Feature-Based Fashion Accessory Clustering: A Traditional Computer Vision Approach for Product Categorization

Final Project Report

KARAOGUZ Oguz **PODDAR Karam** **RAVELLO Riccardo**

Abstract

This project demonstrates a feature-based approach to clustering fashion accessory images without using deep learning techniques. We developed a comprehensive pipeline that combines background removal, SIFT key point detection, and extraction of shape, colour, and texture features with differential weighting to create visually coherent product groupings. Applied to 750 men's accessories from the Fashion Product Images dataset, our method successfully distinguished between different product types using K-means and hierarchical clustering. While statistical metrics favoured more straightforward partitioning (k=2), visual inspection revealed that alternative cluster counts (k=5 for K-means, k=3 for hierarchical) yielded more practical and meaningful groupings. Our approach achieves reasonable computational efficiency (2.36 seconds per image) and demonstrates that traditional computer vision techniques remain valuable for fashion retail applications where labelled training data may be limited.

1. Introduction

The fashion industry faces significant challenges in organising and managing extensive collections of product images, particularly for accessories, which often have subtle visual distinctions yet need precise categorisation. With e-commerce platforms requiring thousands of product images and fashion retailers continuously updating their inventories, efficient image organisation becomes critical for business operations. Currently, most image organisation systems rely on manual tagging or deep learning approaches, both of which have limitations. Manual tagging is labour-intensive and prone to inconsistencies, while deep learning methods can be computationally expensive and require substantial training data.

Several key questions guide our project:

1. How effectively can traditional feature extraction methods identify meaningful visual patterns in fashion accessory images?
2. Can these extracted features provide sufficient discrimination between different accessory categories?
3. How does the performance of K-means clustering compare to hierarchical clustering for this domain?
4. What is the optimal balance between computational efficiency and clustering accuracy?

The significance of this problem extends beyond mere organisation. Efficient image clustering can:

1. Enhance searchability in product databases, allowing designers and merchandisers to locate similar accessories quickly
2. Improve recommendation systems by identifying visually similar products for customers
3. Reduce redundancy in product catalogues by identifying nearly identical items

2. Problem Definition

We define our problem as an unsupervised learning task that aims to partition a set of fashion accessory images into visually coherent clusters without prior category information. Given a set of pictures in $I = \{I_1, I_2, \dots, I_n\}$ the Field, our goal is to assign each image to one of the k clusters $C = \{C_1, C_2, \dots, C_k\}$ such that images with similar visual characteristics belong to the same cluster.

For each image, we need to extract distinctive visual features that capture the essence of the fashion accessories. The feature extraction must be robust to lighting, pose, scale, and background variations. We denote the feature vector for image I_i as F_i , representing a point in a high-dimensional feature space.

The clustering problem can be formulated as an optimisation task to maximise inter-cluster distance and intra-cluster similarity. We formalise this using the silhouette score *S*, defined as:

$$S = \frac{1}{n}\sum_{i=1}^{n} \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Where *a(i)* is the average distance between image *i* and all other photos in the same cluster, and *b(i)* is the minimum average distance between image *i* and images in any other cluster.

A critical aspect of the problem is determining the optimal number of clusters *k* without prior knowledge, which is challenging in unsupervised learning settings. This requires evaluating clustering quality metrics across different values of *k* to identify the most appropriate partition of the feature space.

The problem involves several constraints and challenges:

1. Computational efficiency constraint**:** The algorithm must process large image collections with reasonable time complexity
2. Feature invariance constraint**:** The extracted features should be invariant to rotation, scaling, and minor illumination changes typical in product photography
3. Background complexity**:** Fashion accessory images often include complex backgrounds or models wearing the items
4. Non-semantic gap challenge**:** Bridging the gap between low-level visual features and higher-level semantic categories
5. Unsupervised nature: No labelled training data is available to guide the clustering process

The computational hardness of this problem stems from multiple factors. The high dimensionality of image data makes direct comparison difficult and necessitates sophisticated feature extraction techniques. Additionally, clustering in high-dimensional spaces is complicated by the "curse of dimensionality," where distance measures become less meaningful as dimensionality increases. Finally, evaluating cluster quality without ground truth labels requires careful analysis of internal validation metrics that may not always align with human perceptual judgments of similarity.

