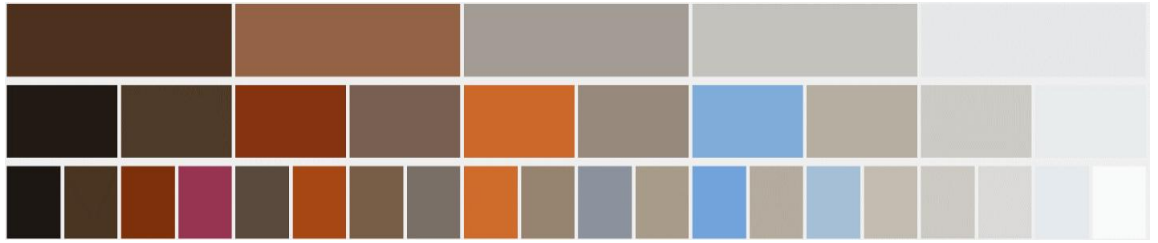
Criteria for Initial Colour Extractions

- Leveraged python colour palette extraction tool, Pylette[1]
- Extracted **5, 10 and 20 dominant colours** per image
- Used the **RGB colour model (Red, Green, Blue)**.
- Used **Median-Cut** to select **the dominant colours**
- Sorted by **frequency.**

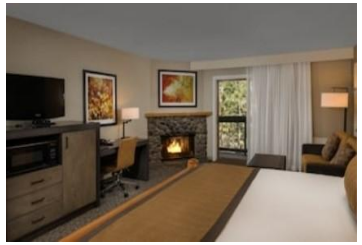
[1] https://github.com/qTipTip/Pylette

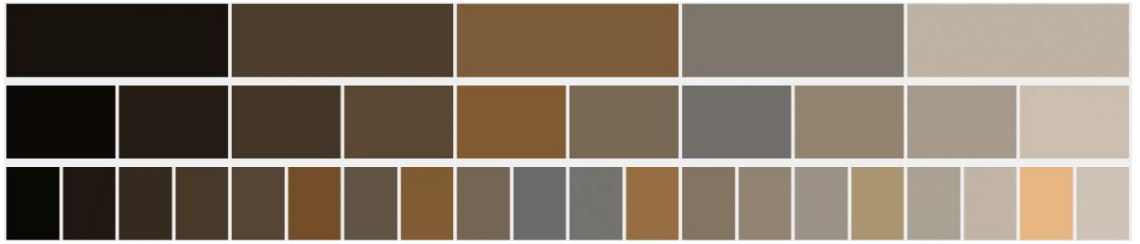
Colour Extraction

451, Hyatt Regency Waikiki Beach Resort & Spa



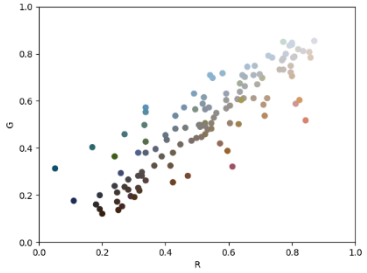
398, The Riverhouse Hotel & Convention Center

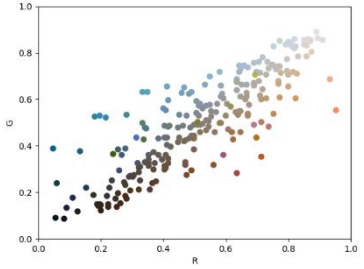


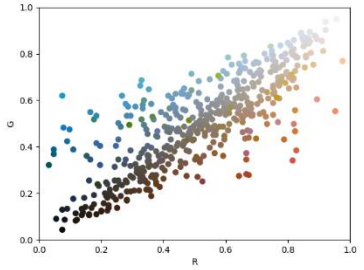


Scatter Plots 2D & 3D

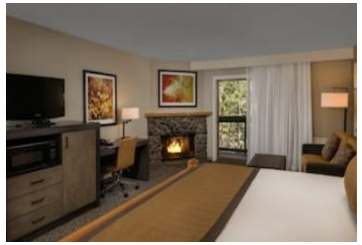
451, Hyatt Regency Waikiki Beach Resort & Spa

