big data and cognitive computing

MDPI

Article

Two-Stage Method for Clothing Feature Detection

Xinwei Lyu [1,*], Xinjia Li [2], Yuexin Zhang [2] and Wenlian Lu [1,2,3,4,*]

[1] Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Yangpu District, Shanghai 200433, China
[2] School of Mathematical Sciences, Fudan University, Yangpu District, Shanghai 200433, China
[3] Shanghai Center for Mathematical Sciences, Fudan University, Yangpu District, Shanghai 200433, China
[4] Shanghai Key Laboratory for Contemporary Applied Mathematics, Fudan University, Yangpu District, Shanghai 200433, China
* Correspondence: 19210850005@fudan.edu.cn (X.L.); wenlian@fudan.edu.cn (W.L.)

Abstract: The rapid expansion of e-commerce, particularly in the clothing sector, has led to a significant demand for an effective clothing industry. This study presents a novel two-stage image recognition method. Our approach distinctively combines human keypoint detection, object detection, and classification methods into a two-stage structure. Initially, we utilize open-source libraries, namely OpenPose and Dlib, for accurate human keypoint detection, followed by a custom cropping logic for extracting body part boxes. In the second stage, we employ a blend of Harris Corner, Canny Edge, and skin pixel detection integrated with VGG16 and support vector machine (SVM) models. This configuration allows the bounding boxes to identify ten unique attributes, encompassing facial features and detailed aspects of clothing. Conclusively, the experiment yielded an overall recognition accuracy of 81.4% for tops and 85.72% for bottoms, highlighting the efficacy of the applied methodologies in garment categorization.

Keywords: facial recognition; two-stage object detection; VGG16; SVM

Citation: Lyu, X.; Li, X.; Zhang, Y.; Lu, W. Two-Stage Method for Clothing Feature Detection. *Big Data Cogn. Comput.* **2024**, *8*, 35. https://doi.org/10.3390/bdcc8040035

Academic Editor: Moulay A. Akhloufi

Received: 8 January 2024
Revised: 14 March 2024
Accepted: 15 March 2024
Published: 26 March 2024

1. Introduction

The fashion apparel industry's reliance on imagery across various platforms, such as print media, e-commerce, and social media, has spurred the development of numerous object detection applications [1–4]. These applications aim to enhance apparel recognition, recommendation, and online search, ultimately improving the consumer experience. In the realm of computer vision, advancements in deep learning technology have significantly impacted image classification, object detection, and instance segmentation [5–7], with object detection emerging as a pivotal research area.

One of the primary challenges in clothing recognition stems from the immense diversity in the style, texture, and cut of clothing. This diversity is influenced by cultural backgrounds, individual preferences, and constantly evolving fashion trends, leading to a wide variation in the visual appearance of clothing items. Such variation presents significant challenges for recognition systems, which often struggle to accurately identify styles due to this high degree of diversity.

Moreover, clothing items in real-world scenarios frequently undergo deformation and occlusion. Garments can change shape with the movement of the human body and may be partially obscured by other objects in photographs [8]. These conditions complicate recognition efforts, as they modify the appearance features of clothing, making it difficult for models based on static images to accurately match clothing items.

Additionally, existing clothing recognition models are often characterized by high complexity, necessitating significant computational resources and storage space [9,10]. This limitation restricts the use of models on resource-constrained devices and escalates the costs associated with model training and deployment. Consequently, there is a pressing need

to develop efficient and lightweight clothing recognition models that can accommodate various application scenarios and device constraints.

Despite advancements in clothing recognition and detection technologies, significant limitations remain, including accuracy in style recognition, model complexity, and adaptability to real-world variations [11,12]. Addressing these challenges, this paper seeks to explore new methods and strategies aimed at ensuring the accuracy and efficiency of clothing recognition while reducing model complexity to better meet real-world application demands.

This paper describes a two-stage object detection structure that identifies facial features and clothing attributes. In the first stage, we use OpenPose [13] for detecting human keypoints, and Dlib [14] is employed for facial keypoint detection. Then, we complement a custom logic segmentation to yield precise body bounding boxes. The second stage involves using the VGG16 [15] model to classify multi-class attributes such as age and clothing material, while a standard SVM model is utilized for binary attributes like the presence of a zipper or collar type. This paper includes a review of related work on object detection in Section 2, an introduction to the two-stage method and related models in Section 3, a description of the dataset, experiments, and results in Section 4, and a conclusion of the study in Section 5.

2. Related Work

With the rise of e-commerce and social media, fashion image recognition has become a priority for computer vision research. This field focuses on traditional object detection and image classification. It can also be extended to image segmentation [16] and multi-label classification [17]. Recent advancements in deep learning have significantly enhanced the precision and detail of image analysis.

Fashion image recognition has historically depended on two-stage detection methods [18–20]. These methods begin by generating candidate regions and then localizing and classifying them. Despite their accuracy, these methods often require a longer processing time, which makes them less suitable for real-time applications. However, deep learning brings notable changes to these two-stage object detection methods. Deep CNNs and transformer models [21,22], among others, have improved candidate region generation and classification. For example, introducing a region proposal network (RPN) in faster R-CNN [23] has markedly enhanced the speed and quality of region proposals.

One-stage detection methods such as YOLO [24] and SSD [25] have been introduced to address the speed limitations of two-stage methods. These methods combine region proposal and classification into a single step, significantly increasing processing speed while maintaining reasonable accuracy. Recent advancements in object detection have seen the YOLOv5 algorithm being extensively applied across various domains, particularly in clothing recognition [26,27] and personal protective equipment (PPE) detection, showcasing its versatility and efficiency. The YOLOv5 algorithm, known for its speed and accuracy, has been adapted and improved to meet the specific needs of different applications. While two-stage methods generally offer higher accuracy than one-stage methods, they are slower than one-stage methods, which may compromise accuracy in complex scenarios.