3. Related Work

Image clustering, particularly for fashion products, has seen significant developments. Our work builds upon several key research areas and methodologies while offering distinctive approaches to the specific challenge of accessory categorisation.

3.1 Feature Extraction Techniques

Scale-Invariant Feature Transform (SIFT) introduced by Lowe [1] has proven to be one of the most robust approaches for detecting and describing local features in images. Lowe's method enables identifying key points invariant to scaling and rotation and partially invariant to illumination changes, which is crucial for identifying fashion accessories across various presentations. Our implementation leverages SIFT's capabilities for initial feature detection, though we extend beyond Lowe's original approach with domain-specific enhancements.

Harris corner detection, developed by Harris and Stephens [2], provides complementary feature extraction capabilities by identifying corners and interest points based on local image gradients. While SIFT excels at capturing distinctive features at various scales, Harris detection is particularly effective at identifying structural elements in accessories like watches, glasses, and jewellery. Our work combines these approaches to capture a more comprehensive set of visual features.

In the domain of colour and texture representation, Duda et al. [3] ‘s work on pattern classification provides foundational techniques for feature quantization and histogram-based representations. Their approaches inform our colour and texture feature extraction methodology, particularly in creating HSV histograms and texture descriptors crucial for distinguishing between visually similar accessories.

3.2 Clustering Methodologies

Jain [4] extensively analysed K-means clustering in his comprehensive review of clustering algorithms, and it remains one of the most widely used techniques for unsupervised learning. Jain's work highlights its strengths and limitations, particularly its sensitivity to initialisation and the challenge of determining optimal cluster numbers. Our implementation addresses these limitations through careful initialisation strategies and multiple evaluation metrics to determine the optimal number of clusters.

Hierarchical clustering approaches, as discussed in the pattern recognition literature [3], offer an alternative perspective on grouping similar items without requiring a predefined number of clusters. Our comparative analysis between K-means and hierarchical clustering extends the discussion in these foundational works by examining their relative performance specifically for fashion accessory categorisation.

3.3 Fashion Image Analysis

Recently, Liu et al. [5] proposed DeepFashion, a large-scale fashion database with annotations for various fashion-related tasks, including category and attribute prediction. While DeepFashion represents a significant advance in fashion image analysis, it relies heavily on deep learning techniques and extensive labelled data. Our approach differs by focusing on traditional computer vision methods that require no training data, making it more accessible for applications with limited computational resources or training datasets.

Hsiao and Grauman [6] introduced a method for creating fashion compatibility recommendations based on visual features. Their work demonstrates the value of understanding visual relationships between fashion items, which aligns with our goal of clustering similar accessories. However, while their approach focuses on cross-category compatibility, our work emphasises intra-category similarity for organisation and retrieval purposes.

3.4 Background Removal and Object Isolation

Rother et al. [7] developed "GrabCut," an interactive foreground extraction technique that influenced subsequent object isolation approaches. While GrabCut requires user interaction, our automated background removal process draws inspiration from its use of iterative graph cuts, adapting the concept to fashion product images through thresholding, morphological operations, and skin detection.

3.5 Our Contribution

Our work differs from previous research in several key aspects:

1. We focus on fashion accessories, which present unique challenges due to their small size, intricate details, and often complex presentations (worn by models or displayed against varied backgrounds).
2. Unlike most general-purpose image clustering systems that treat all features equally, we implement a multi-feature approach that gives higher weights to colour and shape features, which are particularly important for accessory discrimination.
3. We developed a specialised background removal technique incorporating skin detection to handle cases where models wear accessories, a common scenario in fashion photography.
4. While many recent approaches rely exclusively on deep learning, our method demonstrates that traditional computer vision techniques can still achieve effective clustering for specialised domains with the proper feature engineering.
5. We provide a comparative analysis of K-means and hierarchical clustering specifically for fashion accessory images, offering insights into which approach may be more suitable for different organisational goals.