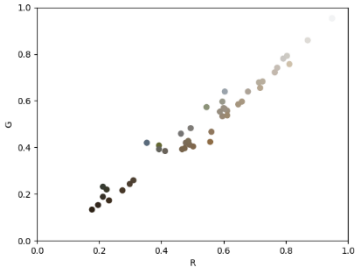


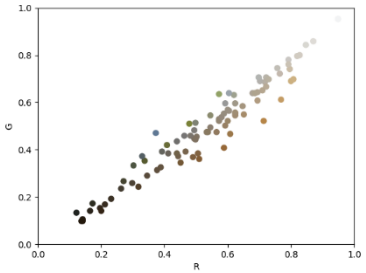


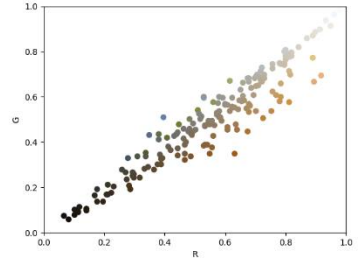


398, The Riverhouse Hotel & Convention Center

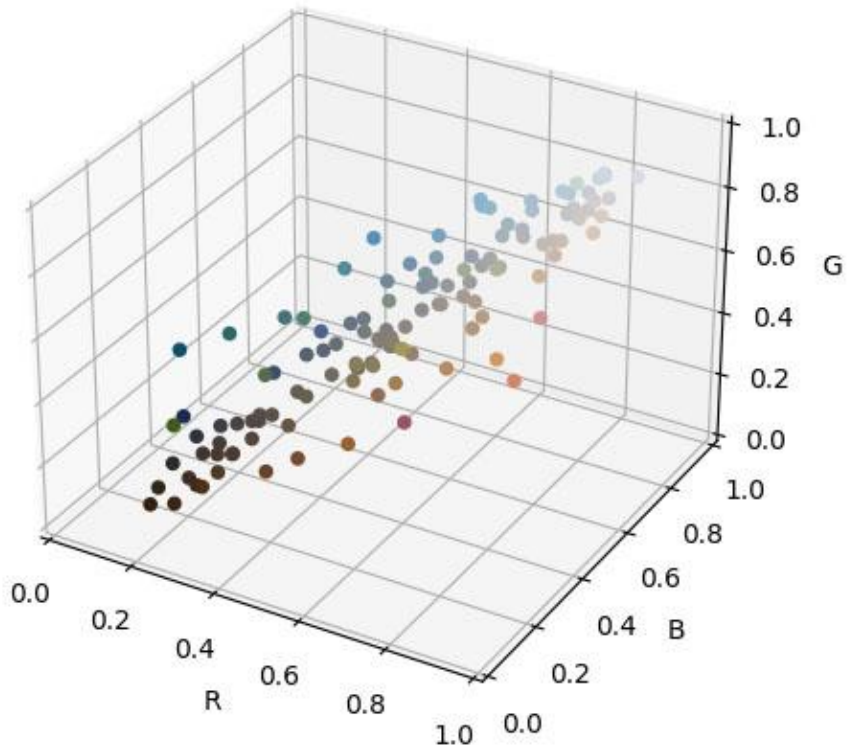








Curse of Dimensionality

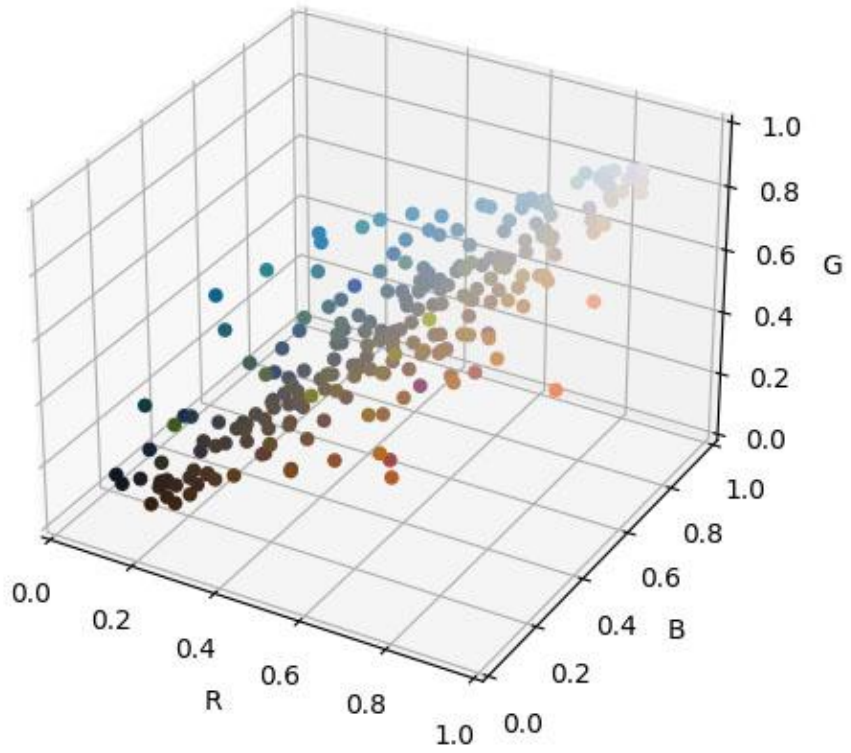


Increasing the number of extracted colours adds dimensional complexity

Too few colours: Risk of losing important visual information

5 Values

Curse of Dimensionality



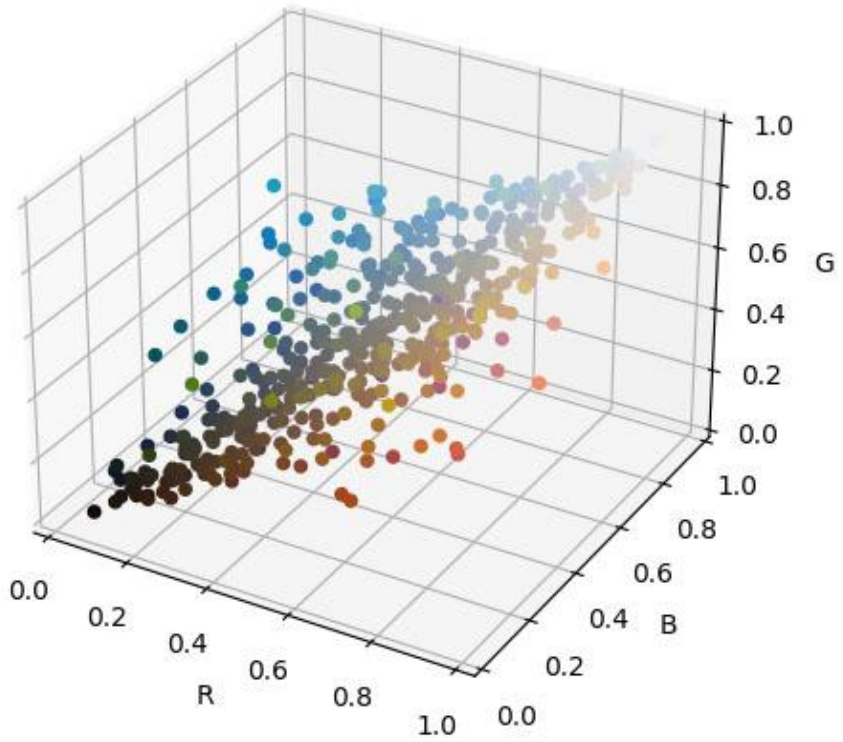