Inspired by Liu Ziwei et al.'s work [28], this study proposes a two-stage method incorporating the latest advancements in deep learning, particularly in facial and human keypoint detection. They developed the DeepFashion dataset and introduced the VGG-based FashionNet network for classification. Meanwhile, they identified challenges in apparel recognition, such as clothing deformation and image variation under different conditions. Our study addresses these challenges by simplifying apparel characteristics such as collars, buttons, zippers, and clothing colors. We employ custom image segmentation based on the keypoint distribution for precise individual figure structuring. This study combines VGG and SVM to enhance the accuracy and efficiency of apparel recognition.

3. Proposed Method

This study adopts a two-stage object detection approach that distinguishes it from traditional machine learning algorithms which exhaustively search for candidate regions. Traditional methods often generate redundant regions, demanding high computational resources and slow processing speeds. The two-stage approach effectively reduces the number of candidate bounding boxes, thereby lowering the computational requirements for classification. In the first stage of this study, the input image was processed to extract bounding boxes. We combined facial and body keypoint analysis and then divided body parts through custom logic. The second stage applies deep neural networks and support vector machines (SVMs) to classify different parts, with the collective attributes of all parts constituting the detection outcome. The research explicitly analyzes attributes such as gender, age, collar, zipper, top material, top pattern, bottom type, sleeve length, and bottom length. The examples of top material and top pattern are shown in Figure 1.

(**a**) (**b**)

Figure 1. (**a**) Examples of top material, including cotton, denim, fur, lace, leather, and tweed; (**b**) examples of top pattern, including floral, plaid, graphic, solid, spotted, and stripe.

Firstly, gender and age are associated with facial information. Original images are marked with facial keypoints, identifying coordinates of eye corners, nose tips, and mouth corners. These keypoints play a crucial role in facial expressions and individual characteristics. After facial detection, the original images are cropped to obtain facial images. These cropped facial images, with varied orientations, require facial calibration, which is determined by the bridge of the nose keypoints to align the facial images in a uniform direction. These facial images and corrected keypoint data are input into classifiers for training to learn gender and age attribute information.

Additionally, this study introduces a discrete differential operator for edge detection, the Scharrx operator, for preprocessing images to enhance horizontal edges in facial images. This enhancement helps highlight features such as the jawline, eyebrows, and lip edges, essential for learning age-related features (like wrinkles or structural changes in the face) and gender-related features (such as facial hair or jawline shape), aiding in achieving more precise classification outcomes. Thus, Scharrx edge images, as an additional channel to the original images and facial keypoint information, serve as inputs for classification learning.

The structure of the facial feature model is illustrated in Figure 2.

Clothing features are also obtained by identifying essential body parts, such as the head, elbows, knees, and ankles, to obtain corresponding coordinate information. We propose a custom segmentation logic based on keypoints, dividing the body into detailed areas like upper arms, forearms, neck, torso, collar, thighs, and calves to obtain respective body part frames. These parts are then cropped to obtain detailed images of each section, facilitating the subsequent classification learning of detailed features for each part.

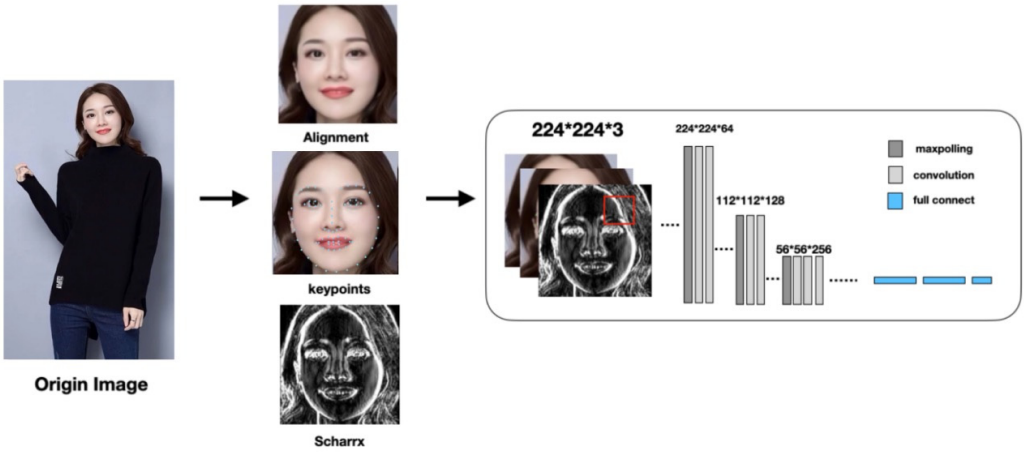

Figure 2. The structure of the facial feature model.

We also combine Harris Corner Detection and Canny Edge Detection [29–31] to meet the demands of learning clothing features. Both are used for detecting corners and edges in images, providing preliminary feature extraction for clothing attributes such as collar shapes, zippers, and bottom type, and offering enriched features for classifiers to improve accuracy. Harris Corner Detection and Canny Edge Detection are first used to identify key features like corners, edges, and contours in clothing images. These features are then transformed into feature vectors and combined with bounding boxes to create a rich composite input. Support vector machines (SVMs) use this input for classification, finding the optimal hyperplane in the feature space to separate different categories. This method achieves high accuracy by blending local features with the global structure of images, making SVMs more effective in classifying complex clothing images. However, Harris Corner Detection and Canny Edge Detection focus on corners and edges in images, potentially causing overfitting. Therefore, the input for classifying patterns and materials of tops includes only the original pixels of the corresponding torso region. The structure of the clothing feature model is illustrated in Figure 3.