By combining these elements, our work contributes to understanding how traditional computer vision techniques can be effectively applied to practical fashion inventory organisation problems. It complements the recent surge in deep learning approaches while offering a more accessible alternative for scenarios with limited computational resources or training data.

4. Methodology

4.1 Data Collection and Preprocessing

For this project, we utilised the Fashion Product Images dataset from Kaggle, which contains many product images across multiple categories. We specifically targeted men's accessories to focus our analysis and avoid excessive complexity. The dataset included 4,413 men's accessory images spread across six subcategories:

{Socks, Watches, Sandals, Lips (lip care products), Eyes (eyewear), and Gloves}

We randomly sampled 750 images from this subset for computational efficiency and practical demonstration.

The product images in the dataset feature varying backgrounds, lighting conditions, and presentation styles. Some accessories are shown in isolation against clean backgrounds, while others are worn by models or displayed with props. This heterogeneity presents a significant challenge for clustering algorithms, necessitating robust pre-processing and feature extraction methods.

4.2 Image Preprocessing Pipeline

Our approach begins with a comprehensive pre-processing pipeline designed to isolate the accessory from its background, creating a standardised representation for feature extraction:

1. **Grayscale Conversion and Noise Reduction**: Each RGB image is converted to grayscale, and a Gaussian blur is applied to reduce noise while preserving critical structural details.
2. **Automatic Thresholding**: We employ Otsu's method for automatic thresholding, which determines an optimal threshold value by minimising the intra-class variance between foreground and background pixels. Mathematically, Otsu's method selects the threshold *t* that maximises:
$$\sigma_B^2(t) = \omega_1(t)\omega_2(t)[\mu_1(t) - \mu_2(t)]^2$$
Where $\omega_1(t)$ and $\omega_2(t)$ are the probabilities of the two classes separated by threshold *t*, and $\mu_1(t)$ $\mu_2(t)$ are their means.
3. **Morphological Operations**: We apply opening and closing operations to refine the binary mask, eliminating minor artefacts and filling holes:
$$A \circ B = (A \ominus B) \oplus B (\text{opening})$$

$$\boxed{A \cdot B = (A \oplus B) \ominus B(\text{closing})}$$

Where $\oplus$ and $\ominus$ denote dilation and erosion operations, respectively, and B is a structuring element.

4. **Contour Analysis**: We identify the most prominent contour in the binary mask, assuming it corresponds to the main accessory. This helps separate the primary object from more minor background elements.
5. **Skin Detection**: Since models often wear many accessories, we implement skin detection in the YCrCb colour space using the following thresholds: Y: [0, 235], Cr: [133, 173], Cb: [77, 127]. This creates a skin mask that is subsequently inverted and combined with the object mask to reduce the influence of skin regions on feature extraction.
6. **Background Replacement**: Finally, we replace the background with white while preserving the foreground object, resulting in a clean, standardised image for feature extraction.

4.2 Feature Extraction

Our feature extraction framework employs multiple complementary approaches to capture diverse aspects of each accessory:

SIFT Features: We implement the Scale-Invariant Feature Transform described by Lowe (2004). SIFT detects key points by:

- Constructing a scale space using Difference of Gaussian (DoG) filters
- Locating maxima/minima in the DoG scale space
- Eliminating low-contrast and edge responses
- Assigning orientations based on local image gradients
- Computing descriptors as normalised histograms of gradients
- Each SIFT descriptor is a 128-dimensional vector representing the local appearance around a key point.

Shape Features: We extract five key shape metrics:

- Aspect ratio: width/height of the bounding rectangle
- Extent: contour area/bounding rectangle area
- Solidity: contour area/convex hull area
- Equivalent diameter: diameter of a circle with the same area as the contours
- Convexity: convex hull perimeter/contour perimeter

Colour Features: We capture colour information using:

- HSV colour space histograms (36 bins for hue, 32 for saturation, 32 for value)
- Dominant colour extraction using K-means clustering in colour space
- Spatial colour distribution by dividing the image into a 4×4 grid and computing colour statistics for each cell

Texture Features: We compute the Histogram of Oriented Gradients (HOG) features by:

- Calculating image gradients using Sobel operators
- Dividing the image into 8×8 pixel cells
- Creating histograms of gradient orientations for each cell (9 bins spanning 180°)
- Normalising histograms over partially overlapping blocks

4.3 Bag of Visual Words Representation

To create a unified representation from the variable number of SIFT descriptors extracted from each image, we implement the Bag of Visual Words (BoVW) model:

Visual Vocabulary Creation: We collect SIFT descriptors from all images and apply K-means clustering with K=500 to define a "visual vocabulary." Each cluster centre represents a "visual word" corresponding to a distinctive local pattern.