Increasing the number of extracted colours adds dimensional complexity

Too few colours: Risk of losing important visual information

Too many colours: Risk of Over Fitting due to the curse of dimensionality

10 Values

Curse of Dimensionality



Increasing the number of extracted colours adds dimensional complexity

Too few colours: Risk of losing important visual information

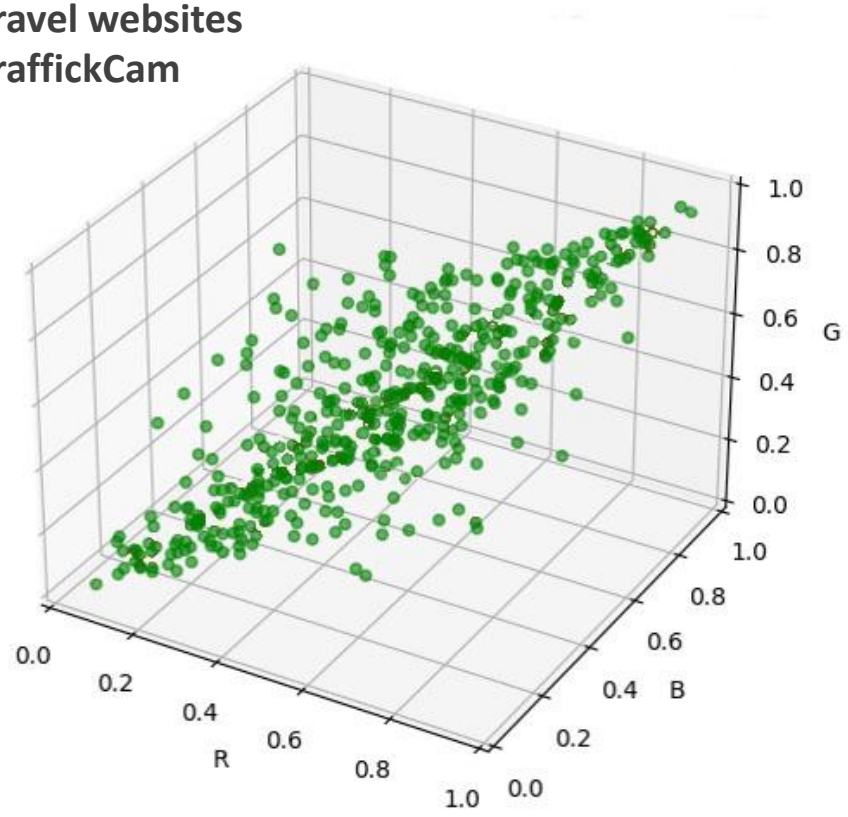
Too many colours: Risk of Over Fitting due to the curse of dimensionality