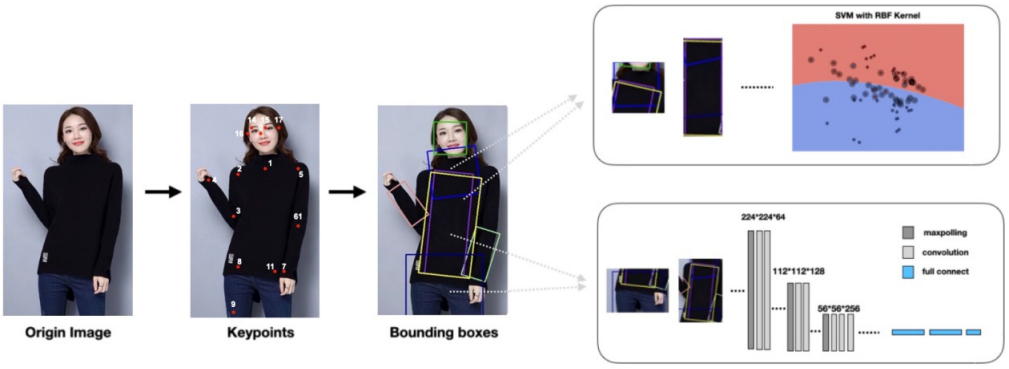

Figure 3. The structure of the clothing feature model.

In this research, keypoint detection methods utilize OpenPose and Dlib. OpenPose outputs a set of 2D coordinates for each keypoint on each individual in the image, along with confidence scores for each detection, while Dlib outputs keypoint coordinates and the bounding boxes of detected objects.

The primary classification model used in this study is VGG16, renowned for its high accuracy in object detection algorithms and proven high performance in various image recognition tasks. As backbone networks, VGG models are particularly effective in recognizing and classifying facial features, capable of capturing complex patterns and features in images. Moreover, VGG models pre-trained on large datasets can be easily applied to other image tasks through transfer learning, achieving good performance even

on smaller datasets. Additionally, SVM classifiers recognize features such as zippers and collar shapes. The datasets for zipper and collar type are relatively small, whereas SVMs can also perform well, offering efficiency compared to smaller neural network classifiers.

Additionally, we determine skin exposure based on the ratio of skin pixels within a threshold range for sleeve length and bottom length attributes, employing empirical values for skin pixel ranges [32] ($77 < Cb < 127$, $133 < Cr < 173$). For instance, if the ratio of pixels within the designated range in the forearm and lower leg regions exceeds 0.3, the garment is classified as long-sleeved or long pants, respectively; conversely, a lower ratio indicates short sleeves or shorts. This range typically covers individuals with lighter complexions well, but it may not be as applicable for those with darker skin tones. To accommodate a variety of skin colors, this paper employs a combined approach of leveraging the YCbCr color space range and statistical analysis of facial skin tones to detect skin pixels comprehensively. Facial regions are converted from the RGB color space to the YCbCr color space, and the values in the Cb and Cr channels are statistically analyzed. The median (Md) and standard deviation (σ) of these values are calculated, and the threshold range for the Cb and Cr channels is set to $Md \pm 2\sigma$, resulting in a newly defined skin pixel range. The specific process for skin pixel recognition is illustrated in Figure 4.

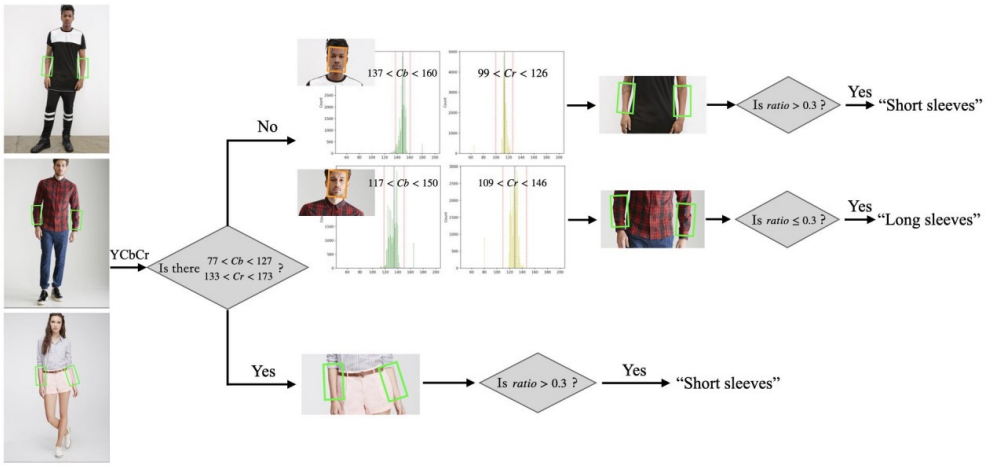

Figure 4. The process for skin pixel recognition.

4. Experiment

4.1. Dataset

In this research, data were sourced from the DeepFusion dataset, a publicly accessible resource developed by the Multimedia Laboratory at The Chinese University of Hong Kong [33]. This dataset comprises images obtained from various fashion shopping websites. The study initially focused on object detection and then extracted and inferred facial and clothing features using VGG16 and SVM. Notably, the data are manually annotated with specific attributes such as age, gender, collar type, zipper, sleeve length, top material, top pattern, bottom type, and bottom length.

In this study, we employed random preprocessing adjustments to images from the same dataset to enhance the model's generalization capability. By applying random transformations such as cropping, rotation, and color adjustment, we generated a diversified training sample set that simulates various scenarios and conditions encountered in the real world. This data augmentation strategy is instrumental in mitigating model overfitting, bolstering its robustness against new and unseen images.