Histogram Generation: We assign each image’s SIFT descriptors to the nearest visual words and construct a normalised histogram representing the distribution of visual words in the image.

Feature Vector Combination: We concatenate the BoVW histogram with weighted versions of the shape, colour, and texture features. Through experimentation, we determined optimal weights to be:

- SIFT BoVW features: weight of 1.0 (baseline)
- Texture features: weight of 0.8
- Shape features: weight of 6.0 (significantly higher for accessories)
- Colour features: weight of 5.0 (highest weight due to importance in accessory discrimination)

Standardisation: We standardise the combined feature vectors to zero mean and unit variance to ensure features with different scales contribute equally to the clustering process.

4.4 Clustering Algorithms

We implemented and compared two clustering approaches:

K-means Clustering: Starting with randomly selected centroids, K-means iteratively:

- Assign each feature vector to the nearest centroid

- Recalculates centroids as the mean of all vectors in each cluster
- Repeats until convergence or maximum iterations are reached

Hierarchical Agglomerative Clustering: This bottom-up approach begins with each feature vector as its cluster and progressively merges the closest clusters. We experimented with different linkage criteria:

- Average linkage: distance between clusters is the average distance between all pairs of points
- Ward's method minimizes the variance increase after merging
- Complete linkage: distance between clusters is the maximum distance between any points

4.5 Determining the Optimal Number of Clusters

To identify the optimal number of clusters, we evaluated:

1. **Silhouette Score**
2. **Elbow Method**: Plots distortion (sum of squared distances to nearest centroid) against several clusters. The "elbow", where distortion reduction slows, represents a good compromise between complexity and fit.

4.6 Cluster Quality Assessment

We systematically evaluated K values ranging from 2 to 40 for K-means clustering, comparing silhouette scores to determine the optimal number of clusters. Similarly, for hierarchical clustering, we evaluated different levels of the dendrogram to find the optimal cut point.

To compare the K-means and hierarchical clustering results, we computed the Adjusted Rand Index (ARI), which measures the similarity between two clustering while accounting for chance. The ARI is calculated as:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI}$$

Where RI is the Rand Index, and E[RI] is its expected value under random clustering.

5. Evaluation

Our feature-based image clustering approach was evaluated using quantitative metrics and visual inspection of results, focusing on cluster quality and computational efficiency.

5.1 Optimal Number of Clusters and Visual Results

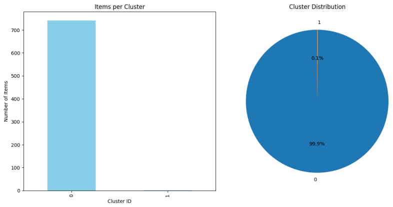

Figure 1: Cluster Distribution with recommended clusters per silhouette score (K-2)

While statistical analysis through silhouette scores strongly favoured k=2 (0.7943), the resulting clusters were highly imbalanced, with 99.9% of images in a single cluster (Figure 1). This prompted us to explore alternative cluster counts with more practical utility:

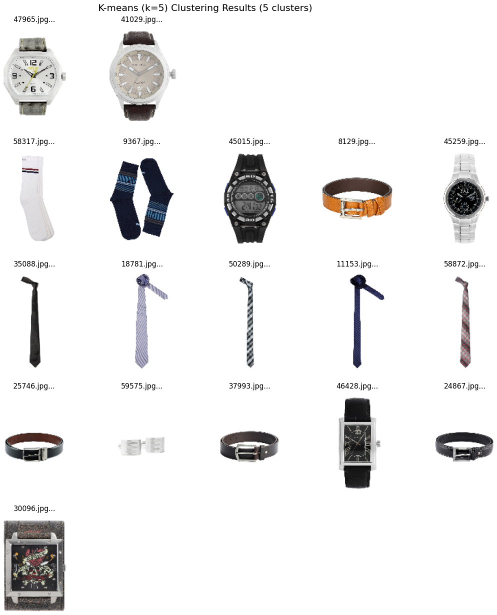

Figure 2: K means with 5 clusters

K-means with k=5 (Silhouette score: 0.0298) showed impressive visual coherence, as seen in figure 2:

- Cluster 1: Watches with light-coloured faces
- Cluster 2: Mixed accessories, including socks, digital watches, and belts
- Cluster 3: Ties with various patterns and colours
- Cluster 4: Belts with different buckle styles and a watch
- Cluster 5: A distinctive decorative watch

This clustering successfully separated ties, belts, and different types of watches based on visual appearance, demonstrating effective discrimination despite the lower silhouette score.

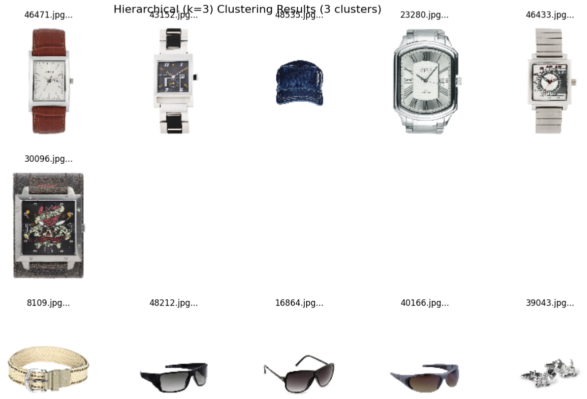

Figure 3: Hierarchical **wi**th 3 clusters

Hierarchical clustering with k=3 (Silhouette score: 0.0853) provided an alternative grouping (Figure 3):

- Group 1: Primarily watches with varied faces and bands
- Group 2: A distinctive decorative watch
- Group 3: Eyewear, including multiple styles of sunglasses

The hierarchical approach was particularly effective in grouping functionally similar items (watches, eyewear, belts) while maintaining visual coherence within each group.

5.2 Preprocessing Performance

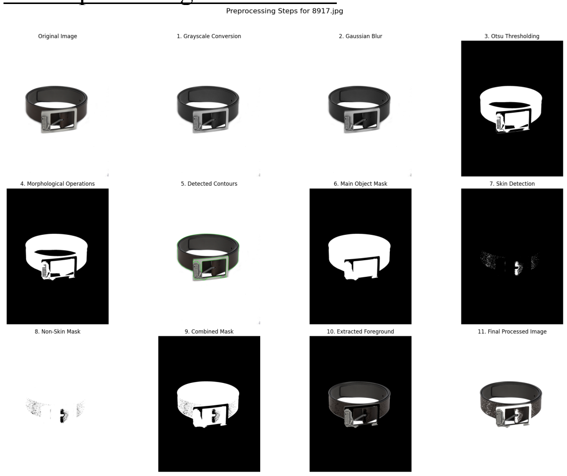

Figure 4: Pre-processing stages on a sample image

The effectiveness of our multi-stage pre-processing pipeline is demonstrated in Figure 4, which shows the step-by-step transformation for a belt (image 8917.jpg):

1. The original image is successfully converted to grayscale
2. Gaussian blur effectively reduces noise while preserving structural details
3. Otsu's thresholding creates a clear separation between the belt and the background
4. Morphological operations refine the binary mask, eliminating minor artefacts
5. Contour detection (green outline) accurately identifies the belt's shape
6. The primary object mask cleanly isolates the accessory
7. Skin detection identifies potential skin regions
8. The non-skin mask removes these regions
9. The combined mask provides the final segmentation 10-11. The foreground extraction and final processed image show a clean belt on a white background

This visual sequence confirms that our pre-processing approach effectively handles accessory isolation, even for challenging items with intricate details.