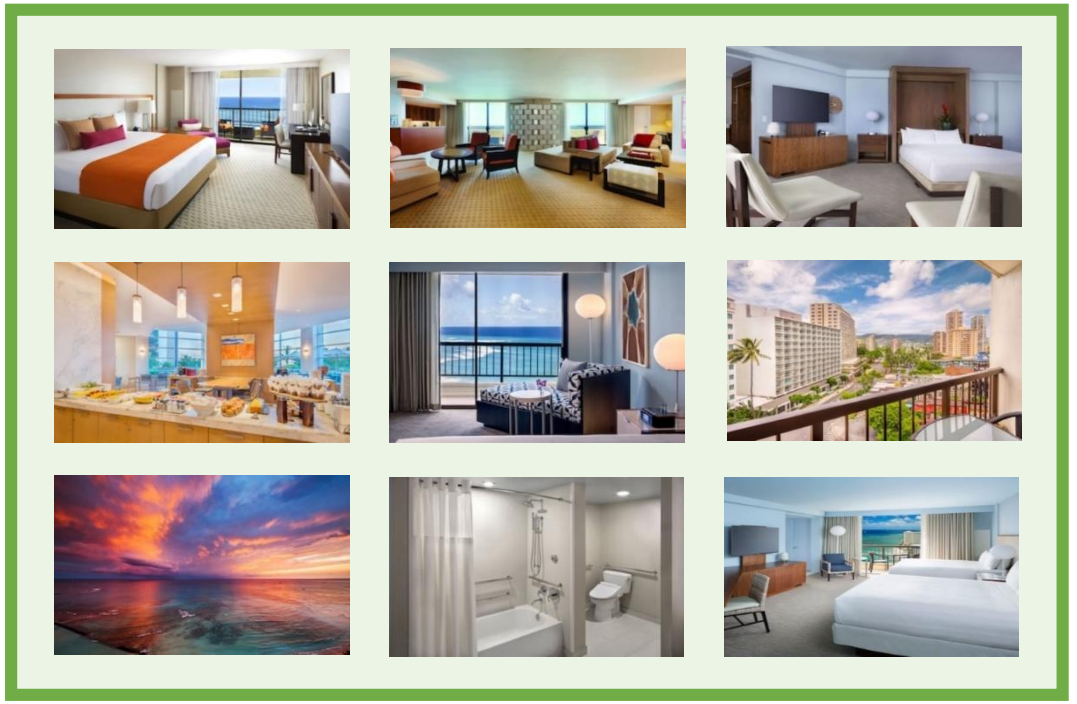
- **Finding the right number of colour values is crucial to balance detail and computational complexity**

