Universidad Politécnica
de Madrid

**Escuela Técnica Superior de
Ingenieros Informáticos**

Master in Data Science

# NON-TEXTUAL DATA EXTRACTION

Javier Arteaga Puell
Rodrigo Castañón Martínez
Dakota James Mellish

Madrid, February 2025

## 1. Motivation

Content-Based Image Retrieval (CBIR) is an approach used in searching and classifying images based on visual features such as color, texture, and shape. Unlike traditional metadata-based retrieval methods, CBIR allows direct analysis of the visual information contained in images, which is essential in applications such as pattern recognition, computer vision, and multimedia retrieval (Datta et al., 2008).

This project applies CBIR techniques to identify and compare basketball jerseys by analyzing both their color distributions and structural features. The retrieval system is designed to improve accuracy by first filtering images using color histograms and then refining results through shape-based descriptors such as HOG and Harris corner detection or vice versa. This combination enables more precise jersey matching, taking into account both visual similarity and distinctive text-based details like team names and numbers.

By leveraging this multi-stage approach, the project aims to provide a more robust and efficient solution for automated jersey recognition, sports analytics, and merchandise classification. Different color spaces and similarity metrics have been explored to enhance retrieval performance, ensuring the system effectively differentiates jerseys with similar colors but distinct textual or structural characteristics (Jain, Duin, & Mao, 2000).

The images used for this assignment where extracted from the Basketball Jersey Archive

## 2. State of the art

### 2.1 Color Space

Color spaces are mathematical representations that describe the colors in an image. In this project, the following models have been used:

**RGB (Red, Green, Blue)**: An additive model where colors are represented by the combination of three primary components. It is commonly used in digital image processing (Gonzalez & Woods, 2018).

**HSV (Hue, Saturation, Value)**: This model separates brightness information from color, facilitating tasks such as segmentation and hue comparison (Trimeche et al., 2017).

### 2.2 Distance Measurement

To compare image similarity, two distance measures have been employed:

- **Euclidean Distance**: Measures the difference between the feature vectors of two images. It is defined as (Duda, Hart, & Stork, 2001):

$$d(A, B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

- **Chi-Square Distance**: Measures the difference between two histograms by weighting the magnitude of each difference with the sum of the corresponding bin values:

$$\chi^2(A,B) = \sum_{i=1}^{n} \frac{(A_i - B_i)^2}{A_i + B_i + \epsilon}$$

In this project, the Chi-Square distance has been used to compare the color histograms extracted from the processed images.

## 3. Implementation of the toy CBIR

### 3.1 'Smart' Histogram Descriptor

A histogram descriptor represents the color distribution in an image by counting the frequency of different color values. In this project, we use histograms based on the RGB and HSV color spaces to capture different aspects of color information.

To improve robustness, the following preprocessing steps were applied:

- Color conversion: The images were converted from BGR (OpenCV default) to RGB and HSV.
- Histogram normalization: Each histogram was normalized to account for different image sizes and illumination variations.
- Histogram binning: A predefined number of bins was used to balance detail and computational efficiency.

The histogram descriptor is computed as follows:

$$H(c) = \sum_{i=1}^{N} \delta(I_i = c)$$

where H(c) is frequency of color c, Ii is color value at pixel iii, N is total number of pixels.

### 3.2 HOG+Harris

To capture shape and texture information, we use a combination of Histogram of Oriented Gradients (HOG) and Harris Corner Detection:

- **HOG Descriptor**: Extracts gradient orientation histograms from an image, highlighting edge structures and shapes. In this project, we use OpenCV's 'cv2.HOGDescriptor()' to compute these features on grayscale images.

- **Harris Corner Detection**: Identifies points in an image where the intensity changes significantly in multiple directions. The function 'cv2.cornerHarris()' is applied, followed by thresholding to select the most prominent corners.

These descriptors complement color histograms by providing structural and textural information, enhancing retrieval performance.

### 3.3 Distance Calculation

To compare images, a similarity measure is needed. This project implements the Chi-Square distance, which is effective for comparing histograms:

$$\chi^2(A,B) = \sum_{i=1}^{n} \frac{(A_i - B_i)^2}{A_i + B_i + \epsilon}$$

where A and B are the histograms of two images, and ϵ\epsilonϵ prevents division by zero. This method allows for efficient comparison of images based on their color distributions.

### 3.4 Combination of Both Descriptors

To improve retrieval accuracy, a hybrid approach was implemented by first analyzing color similarity using histograms and then refining the results using shape and text-based features and vice versa. The final similarity score is determined by the sum of both normalized distances this way This method ensures that initial filtering is efficient while maintaining high accuracy by incorporating structural features in the final comparison. By first reducing the search space with color histograms and then applying shape and text-based analysis and vice versa, the system effectively retrieves images that are both visually and semantically similar.

### 3.5 Indexing and Searching in CBIR

The implemented system performs indexing and content-based image retrieval (CBIR) using two main approaches: color analysis and text feature extraction. During the indexing process, all images from the *img* folder are loaded and preprocessed by removing the background. Each image is divided into sub-images, and color histograms are extracted from the BGR channels. For text analysis, OCR is used to detect the text region on the jersey, and features are extracted using HOG descriptors and Harris Corner detection. These features are stored as vectors for efficient comparison.

For the search and comparison process, image similarity is determined by comparing color histograms using the Chi-Square distance metric. Simultaneously, text features are compared using the Euclidean distance between the extracted HOG and Harris descriptors. Both metrics are then normalized and combined to enhance the accuracy of similarity detection. The results are ranked based on similarity scores, and the top five most similar jerseys are visually displayed. This approach enables efficient retrieval of basket jerseys based on their design and typography, optimizing the search process within large image datasets.

## 4. References

Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, *40*(2), 1-60.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4-37.

Gonzalez, R. C., & Woods, R. E. (2018). *Digital image processing*. Pearson.

Trimeche, A., Ziou, D., & Laanaya, A. (2017). A survey on color spaces and their application for skin detection. *Multimedia Tools and Applications*, *76*, 13279-13302.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. John Wiley & Sons.