5.3 Feature Extraction Performance

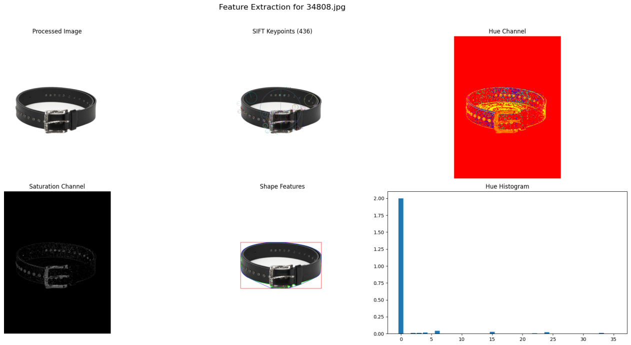

Figure 5: Feature extraction on a sample image

Figure 5 illustrates our comprehensive feature extraction process on a sample belt image (34808.jpg):

- SIFT key point detection successfully identified 436 distinctive points concentrated on the belt structure, holes, and buckle details
- The hue channel visualisation highlights different material components of the belt
- The saturation channel emphasises texture contrasts
- Shape features correctly outline the belt's contour (red rectangle) and identify its key geometric properties
- The hue histogram reveals the dark nature of the belt, with most values concentrated near 0

Extracting these complementary features contributed to the clustering algorithm's ability to group visually similar accessories.

5.4 Runtime Analysis

The computational performance broke down as follows (per image):

- SIFT Feature Extraction: 0.3316 seconds
- Texture Feature Extraction: 0.0302 second
- Colour Feature Extraction: 1.9944 seconds

- Total Feature Extraction: 2.3562 seconds

Colour feature extraction dominated the processing time (85% of the total), suggesting a potential area for optimisation. The overall processing time of 2.36 seconds per image indicates reasonable efficiency for moderately sized datasets but could become a bottleneck for huge collections.

5.5 Feature Vector Characteristics

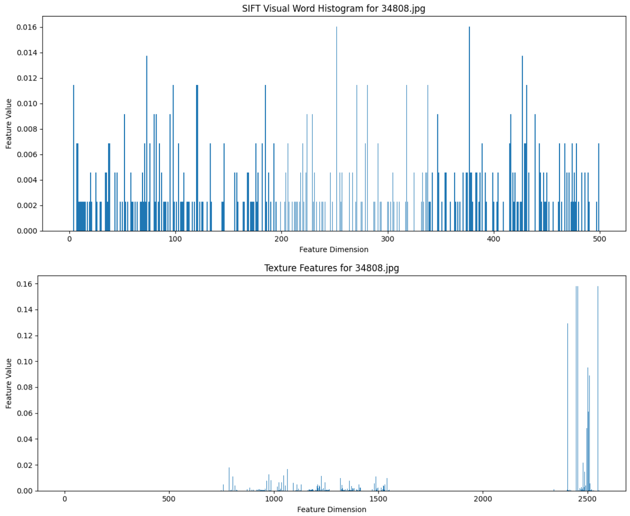

Figure 6: Sample histogram for a bag of visual words and Texture Features

Our final feature vectors combined:

- 500-dimensional Bag of Visual Words representations from SIFT features
- 5-dimensional shape feature vectors (aspect ratio, extent, solidity, equivalent diameter, convexity)
- Colour features including 36-bin hue histograms, 32-bin saturation histograms, and spatial colour information
- HOG-based texture features

These were weighted differently (SIFT: 1.0, texture: 0.8, shape: 6.0, colour: 5.0) to optimise clustering performance, with higher weights assigned to shape and colour features that proved more discriminative for accessories.

5.5 Limitations and Challenges

Despite the effectiveness of our approach for specific groupings, several limitations were observed:

- The binary separation dominant in the statistical analysis (k=2) did not align with functional categories, suggesting a substantial visual distinction that overrode more nuanced groupings.
- Visual similarity often crossed functional category boundaries, particularly for items sharing similar materials and colours (black leather items appeared identical regardless of whether they were belts, watches, or other accessories).
- The high concentration of dark-coloured accessories (as seen in the hue histograms) limited the discriminative power of colour features for many items.
- While generally effective, the pre-processing pipeline occasionally struggled with fragile accessories or items with similar colours to the background.