20 Values

Comparing Trainsets and Teststes

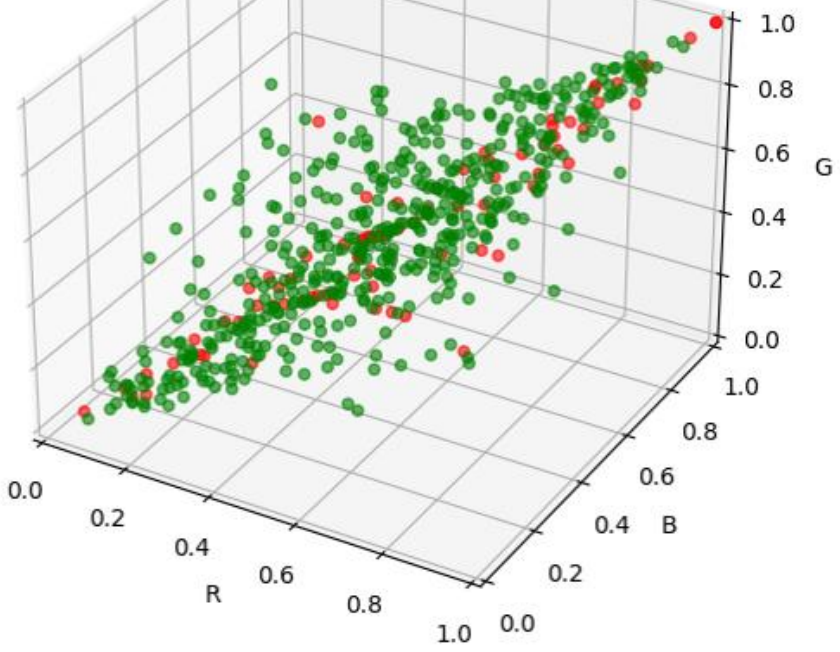
- Travel websites
- TraffickCam





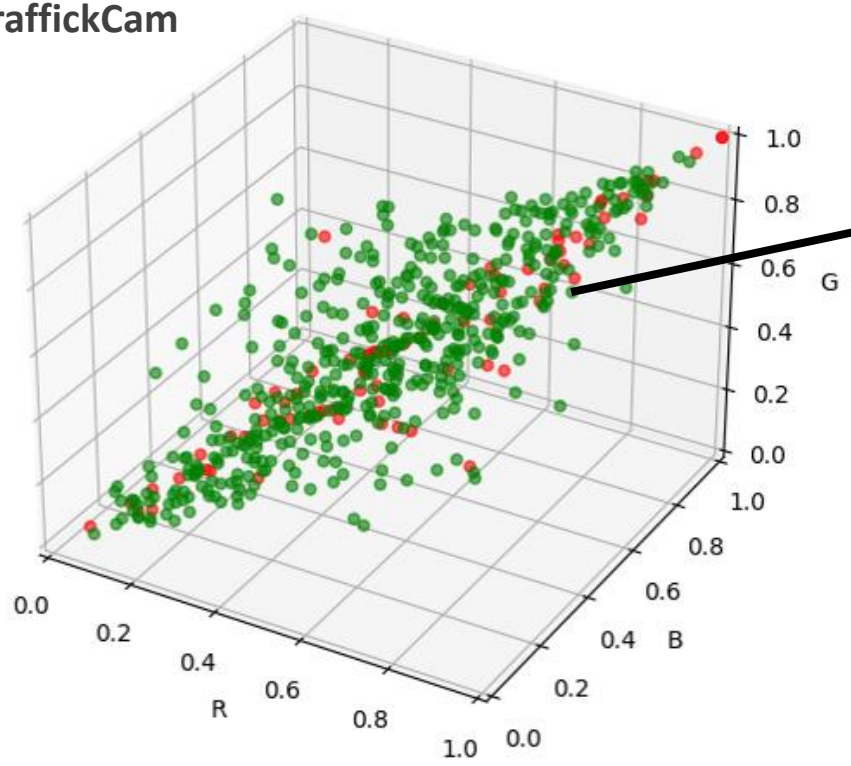
Comparing Trainsets and Teststes

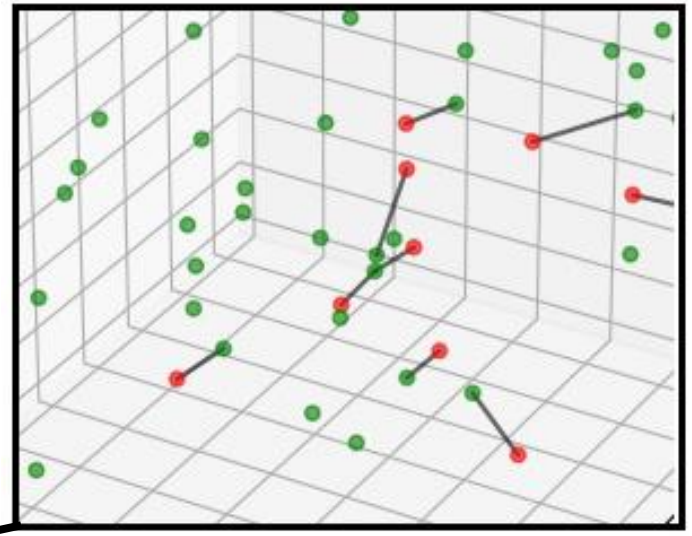
- Travel websites
- TraffickCam



Euclidean Distance

- Travel websites
- TraffickCam





- Measures the straight-line distance between two points in RGB space.
- Calculated as the square root of the sum of squared differences between the corresponding color values (Red, Green, Blue).

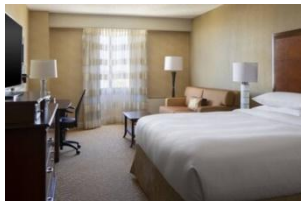
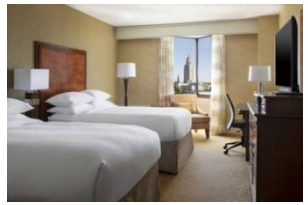
$$d = \sqrt{(R_2 - R_1)^2 + (G_2 - G_1)^2 + (B_2 - B_1)^2}$$

Dataset 1: „Unclean“

3 x 100 different hotels

Travel Websites (Train)



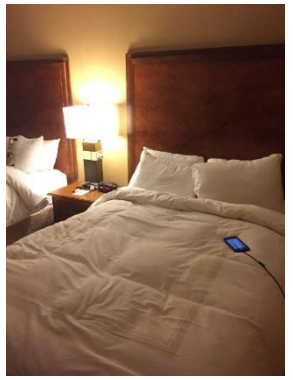









TraffickCam (Test)




Dataset 1: „Unclean“

Euclidean distances

Extracted Values	Top-10	Top-20	Top-30	Top-40	Top-50	Top-60	Top-70	Top-80	Top-90	Top-100
1 Value	11,33	22,67	30,33	43,33	53,33	63,67	71,33	81,33	91,67	100,00
2 Values	10,33	22,67	32,67	44,33	52,33	62,00	74,00	84,00	91,67	100,00
5 Values	12,33	23,00	33,00	43,67	52,67	63,67	72,67	83,00	90,33	100,00
10 Values	12,33	22,00	33,00	42,67	51,67	62,67	71,33	82,33	90,00	100,00
20 Values	13,00	22,67	35,00	45,00	54,00	63,33	70,33	82,00	90,33	100,00

Summary:

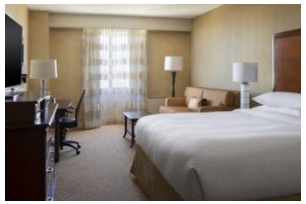
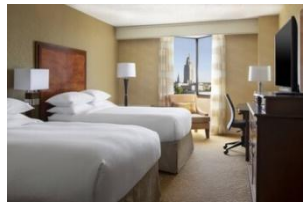
- **Top-100 accuracy exceeded** the **95% threshold**, but **Top-50 accuracy** was **only** about **50%**.
- The **inclusion** of **out-of-scene images** (e.g., tourist images with sea views) likely **caused** a **high false positive rate**.

Dataset 1: „Unclean“

3 x 100 different hotels

Travel Websites (Train)



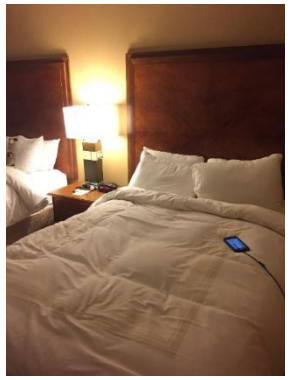









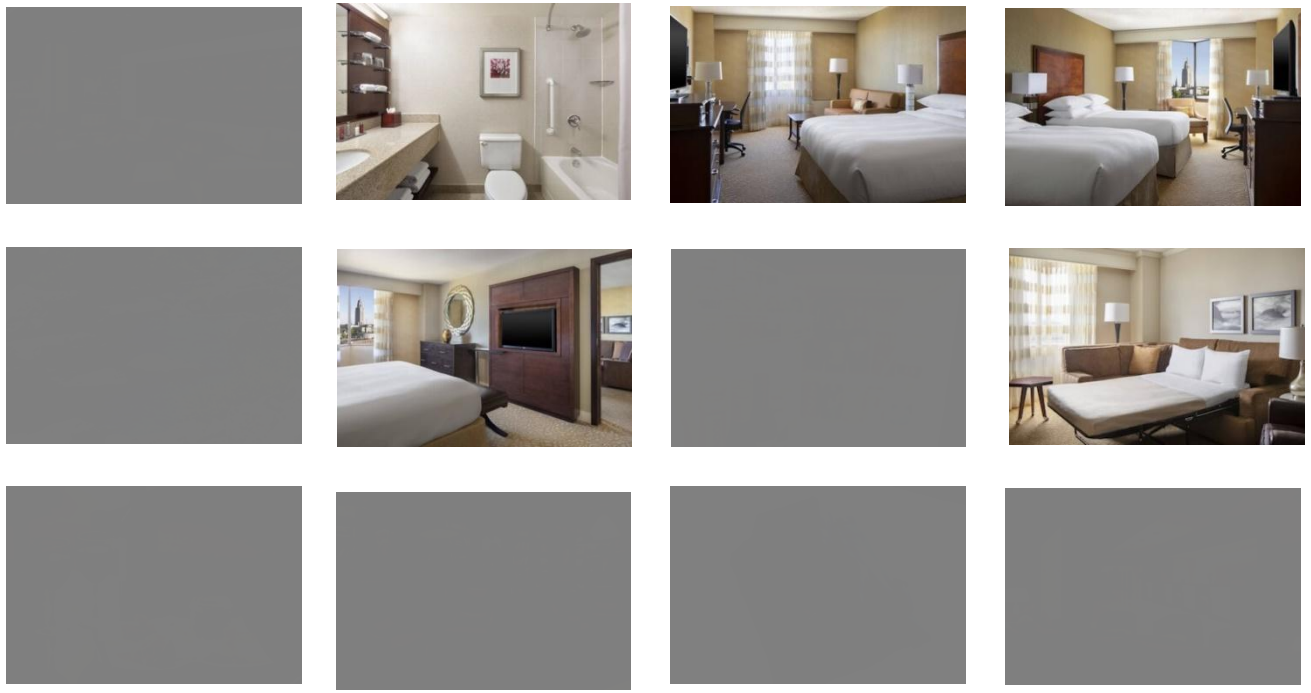
TraffickCam (Test)



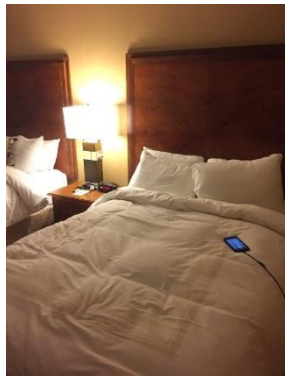

Dataset 2: „Clean“

“Clean” datasets = “Unclean” datasets – TraffickCam and unwanted scenarios

Travel Websites (Train)



TraffickCam (Test)

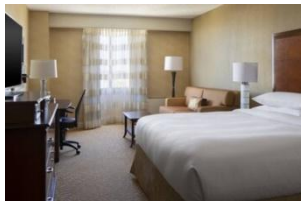
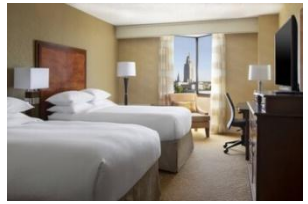


Dataset 2: „Clean“

“Clean” datasets = “Unclean” datasets – TraffickCam and unwanted scenarios

Travel Websites (Train)



Travel Websites (Test)

Dataset 2: „Clean“

Euclidean distances

Extracted Values	Top-10	Top-20	Top-30	Top-40	Top-50	Top-60	Top-70	Top-80	Top-90	Top-100
1 Value	42,00	62,00	72,67	81,00	89,33	92,67	95,67	97,33	99,33	100,00
2 Values	48,33	66,67	75,33	83,00	88,00	92,33	96,00	98,00	99,33	100,00
5 Values	44,33	59,67	68,00	78,00	85,00	90,33	94,67	96,67	99,00	100,00
10 Values	42,33	60,00	70,00	79,33	86,67	90,67	93,33	96,33	98,67	100,00
20 Values	43,33	57,33	69,33	80,67	87,00	92,00	96,00	97,33	99,00	100,00

Summary:

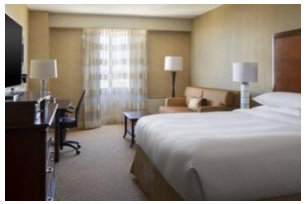
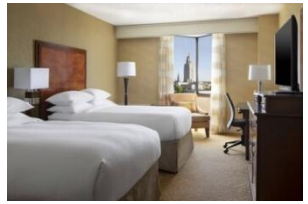
- Significant improvement; **Top-70 accuracy** reached **95 %** for both **2** and **20 descriptors.**

Dataset 2: „Clean“

“Clean” datasets = “Unclean” datasets – TraffickCam and unwanted scenarios

Travel Websites (Train)



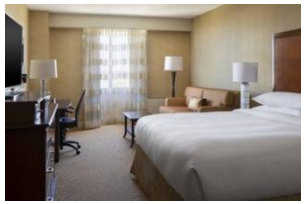
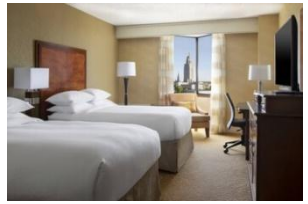
Travel Websites (Test)

Dataset 3: „Three Beds“

“Three Beds” datasets = Optimised selection out of the “Clean” datasets

Travel Websites (Train)



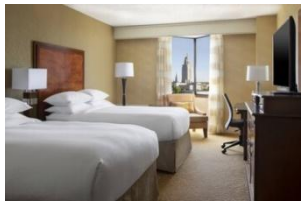
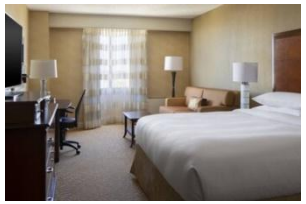



Travel Websites (Test)

Dataset 3: „Three Beds“

“Three Beds” datasets = Optimised selection out of the “Clean” datasets

Travel Websites (Train)

Travel Websites (Test)

Dataset 3: „Three Beds“

Euclidean distances

Extracted Values	Top-10	Top-20	Top-30	Top-40	Top-50	Top-60	Top-70	Top-80	Top-90	Top-100
1 Value	51,00	68,33	80,33	88,00	92,33	95,33	96,67	99,00	99,67	100,00
2 Values	62,00	74,33	84,33	92,00	95,67	96,67	97,00	99,33	99,67	100,00
5 Values	55,33	66,33	77,00	83,00	89,67	93,33	96,00	98,33	99,33	100,00
10 Values	55,67	72,00	82,67	88,67	94,00	96,00	97,00	98,67	99,33	100,00
20 Values	53,33	70,33	80,33	86,67	90,67	94,33	96,00	98,67	99,67	100,00

Summary:

- **2 descriptors** achieved the best results with a **Top-50 accuracy over 95 %.**

Key Insights

Colour palettes can **greatly improve** the accuracy of hotel room identification.

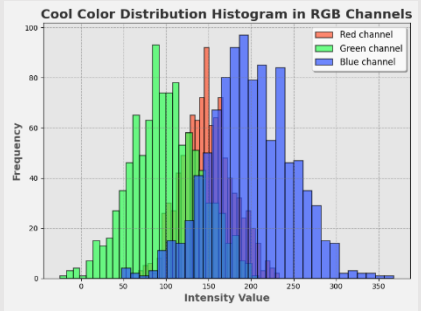
For **Euclidean Distance** the **2-descriptor** model was the **most reliable**, offering high accuracy with minimal complexity.

Cleaning the **datasets** (removing out-of-scene images) had a **significant impact** on accuracy.

Too many descriptors introduced the risk of the **curse of dimensionality**.

Further Improvements

Explore a **histogram-based analysis** to understand **color distribution** and introduce **weighting** within color spectra.



Further classification of training and test datasets based off **perceptual context**, such as bed, bathroom, & minibar.

Incorporating **texture-based methods.**

Exploring **alternative color spaces,** such as HSV/HLS.

White balance in the RGB space.

Test under **real-world conditions** by incorporating **occlusion scenarios.**

Advantages

Independent of different angles or perspectives in room images.

Publicly available dataset.

Reproducible and **traceable** method, particularly in light of regulatory frameworks like the **AI Act**, where transparency and explainability of results are key requirements.

Methods are **highly flexible** and can be adapted to different datasets and applications by adjusting the **number of descriptors** and the **distance metrics.**

Computationally simple, making it easy to implement and scale across large volumes of images.

UCD Forensics and
Security Research Group

Let‘s dive deeper!

Jessica Herrmann

MSc FCCI | Cybercrime Expert | AI Project Office KIRKE | Police Baden-Württemberg