Data augmentation is a widely used technique in machine learning, particularly in the training of deep learning models, aimed at increasing the diversity of the training set through the application of random transformations to images. These transformations include rotations, scaling, cropping, flipping, and alterations in brightness and contrast. The primary objective of data augmentation is to artificially enhance the diversity of training

data by emulating different factors that might affect the appearance of images in the real world, thus producing new images derived from the original training data. This approach not only simulates environments with varying lighting conditions and color variations, aiding the model in learning features independent of color changes, but also introduces variability in the orientation of objects or the camera by flipping images horizontally or vertically, thereby increasing the model's invariance to image flipping. Figure 5 illustrates the comparative images before and after image enhancement. The specific techniques used for augmentation were as follows:

- Randomly mirror the image with a probability of 0.5.
- Randomly adjust the brightness between 0.9 and 1.1 times the original image.
- Randomly adjust the contrast between 0.9 and 1.1 times the original image.
- Randomly adjust the hue between 0.9 and 1.1 times the original image.
- Randomly adjust the saturation between 0.9 and 1.1 times the original image.

(**a**) (**b**)

Figure 5. (**a**) Original image; (**b**) augmented image.

4.2. Obtaining Bounding Boxes

Initially, body and facial keypoints are detected using OpenPose and Dlib. OpenPose utilizes the "CoCo18keypoints" neural network structure to detect body keypoints. The OpenPose website provides a Caffe framework for this neural network structure. The pre-trained network is accessed using the readNetFromCaffe function [34]. The Dlib library [35] in OpenCV identifies 68 crucial facial points, offering a more comprehensive representation of facial features than OpenPose. These 68 keypoints together outline the facial contour. OpenPose's detection of keypoints and the corresponding pose map of the original photo are illustrated in Figure 6. The facial recognition box and 68 keypoints generated by Dlib's facial recognition in the original photo are displayed in Figure 7.

In human images, variations like tilted head angles are common. Effectively calibrating faces that are not fully oriented enhances facial feature recognition accuracy. Standard calibration methods include DeepFace and DEX. This study adopts a self-improved calibration method based on facial keypoints.

For facial calibration, the rotation angle θ is calculated based on the nose bridge keypoints (27, 28, 29, 30). The angle θ is then calculated as follows:

$$\theta = \arctan\left(\frac{y_{30} - y_{27}}{x_{30} - x_{27}}\right)$$

This formula computes the arctangent of the slope formed by the line connecting keypoints 27 and 30. The resulting angle θ is used for image rotation alignment. Notably, during this process, the original image is rotated, and the facial recognition box is recalibrated using the line between keypoint 27 and 30 as the rotation axis. This method involves rotating the entire original image instead of just the cropped face portion. Rotating the original image introduces edge information beyond the facial recognition box, including hair, ears, and collars. This extra context, extending outside the original facial bounding box, enhances facial recognition algorithms. It improves the algorithms' ability to identify facial features by providing more contextual information.

(**a**) (**b**)

Figure 6. (**a**) Eighteen keypoints distribution map of OpenPose; (**b**) connecting keypoints by direction to obtaining pose map.

(**a**) (**b**)

Figure 7. (**a**) Dlib facial detection box; (**b**) a diagram of 68 keypoints generated by Dlib.

A similar method is used for the body part boxes. Key points identify specific body parts, and then the corresponding body part boxes are cropped. The principal logic for cropping various body parts includes the following:

- Torso: Keypoints 2, 5, 8, 11, and 1 are primarily used. The logic involves forming a rectangle using the line between keypoints 2 and 5 as the width and the line between 1 and 8 as the length. The length direction is determined by the line connecting the midpoint of keypoints 8 and 11 with keypoint 1.
- Collar: Keypoints 0 and 1 are primarily used. The logic is to form a square with the neck as the intersection point of the diagonals and the distance between the neck and nose as the side length, oriented from the neck to the nose.
- Zipper: Primarily using keypoints 1, 8, and 11, the logic forms a rectangle with the midpoint of the line between 8 and 11 as the central axis of the longer side. The length is determined by the line connecting keypoint 1 and the midpoint, and the width is a quarter of the line length between 1 and 8.
- Upper Arm: For the right upper arm, keypoints 2 and 3 are used, and for the left, keypoints 5 and 6 are used. The method creates a rectangle with the line between keypoints 2 and 3 (or 5 and 6) as the central axis of the longer side, and the width is half the length of the longer side.
- Forearm: Keypoints 3 and 4 are used for the right forearm, and keypoints 6 and 7 are used for the left. The logic forms a rectangle with the line between keypoints 3 and 4 (or 6 and 7) as the central axis of the longer side.
- Bottom: Keypoints 8, 10, 11, and 13 are used. The distance between the midpoint of the line connecting keypoints 8 and 11 and the midpoint of the line connecting keypoints 10 and 13 is used as the length of the rectangle, with half of this distance serving as the width.
- Thigh: For the right thigh, keypoints 8 and 9 are used, and keypoints 11 and 12 are used for the left. The method involves forming a rectangle with the line between keypoints 8 and 9 (or 11 and 12) as the central axis of the longer side.
- Lower Leg: Keypoints 9 and 10 are used for the right lower leg, and keypoints 12 and 13 are used for the left. The logic is similar to the thigh, forming a rectangle with the line between keypoints 9 and 10 (or 12 and 13) as the central axis of the longer side.

During the final testing phase, 85 randomly selected images were used, with 71 successfully passing the detection, resulting in an 84% pass rate. An example image of human body part boxes is shown in Figure 8.

(**a**) (**b**)

Figure 8. (**a**) Original image; (**b**) an example image including body part boxes.

4.3. Facial Attributes

Models for age and gender recognition were developed by employing a randomized shuffling of the training dataset. The age classification model accurately identified features in 6698 out of 8667 test photographs, resulting in an accuracy rate of 77.28%. The gender classification model achieved recognition accuracy of 95.89% on a test dataset comprising 5859 images. An age confusion matrix heatmap is displayed in Figure 9a. A gender confusion matrix heatmap is displayed in Figure 9b.

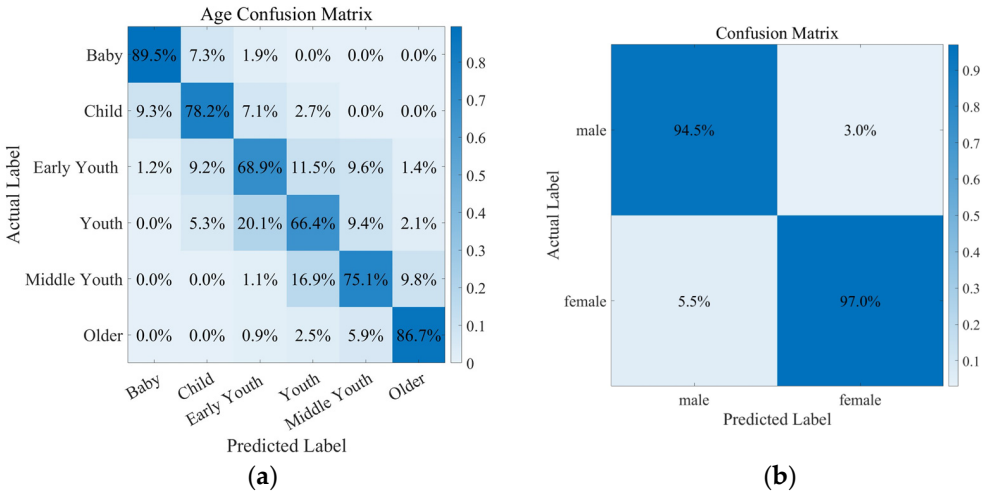

Figure 9. (**a**) Age confusion matrix heatmap; (**b**) gender confusion matrix heatmap.

4.4. Clothing Attributes

The attribute of zippers and buttons was identified using an SVM classifier, which demonstrated a 70% accuracy rate across a test set of 294 images. Collar shapes were classified as either V-neck or round neck based on the features extracted from the collar area, with the final accuracy reaching 78%. The performance of SVM with zipper on the test dataset is presented in Table 1. The performance of SVM with collar type is detailed in Table 2.

Table 1. Performance results table of SVM with zippers on the test set.

Zipper Attribute	Precision	Recall	F1-Score	Image Amount
with	0.71	0.66	0.69	149
without	0.68	0.72	0.70	145

Table 2. Performance results table of SVM with collar type on the test set.

Collar Type Attribute	Precision	Recall	F1-Score	Image Amount
V-neck	0.81	0.77	0.79	247
round neck	0.75	0.79	0.77	213

In terms of top materials (cotton, denim, fur, lace, leather, and tweed), the model achieved an accuracy rate of 57.04% on the test dataset. For the classification of top pattern (floral, plaid, graphic, solid, spotted, and stripe), the model showed an accuracy rate of 89.81% on the test set. The confusion matrix heatmap for top material classification is depicted in Figure 10a. The confusion matrix heatmap for top pattern classification is presented in Figure 10b. Bottoms are mainly categorized into "Skirt" and "Trousers". On the test dataset, the model records an average accuracy of 91.2%. The confusion matrix heatmap for bottom type classification is presented in Figure 11.

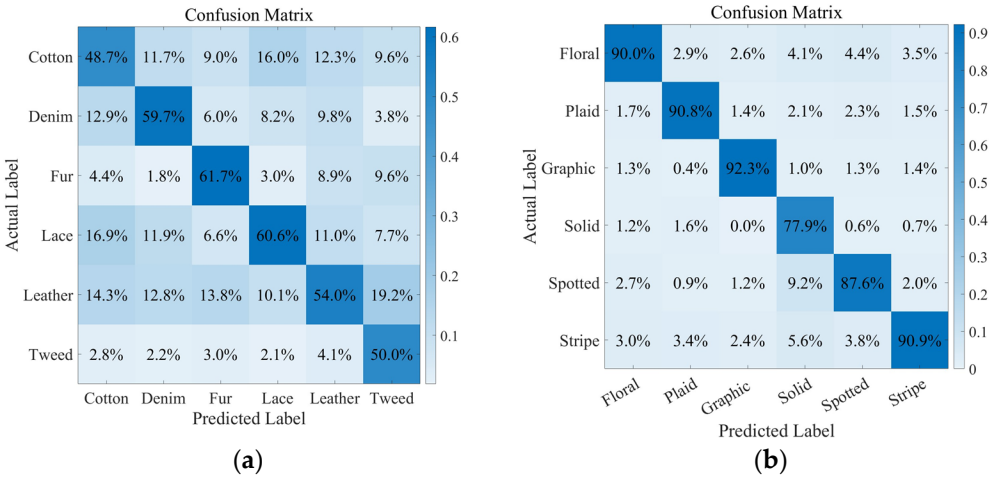

Figure 10. (**a**) Top material confusion matrix heatmap; (**b**) top pattern confusion matrix heatmap.

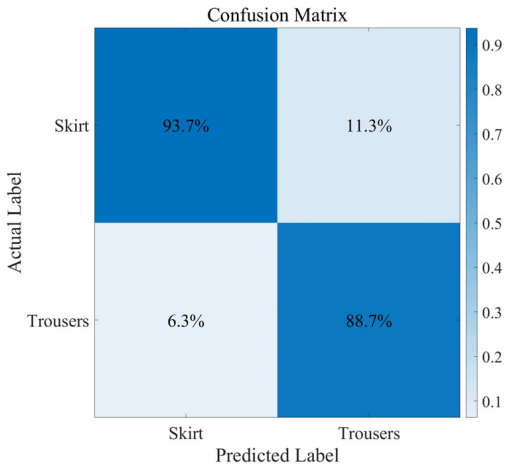

Figure 11. Bottom type confusion matrix heatmap.

The determination of sleeve length relied on the ratio of exposed skin pixels within the detected lower arm frames. Sleeves were classified as short if the bare skin ratio exceeded 0.3 and as long sleeves if the ratio was below 0.3, with accuracy rates of 93.72% and 84.49% for short and long sleeves, respectively. The methodology for determining the length of bottoms was analogous, initially checking for the presence of skin pixels within thigh frames. In the absence of skin pixels in thigh frames, calf frames were examined. Bottoms were categorized as short or long based on the skin exposure ratio calculated within the calf frames, with the model achieving an accuracy rate of 80.23% on the test dataset. Table 3 displays the performance of the sleeve and bottom length results.

Table 3. Performance results table of sleeve/bottom length on the test set.

Sleeve Length	Precision	Recall	F1-Score	Image Amount
long	0.85	0.79	0.82	1025
short	0.94	0.81	0.87	5935
Bottom Length	**Precision**	**Recall**	**F1-Score**	**Image Amount**
long	0.84	0.77	0.80	3186
short	0.80	0.78	0.79	2307

4.5. Proceeding Speed

The experiments used the Pytorch 0.4.0 deep learning framework on a 64-bit Ubuntu 16.04 operating system, facilitated by the Nvidia TITAN Xp GPU for training. The execution

time of the code for each segment in the experimental setup of this study is presented in Table 4. The final results of an example are shown in Figure 12.

Table 4. Running time of each section.

Stage	Step	Operation	Average Cost Time
First	Body bounding boxes	Body keypoint detection and segment logic	2.9022 s
Second	Facial bounding boxes	Facial keypoint detection and segment logic	1.4905 s
Third	Age/Gender	Facial bounding box and VGG16 classifier	1.6338 s
Forth	Zipper	Zipper bounding box and SVM classifier	0.0034 s
	Collar	Collar bounding box and SVM classifier	0.0063 s
	Collar shape	Collar bounding box and SVM classifier	0.003 s
	Sleeves length	Arm bounding box and skin detector	<0.0001 s
	Top material	Torso bounding box and VGG16 classifier	1.6455 s
	Top pattern	Torso bounding box and VGG16 classifier	1.6367 s
	Bottom type	Bottoms bounding box and VGG16 classifier	0.8359 s
.....	Bottom length	Leg bounding box and skin detector	<0.0001 s

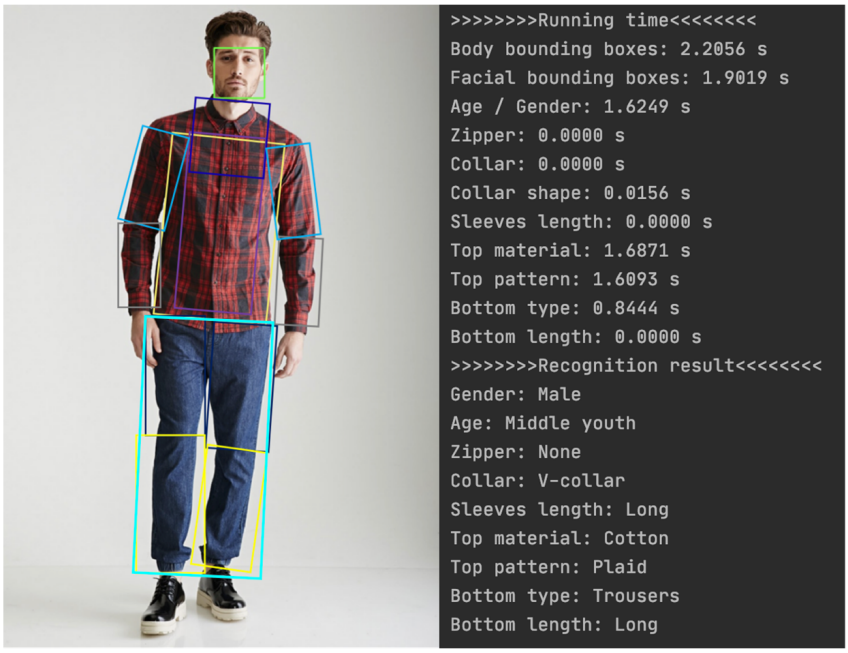

Figure 12. The final result of an example displayed.

4.6. Discussion and Results

In the realm of facial attribute recognition, the accuracy rate for gender identification reached an impressive 95.89%, while age recognition achieved an accuracy of 77.28%. This indicates that the employed models are adept at capturing the key features distinguishing gender and age. For age recognition, the accuracy may be affected by factors such as

makeup and obstructions. Obstructions like glasses, hats, or hair could conceal key age-indicative features, thereby reducing the accuracy of recognition.

Regarding clothing attribute recognition, a significant variance in accuracy rates for different attributes was observed. The recognition accuracy for top pattern in the test set was 89.81%, whereas the accuracy for identifying top materials was only 57.04%. The lower accuracy in recognizing clothing materials may be linked to the inherent constraints of the VGG model. Patterns typically exhibit unique colors and shapes that are easily identifiable, even in lower-resolution images. In contrast, the texture details of materials may become indiscernible at lower image qualities. Material characteristics pose more challenges than pattern features under the same data quality. The lower performance of clothing materials in this study is also limited by the inherent limitations of the VGG model. The model's focus on extracting global features and shape information may not sufficiently capture the nuanced differences in material textures, which are crucial for distinguishing fabrics. These subtle textures and details within images, pivotal for material differentiation, might be overlooked during the convolutional processing in VGG. Moreover, the model's performance is notably influenced by the image's quality and resolution; images of lower resolution or subpar quality might lack the necessary detail to differentiate between material textures effectively.

To address these limitations, future research could consider ensuring high-resolution images. The model's generalization capability could be enhanced through increasing the diversity of material samples, including variations in lighting, angles, and potential obstructions. Additionally, more advanced deep learning architectures could be explored which capture the fine details of material textures.

For specific attributes like sleeve and bottom lengths, a logic-based method relying on the ratio of exposed skin pixels was employed, achieving an average accuracy of 89.1% for sleeve length and 80.23% for bottom length, thereby validating the effectiveness of this approach. Our study excluded the effects of stockings and tattoos because they could interfere with the judgment of skin pixels. Future models will need to account for these elements, facing challenges such as accurately distinguishing between tattoos and clothing patterns, as well as addressing changes in skin color and texture caused by stockings.

Due to the feature segmentation of clothing in this study, the accuracy of individual segmented features contributes to the overall recognition error for tops. In this experiment, the overall recognition accuracy for tops was determined to be 81.4%, and the overall recognition accuracy for bottoms was 85.72%. The overall recognition accuracy for tops is calculated as the average accuracy of segmented features, including collar, zipper, pattern, material, and sleeve length. Conversely, the overall recognition accuracy for bottoms is determined by averaging the accuracies of two categories: type and length of the bottom wear. By integrating different models for various features, we attained an above-average recognition accuracy for complete top categories, underscoring the efficacy and practicality of our method.

In terms of performance and processing time, our approach also demonstrated commendable results. The experimental findings revealed that the complete recognition process for top categories averaged only 1.6582 s, and the bottom category recognition process took merely 0.8359 s, keeping the average total processing time within a reasonable range. This ensures the practicality and operability of the model. These results suggest that, despite the typically high computational demands of deep learning models, our method has been optimized for performance, maintaining reasonable processing times.

In summary, our method, through the meticulous segmentation of clothing features and the adaptive application of the most suitable classification models for different features, not only enhances recognition accuracy but also ensures the efficiency of the model's operation. This showcases the potential and practical value of our approach in the fields of clothing and facial attribute recognition.

5. Conclusions

In this study, we have developed an innovative approach for identifying and classifying clothing and facial attributes, with a particular emphasis on the detailed segmentation of features on tops. Contrary to traditional research that classifies entire tops or lower body attire directly, we further dissect tops into several feature categories, such as collars, zippers, materials, and patterns. Depending on the specific characteristics of these features, such as binary or multi-class issues, we select the most suitable classification model for processing.

The methodologies employed in this research, such as data augmentation, SVM classifiers, and deep learning models like VGG16, provide several advantages. Data augmentation enhances the model's generalization capability by simulating real-world variations in training data, reducing the risk of overfitting. SVM classifiers have been proven effective in recognizing specific attributes such as zippers and collars. The VGG16 model is renowned for its deep architecture, adept at capturing complex patterns within facial and clothing features, which aids in improving the accuracy of gender recognition and certain clothing attributes.

Conclusively, the experiment yielded an overall recognition accuracy of 81.4% for tops and 85.72% for bottoms, highlighting the efficacy of the applied methodologies in garment categorization. The use of data augmentation and the combination of SVM and deep learning approaches represent methodological advancements, offering a more nuanced understanding of the complex interactions between different attributes in fashion images.

However, these methods also have limitations. SVM classifiers may not capture hierarchical feature representations as effectively as deep learning models, potentially limiting their performance in more complex classification tasks. The performance of the VGG16 model may be influenced by the quality and diversity of the training data.

In summary, the method proposed in this study performs well in terms of overall accuracy. In terms of computational performance, by optimizing models and algorithms, as well as leveraging high-performance hardware, this approach can achieve rapid processing times while maintaining high accuracy, making it suitable for practical application scenarios. Future work could further explore the optimization of models and algorithms to improve the accuracy of age recognition and certain clothing attribute identifications while maintaining or enhancing computational efficiency.

Author Contributions: Conceptualization, X.L. (Xinwei Lyu) and X.L. (Xinjia Li); methodology, X.L. (Xinwei Lyu), X.L. (Xinjia Li), and Y.Z.; validation, X.L. (Xinwei Lyu); formal analysis, X.L. (Xinwei Lyu); data curation, X.L. (Xinwei Lyu); writing—original draft preparation, X.L. (Xinwei Lyu) and Y.Z.; writing—review and editing, X.L. (Xinjia Li) and W.L.; visualization, X.L. (Xinwei Lyu); supervision, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study: http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html (accessed on 1 August 2022).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, K.; Yuan, C.; Jiang, Y.; Luo, L. Weakly supervised instance segmentation by exploring entire object regions. *IEEE Trans. Multimed.* **2021**, *25*, 352–363. [CrossRef]
2. Lee, C.-H.; Lin, C.-W. A Two-Phase Fashion Apparel Detection Method Based on YOLOv4. *Appl. Sci.* **2021**, *11*, 3782. [CrossRef]
3. Badola, K.; Joshi, A.; Sengar, D. Product Recommendation using Object Detection from Video, Based on Facial Emotions. In Proceedings of the 2020 International Conference on Computer Science and Information Technology, Chennai, India, 26–27 December 2020; pp. 51–56. [CrossRef]
4. Wang, S.; Du, Z.; Du, Y.; Chen, J. Online object detection task offloading in UAV ad hoc networks. In Proceedings of the 2022 IEEE International Conference on Unmanned Systems (ICUS), Guangzhou, China, 28–30 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 758–763.

5. Michelucci, U.; Michelucci, U. Object Classification: An Introduction. In *Advanced Applied Deep Learning: Convolutional Neural Networks and Object Detection*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 195–220.
6. Hechun, W.; Xiaohong, Z. Survey of Deep Learning Based Object Detection. In Proceedings of the ICBDT 2019: 2nd International Conference on Big Data Technologies, Jinan, China, 28–30 August 2019; pp. 149–153. [CrossRef]
7. Li, K. Applications of Deep Learning in Object Detection. In Proceedings of the 2022 International Conference on Computers, Information Processing and Advanced Education (CIPAE), Ottawa, ON, Canada, 26–28 August 2022; pp. 436–442. [CrossRef]
8. Wang, Y.-H.; Wang, T.-W.; Yen, J.-Y.; Wang, F.-C. Dynamic human object recognition by combining color and depth information with a clothing image histogram. *Int. J. Adv. Robot. Syst.* **2019**, *16*, 1729881419828105. [CrossRef]
9. Wang, C.; Yao, B.; Liu, L.; Peng, Y. A Lightweight Serial CNN Model for Remote Sensing Ship Target Recognition on FPGA. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Guiyang, China, 17–22 July 2022.
10. Zhang, X. Design and Implementation of the Chinese Character Font Recognition System Based on Binary Convolutional Encoding and Decoding Network. In Proceedings of the IEEE International Conference on Pattern Recognition and Computer Vision (PRCV), Shenyang, China, 14–16 July 2023.
11. Chen, X.; Deng, Y.; Di, C.; Li, H.; Tang, G.; Cai, H. High-Accuracy Clothing and Style Classification via Multi-Feature Fusion. *Appl. Sci.* **2022**, *12*, 10062. [CrossRef]
12. Phyu, K.W.; Funakubo, R.; Hagiwara, R.; Tian, H.; Minami, M. Verification of illumination tolerance for photo-model-based cloth recognition. *Artif. Life Robot.* **2018**, *23*, 118–130. [CrossRef]
13. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
14. Yang, J.; Adu, J.; Chen, H.; Zhang, J.; Tang, J. A facial expression recongnition method based on dlib, ri-lbp and resnet. *J. Phys. Conf. Ser.* **2020**, *1634*, 012080. [CrossRef]
15. Itkare, S.; Manjaramkar, A. Fashion classification and object detection using CNN. In Proceedings of the Information and Communication Technology for Competitive Strategies (ICTCS 2020) Intelligent Strategies for ICT, Jaipur, India, 11–12 December 2020; Springer: Berlin/Heidelberg, Germany, 2021; pp. 227–236.
16. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3523–3542. [CrossRef] [PubMed]
17. Dao, S.D.; Zhao, E.; Phung, D.; Cai, J. Multi-label image classification with contrastive learning. *arXiv* **2021**, arXiv:2107.11626.
18. Bai, T. Analysis on Two-stage Object Detection based on Convolutional Neural Networks. In Proceedings of the 2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), Bangkok, Thailand, 30 October–1 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 321–325.
19. Ansari, M.F.; Lodi, K.A. A survey of recent trends in two-stage object detection methods. In Proceedings of the Renewable Power for Sustainable Growth: Proceedings of International Conference on Renewal Power (ICRP 2020), Jammu, India, 17–18 April 2020; Springer: Berlin/Heidelberg, Germany, 2021; pp. 669–677.
20. Shi, Z. Object Detection Algorithms: A Comparison. In Proceedings of the 2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Dali, China, 12–14 October 2022; pp. 861–865. [CrossRef]
21. Lu, W.; Lan, C.; Niu, C.; Liu, W.; Lyu, L.; Shi, Q.; Wang, S. A CNN-Transformer Hybrid Model Based on CSWin Transformer for UAV Image Object Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1211–1231. [CrossRef]
22. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
23. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
24. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
25. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
26. Chang, Y.; Zhang, Y.-Y. Deep Learning for Clothing Style Recognition Using YOLOv5. *Micromachines* **2022**, *13*, 1678. [CrossRef] [PubMed]
27. Wanlin, E.; Yang, Z.; Yu, J. Detection and Recognition of Personal Protection Equipment Wearing Based on an Improved YOLOv5 Algorithm. In Proceedings of the 2023 4th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC), Nanjing, China, 18–20 August 2023.
28. Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; Tang, X. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1096–1104.
29. Kitti, T.; Jaruwan, T.; Chaiyapon, T. An object recognition and identification system using the harris corner detection method. *Int. J. Mach. Learn. Comput.* **2012**, *2*, 462. [CrossRef]
30. Harris, C.; Stephens, M. A Combined Corner and Edge Detector. Available online: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=88cdfbeb78058e0eb2613e79d1818c567f0920e2 (accessed on 1 August 2022).
31. Canny, J. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *PAMI-8*, 679–698. [CrossRef]

32. Patravali, S.D.; Wayakule, J.M.; Katre, A.D. Skin segmentation using YCBCR and RGB color models. *Int. J.* **2014**, *4*, 341–346.
33. MMLab, The Chinese University of Hong Kong. DeepFashion. Available online: http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html (accessed on 1 August 2022).
34. CMU Perceptual Computing Lab. OpenPose. Available online: https://github.com/CMU-Perceptual-Computing-Lab/openpose (accessed on 1 August 2022).
35. Dlib C++ Library. Available online: http://dlib.net (accessed on 1 August 2022).