6. Conclusions

This project demonstrates that traditional computer vision techniques can successfully organise fashion accessories based on visual similarity without relying on deep learning approaches. Our feature-based image clustering approach revealed several important insights:

First, combining SIFT, shape, colour, and texture features provides sufficient discriminative power to create visually coherent clusters of accessories. The differential weighting of features proved crucial, with shape and colour characteristics requiring significantly higher weights (6.0 and 5.0, respectively) compared to SIFT and texture features to achieve meaningful groupings.

Second, our comprehensive pre-processing pipeline isolates accessories from complex backgrounds, enabling more accurate feature extraction. The multi-stage approach combining Otsu thresholding, morphological operations, and skin detection successfully handled various accessory types despite their different shapes, materials, and presentations.

Third, the optimal number of clusters differs significantly depending on the evaluation metric. While silhouette scores strongly favoured k=2, visual inspection revealed that k=5 for K-means and k=3 for hierarchical clustering produced more meaningful and practical groupings that better aligned with accessory categories.

Fourth, hierarchical clustering demonstrated particular strength in grouping functionally similar items (watches, eyewear, belts) while maintaining visual coherence within each group, suggesting it may be better suited for applications where functional categorisation is essential.

6.1 Highlights

The most notable achievements of our work include:

1. Successfully developing a robust background removal technique effectively isolates accessories

from complex backgrounds, even for challenging items like thin belts and patterned ties.

2. Creating a weighted feature representation that combines complementary aspects of accessories (shape, colour, texture, and key points) to enable more meaningful clustering.

3. This study demonstrates that alternative cluster counts (k=5 for K-means, k=3 for hierarchical) can provide practically useful groupings despite lower silhouette scores, highlighting the importance of qualitative evaluation alongside quantitative metrics.

4. Achieving reasonable computational efficiency (2.36 seconds per image) without requiring specialised hardware, making our approach accessible for moderate-sized fashion inventories.

6.2 Future Work

Several promising directions for future research emerge from this work:

1. **Neural Network Integration**: While our approach deliberately avoided deep learning, incorporating neural networks for specific components could enhance performance. For example, a CNN-based background removal model could improve pre-processing efficiency or learned feature extractors could replace or complement traditional features while maintaining the interpretability of the clustering stage.

2. **Adaptive Feature Weighting**: Developing methods to automatically determine optimal feature weights for different accessory types could improve performance across diverse inventories. This might involve a semi-supervised approach where a small set of labelled examples guides the weighting.

3. **Hierarchical Category Detection**: A two-stage approach that identifies broad functional categories (watches, belts, eyewear) and performs finer-grained clustering within each category could better align with how fashion retailers typically organise images.

4. **Colour Feature Optimization**: Since colour feature extraction dominated processing time (1.99 of 2.36 seconds per image), optimising this component through more efficient algorithms or parallel processing could significantly improve overall performance.

By building on these foundations, future systems could bridge the gap between unsupervised visual organisation and the semantic categories that are most useful for fashion retail applications, ultimately creating more efficient and intuitive product management tools.

7. References

[1] Lowe, D. G. (2004). "Distinctive Image Features from Scale-Invariant Key points." International Journal of Computer Vision, 60(2), 91-110.

[2] Harris, C., & Stephens, M. (1988). "A Combined Corner and Edge Detector." Proceedings of the Fourth Alvey Vision Conference, 147-151.

[3] Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern Classification (2nd ed.). John Wiley & Sons.

[4] Jain, A. K. (2010). "Data Clustering: 50 Years Beyond K-Means." Pattern Recognition Letters, 31(8), 651-666.

[5] Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations." Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[6] Hsiao, W.-L., & Grauman, K. (2018). "Creating Capsule Wardrobes from Fashion Images." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[7] Rother, C., Kolmogorov, V., & Blake, A. (2004). "GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts." ACM Transactions on Graphics (TOG), 23(3), 309-314.

[8] Aggarwal, P. (2018). "Fashion Product Images Dataset." Kaggle. Retrieved from https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset