Signal, Image and Video Processing (2020) 14:1441–1449
https://doi.org/10.1007/s11760-020-01690-8

ORIGINAL PAPER

Human attribute recognition method based on pose estimation and multiple-feature fusion

Xiao Ke[1,2,3] · Tongan Liu[1,3] · Zhenda Li[1,3]

Received: 15 October 2019 / Revised: 3 February 2020 / Accepted: 2 April 2020 / Published online: 30 April 2020

Abstract
As easy-to-search semantic information, human clothing attributes have important research value in the field of computer vision. Existing attribute recognition methods encounter problems such as interference from environmental factors, and as a result show poor clothing positioning accuracy. To address these problems, a human attribute recognition method based on human pose estimation and multiple-feature fusion is proposed. First, some retrieval results are obtained for subsequent attribute recognition through appearance feature matching. Then, through a deep SSD-based human pose estimation method, the foreground area belonging to the human in the image is located, and the background interference is excluded. Finally, the analytical results of various methods are combined. The iterative smoothing process and the maximum posteriori probability assignment method are adopted to enhance the correlation between attribute labels and pixels, and the final attribute recognition results are obtained. Experiments on the benchmark dataset show that the performance of our model is improved, and solves the problems of inaccurate clothing label recognition and pixel resolution area deviation in a single recognition mode.

Keywords Deep learning · SSD · Pose estimation · Multiple-feature · Human attribute recognition

1 Introduction

Human attribute recognition [1] has important research value in the field of computer vision. Human attributes such as age, gender, and clothing worn can be used as easy-to-search semantic information that can be applied to video surveillance for biometric recognition, and can be applied to face detection [2], automatic image annotation [3], and saliency detection [4]. An important advantage of semantic information based on low-level visual features is its robustness to the diversity of viewpoint changes, which means that human attributes can serve as a basis for subsequent long-term computer vision work.

The recognition of human attributes in video-acquired surveillance images in the real world is challenging for a variety of reasons. First, imaging quality is usually poor, with low resolution and a high susceptibility to motion blur [5]. Secondly, the recognition may be affected by the appearance of the clothing, and because of different human poses in different images, the corresponding attributes can be located in different spatial positions in the images. Finally, the tagged attribute data from the surveillance video images are difficult to collect and can only be obtained in small quantities. These three factors make it very difficult to learn a model of human attributes through training. Early attribute recognition methods relied primarily on manually extracted features, such as the color or text annotation of the clothing items. In recent years, human attribute recognition models based on deep learning have begun to attract more and more attention. This is because deep learning models have more powerful and stable learning abilities when given large-scale datasets. The images obtained through surveillance video are usually of poor quality, low resolution, and complex clothing appearance changes in the surveillance scene. Even in these difficult situations, we can get a well-performing model through deep learning. All these factors undoubtedly increase the position of deep learning in the recognition of human attributes.

To solve these problems, one approach is to investigate the interdependence and correlation between related

✉ Xiao Ke
kex@fzu.edu.cn

1 College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China

2 Key Laboratory of Spatial Data Mining and Information Sharing, Ministry of Education, Fuzhou 350003, China

3 Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, Fuzhou University, Fuzhou 350116, China

attributes [6, 7]. For example, in general people's intuitive impressions, skirts are more likely to appear on women. Thus, there is a correlation between "female" and "skirt," which provides a constraint that acts as a visual complement to appearance recognition. Another approach is to explore the visual context as an additional source of information to assist in the recognition of human attributes. For example, most jockeys on the racecourse wear equestrian hats and riding boots. In this case, different people would share similar attributes in the same scene.

The combination of human detection and pose estimation is of great help to improve the recognition of human attributes. Human detection is an area that has been widely researched in the computational field. Human pose estimation is a top-heavy task in computer vision. In areas of high progress such as action recognition [8], gait recognition [9], human clothing attribute recognition, and video analysis [10], pose estimation of the human body plays a key role. In most cases, because of the above-mentioned interference factors, the result of direct human attribute recognition in the surveillance scene is relatively poor. Human detection helps to get the human regions under the surveillance scene.

Human clothing attributes are undoubtedly important elements, and play a vital role in the fields of surveillance safety, driverless technology, intelligent robots, etc. Related technologies such as human detection, pose estimation, image deblurring, and image enhancement are all important in the field of computer vision. The area obtained by human detection and pose estimation can greatly eliminate the interference of background factors and improve the accuracy of human clothing attribute recognition. With the recognition of human attributes, it is possible to achieve further progress in monitoring security and unmanned automatic driving etc.

Human clothing attribute recognition is equivalent to a multi-label image classification (MLIC) problem. Sequential multi-label prediction has been explored in earlier research [11, 12]. These methods are based on the CNN–RNN model. Crucially, these existing MLIC models assume: (1) the ready availability of large-scale marker training data; and (2) images of sufficient quality. These two assumptions are not valid for the recognition of human clothing attributes in surveillance images. A recent multi-person image annotation approach advances this continuous MLIC paradigm by combining interpersonal social relationships and scene background [13]. The method specifically utilizes the background of family members and friend-centric high-resolution images, but does not extend to open world surveillance scenarios of poor image data. In addition, powerful attribute level labels are required, while human clothing attributes are mostly weak labels at the image level.

Deep learning plays an important role in the study of human attribute recognition. For example, Liu et al. [14] of SenseTime proposed a deep network model HydraPlus-Net based on an attention mechanism. The features are mapped to different feature layers by multi-directionality. The attention-based depth feature obtained by HydraPlus-Net has several advantages: (1) the model can capture attention from the shallow to the semantic layer; (2) the model can construct multi-scale deep attention features, enriching the final human feature representation. Wang et al. [1] proposed a JRL model to mine the context information of human clothing attributes and the correlation between attributes to improve the recognition accuracy. JRL learns the correlation between attributes in a human image, specifically the interrelationship between attribute prediction orders.

The main contributions of this paper are as follows:

(1) Aiming to improve image quality performance under the surveillance scene, we propose an algorithm to improve image quality. The algorithm combines a multiple scale Retinex image enhancement algorithm with color restoration and a blind deconvolution deblurring algorithm.
(2) A feature descriptor incorporating multiple features is proposed to enhance the expressiveness of the clothing attribute.
(3) Aiming at the positioning deviation caused by environmental factors in the existing attribute recognition method, a human clothing attribute recognition method based on human pose estimation and multi-feature fusion is proposed.

2 Preprocessing of human clothing attribute recognition

Because the label category of this article always includes the skin, hair, and background, it is first determined whether it belongs to the clothing label or the fixed three labels. The inputs are the RGB, Lab, MR8, HOG, Boundary Margin, and Pose Margin features, and these features are combined to extract complex features and form skin–hair detection. Skin–hair detection is used in the style descriptor instead of the Pose Margin, because the goal is to find similar styles independently of the pose information, while indirectly including pose-dependent information.

A style descriptor is designed to efficiently find a similar appearance for style retrieval. For an image, the middle joint is inserted between the original 14 joints, and 24 key points are obtained. These key points are used to abstract special space descriptors of the joint parts. That is, for each patch of joint parts, we calculate the mean and standard deviation of the normalized features, including RGB, Lab, MR8, HOG, Boundary Margin, and skin–hair Detection.

2.1 Feature dimension reduction

Because the feature descriptor of the multi-feature fusion is used to extract the image features, the dimension reaches a relatively large state, and such a high feature dimension leads to difficulty in processing and calculation. After reducing the feature dimension, the overhead of the whole algorithm will be greatly reduced, and it is easier to process and use. To address this, this paper uses principal component analysis (PCA). The dimension of fusion feature extracted from each image is 39,168. In order to improve the efficiency of retrieval, we use the PCA method to reduce the dimension to 441.

The main process of PCA dimension reduction in this paper is as follows:

Step 1 The raw data are composed of a matrix X of m rows 39,168 columns;
Step 2 Subtract the average of the row for each row of X (representing an attribute field);
Step 3 Find a covariance matrix;
Step 4 Find the eigenvalue of the covariance matrix and the corresponding eigenvector r;
Step 5 Arranging the feature vectors into a matrix according to the size of the corresponding feature value from top to bottom, taking the first 441 rows to form a new matrix P;
Step 6 The matrix P is the data after dimension reduction to 441 dimensions.

2.2 Style retrieval and candidate label acquisition

Style retrieval is mainly based on a Paper Doll training set marked by Kota et al. [27] The sample size of this dataset exceeds 1 million. Each sample in the dataset has the parsed clothing labels. Therefore, establishing a K-dimension tree (KD-tree) [15] on this dataset, the samples in the dataset are stored and indexed according to different dimensions, and the samples can be located from the source database more quickly. In general, the KD-tree has two searching functions: (1) searching for K neighbors closest to a certain point, and (2) searching for all points in a certain area.

According to the first function of the KD-tree, using the fusion feature extracted by the above style descriptor as the sample feature, combined with L2-distance, we search KD-tree through the clustering method of K Nearest Neighbors (KNN) [16]. Then, the results in the dataset which cluster closest to the sample to be parsed are obtained, and the relevant meta tag information is obtained to aid in the next step of the analysis. The K value in the KNN cluster is chosen to be 25. KNN was chosen as the method for the retrieval because it shows good performance in the high recall rate prediction task, and the goal here is to eliminate the obviously unrelated clothing items while retaining most of the potential related clothing items. For this reason, we adjust the threshold so that the recall rate in training is 0.5. Additionally, because of the bias in the distribution of labels in the dataset, we use the same threshold for all projects to avoid over-fitting the predictive model.

The purpose of label prediction is to obtain a set of labels that may be related with the image to be parsed, while eliminating completely unrelated label results. Then, we remove the illogical (i.e., wrong) predictions. These removed labels will no longer be predicted. We hope to achieve the best predictive performance in the high recall system. The nearest neighbor images are obtained from the database through the KNN method, so the method of label prediction can be based on a simple vote from KNN to get clothing labels that may appear. Through the voting method of KNN, 25 samples are retrieved. In each sample, each label provides a voting coefficient. The coefficients are weighted by the reciprocal of the distance in the feature space between the labels in the samples and the labels in the images to be parsed, which constitutes the confidence of these labels. We set a threshold to remove labels whose confidence is lower than a certain range and get candidate label sets based on nearest neighbor samples.

3 Image quality improvement in surveillance scene

In this paper, the problem of poor image quality in surveillance scenes is improved to some extent by classical image deblurring and image enhancement methods. In our cognition, the image quality of a human image under a surveillance lens is usually poor and susceptible to motion blur. Therefore, before the clothing attribute recognition, improving of image resolution and eliminating the influence of partial motion blur can help to improve the overall quality of the image. In addition, the problem of color distortion and image blurring caused by photographic interference is solved to a certain extent by means of image enhancement. Combining this with image deblurring and enhancement, human clothing attribute recognition in surveillance scenes can be improved.

A simple algorithm which can improve image quality in some content is proposed in this paper. This algorithm combines a blind deconvolution deblurring algorithm [17] and a multiple scale Retinex image enhancement algorithm with color restoration [18, 19]. The image in the surveillance scene is restored, and the image is deblurred by blind deconvolution first, which improves the clarity of the image. Then the Multiple Scale Retinex with Color Restoration (MSRCR) algorithm is applied by introducing a color recovery factor. The color distortion of the image is alleviated, and the color representation of the image is enhanced to provide assistance for the subsequent attribute recognition process. Our image quality improvement method is simple, effective, and forms a streamlined processing method, which can improve the final recognition effect.

4 Pose estimation based on DCNN

Before performing pixel-level clothing attribute recognition, images are processed by a pose estimation model. It can accurately locate the human foreground area in the image. And the subsequent recognition of clothing attribute can be based on the area. We consider a top-down pose estimation method, which is an SSD-based deep human pose estimation. Images are first processed the SSD human detection model, and then, we can get the candidate human regions. Since most of the background is excluded, we only target candidate regions in the next pose estimation, in which the interference of the surrounding object environment will be greatly reduced.

In this paper, the SSD [20] human detection method based on VGG-16 [21] is used to extract the human regions and then we use DCNN [22–24] to estimate poses.

The process of the deep pose estimation based on a mixed articulated limb model is as shown in Fig. 1. Firstly, an articulated limb model is constructed in the form of a graphical model to represent the pose of the image in the dataset. Nodes represent joints, and edges represent the spatial relationship between nodes. A tree structure model [25] is introduced to transform the graphical structure of the data into a tree structure representation to improve computational efficiency.

Secondly, the image represented by the tree structure will be preprocessed, i.e., the spatial relationship model between joints will be pre-trained. A series of local image blocks representing the spatial relationship between nodes are generated with each joint as the center. In order to prevent the phenomenon of over-fitting in the training process and improve the generalization ability of training samples, mirror rotation and other operations are applied.

Because human pose estimation can be regarded as a classification problem, a K-means clustering method is used to cluster the preprocessed local image blocks according to the relative spatial position of the central joint and adjacent joint. Each joint gets T clustering results, and each cluster represents a class of spatial relationship type instances. These instances have similar relative spatial relationships. By combining these examples with the annotated pose information, the corresponding pose types of T clusters are obtained. These examples are combined to represent the mixed pose types of the nodes. These mixed pose types are used to represent the "pose types" of a joint, and the mixed articulated limb model is obtained.

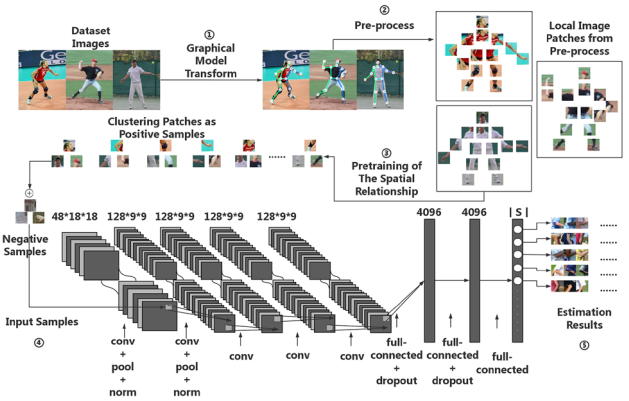

Fig. 1 Process of the deep pose estimation based on a mixed articulated limbs model

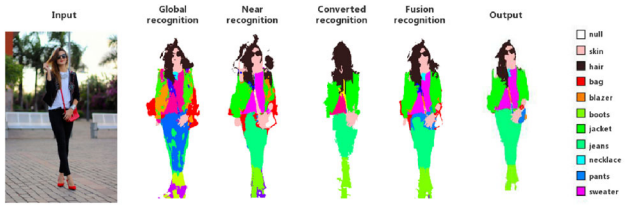

Fig. 2 Example of clothing attribute parsing

Finally, the joint local image blocks obtained from K-means clustering are used as positive examples and combined to establish a mixed articulated limb model, and the local image blocks obtained from the INRIA dataset are used as negative examples. Then, we train the DCNN model. The images are mapped to different pose types by scoring function model, and the weights and parameters are adjusted by a loss function according to the labeling information, so that the scoring result of the mapping is consistent with the actual category, the classification of pose types is completed. The multiple classification model of DCNN is obtained.

5 Human clothing attribute recognition model based on multi-feature fusion

The process of clothing attributes recognition is described here. Firstly, through the pre-trained dataset of Kota et al. [28], the appearance of the K nearest neighbor samples is retrieved. The obtained samples are weighted and voted on to obtain an initial candidate tag set. These samples are extracted to form a nearest neighbor sample set. Secondly, a global recognition model based on the same dataset is built to measure the global confidence of the pixel-label, which form an analytic label set. Thirdly, a nearest group of samples is retrieved by nearest neighbor retrieval to the training sample set, and a nearest neighbor pixel-label confidence is obtained by parsing nearest neighbor similarity. Then, using a mask conversion method similar to super-pixel segmentation-matching, the mask of the results of the nearest neighbor retrieved is match with the image to be recognized. And filter the matched super-pixel blocks according to the threshold to obtain the set of clothing labels. Finally, combining the three kinds of recognition results, and using the iterative smoothing process to further filter the highly probable label results, the final recognition results are obtained. The steps of the process are shown in Fig. 2.

The attribute parsing result is a bit like color clustering, but our method is different from color clustering. Color clustering uses only color features and cannot distinguish between two parts that belong to different parts which have similar colors. The color features are only one part of our approach. In addition to the color features, our proposed method also utilizes multi-dimensional features such as pedestrian pose, clothing texture and gradients, which can effectively distinguish pedestrian foreground and background areas in the image, and then analyze the pedestrian clothing attributes to achieve a more accurate segmentation result.

5.1 Variable definition

The pixel in the sample is defined as i, the predicted garment label of the pixel is l_i, and the complex feature of the pixel is f_i. The nearest neighbor sample set is defined as D, and the tagged label set in the sample set is defined as $\tau(D)$, representing a clothing class target label. Each analysis is defined with a mixed parameter $\Lambda \equiv [\lambda_1, \lambda_2, \lambda_3]$ to perform the final confidence combination.

5.2 Global recognition model based on logical regression

The number of clothing labels considered in this article is 56 and always includes the three labels of hair, skin, and null. In the global recognition stage, the public dataset provided by Kota et al. [28], "the Fashionista dataset," is used. (All samples of the dataset are manually labeled with the clothing label type and its corresponding pixel area.) A logistic regression classifier is trained for each clothing item. We combine them into a fusion classifier. This model calculates a likelihood estimate of assigning a clothing label to a certain pixel to indicate the confidence that a certain clothing label is assigned to a certain pixel. By sorting the confidence, we obtain the label result with the highest confidence of a certain pixel, thereby estimating the result of a preliminary label analysis.

The function model for calculating global recognition pixel-label confidence is:

$$C_{\text{global}}(l_i \,|\, f_i, D) \equiv P(l_i = t \,|\, f_i, \theta_t^g) \cdot 1[t \in \tau(D)] \tag{1}$$

The fusion feature f_i used in the recognition of this part consists of the following simple features: RGB, Lab, MR8, Boundary Margin, HOG, and Pose Margin. P denotes the logistic regression result of the given complex fusion feature f_i and the model parameter θ_t^g, indicating the probability value of a certain tag t in the sample. $1[\cdot]$ is an indication function, indicating that the tag t is one of the tag sets of the nearest neighbor sample. The model parameter θ_t^g uses the Fashionista dataset as a positive sample for training. Simple global recognition may result in the appearance of clothing labels with similar appearance types. The mechanism of label prediction we describe in Sect. 2.2 eliminates similar clothing labels to some extent.

5.3 Near recognition model based on nearest neighbor samples

Similar to the global recognition model, the near recognition model based on nearest neighbor samples is a logistic regression-based classifier for each clothing label, but the training set is the nearest neighbor samples retrieved by nearest neighbor style retrieval. Therefore, the accuracy of the final regression results is improved compared with that of global analysis.

The function model for calculating near recognition pixel-label confidence is as follows:

$$C_{\text{nearest}}(l_i \,|\, f_i, D) \equiv P(l_i = t \,|\, f_i, \theta_t^n) \cdot 1[t \in \tau(D)] \tag{2}$$

The fusion features f_i used in this part are RGB, Lab, Gradient, MR8, Boundary Margin, and Pose Margin. P represents the result of logistic regression given the complex fusion feature f_i and model model parameter θ_t^n. The model parameter θ_t^n uses the nearest neighbor sample set D as training set.

5.4 Converted recognition model based on mask conversion

Images are stored and processed as matrices. The mask likelihood result obtained by the global analysis can be divided into different super pixel regions. The converted recognition model based on mask conversion detects and extracts the area in the nearest neighbor samples that is similar to the mask likelihood results of the samples obtained by global recognition. We use global confidence as a way to convert these super-pixel regions to the image that will be recognized, and obtain the recognition results at this stage.

For the sample to be recognized and the nearest neighbor samples retrieved, the super-pixel blocks of each image region are formed by over-segmentation, and then, we match according to the appearance and pose features. The steps are as follows:

(1) For each super-pixel in the image to be recognized, the L2 pose spacing is used to find five nearest neighbor super-pixels in each retrieved image.
(2) RGB, Lab, MR8, and Gradient are computed in each super-pixel and fuse them to form complex descriptive features.
(3) Based on complex fusion features and L2 pose spacing features, the most matched super-pixel blocks in each nearest neighbor sample are obtained.

A super-pixel block is defined as s, the fusion complex feature of this super-pixel block is represented as $h(s)$, the super-pixel block in which the pixel i is located is s_i, and a sample in the nearest neighbor sample set is $d \in D$. The super-pixel block matched with the nearest neighbor sample is (s_i, d), the label set of the sample is $\tau(d)$, and Z is represents the regularization constant. The function model for calculating converted recognition pixel-label confidence is as follows:

$$C_{\text{transfer}}(l_i | f_i, D) \equiv \frac{1}{Z} \sum_{d \in D} \frac{M(l_i, f_i, d)}{1 + \|h(s_i) - h(s_i, d)\|} \tag{3}$$

$$M(l_i, s_i, d) \equiv \frac{1}{|s_i, d|} \sum_{j \in s_i, d} P\left(l_j = t | f_j, \theta_t^g\right) \cdot 1[t \in \tau(d)] \tag{4}$$

where j denotes the pixel existing in the super-pixel regions of the nearest neighbor samples, and θ_t^g is the model parameter in the global analysis. $M(l_i, s_i, d)$ denotes the average value of logistic regression results obtained by global recognition of super-pixel regions in nearest neighbor samples.

5.5 Fusion score model

The above three recognition methods measure the confidence of a label in a pixel, but a single confidence is not enough to ensure the accuracy of the label assignment result. The three recognition results are fused and analyzed, with $\Lambda \equiv [\lambda_1, \lambda_2, \lambda_3]$ defining the three recognition weight ratios, and the fusion confidence score function model is as follows:

$$C(l_i | f_i, D) \equiv C_{\text{global}}(l_i | f_i, D)^{\lambda_1} \cdot \; C_{\text{nearest}}(l_i | f_i, D)^{\lambda_2} \cdot C_{\text{transfer}}(l_i | f_i, D)^{\lambda_3}. \tag{5}$$

The parameter group $\Lambda \equiv [\lambda_1, \lambda_2, \lambda_3]$ is optimized. Defining the foreground pixel area of the image to be parsed as F, and a pixel $i \in F$, along with the definition $\widetilde{l_i}$ which represents the real label of the pixel i, the optimization function for the above fusion score model is:

$$\max_{\Lambda} \sum_{i \in F} 1\left[\widetilde{l_i} = \arg\max_{l_i} C_{\Lambda}(l_i | f_i, D)\right] \tag{6}$$

The parameter group $\Lambda \equiv [\lambda_1, \lambda_2, \lambda_3]$ which maximizes the fusion function C at a certain pixel is obtained first, and then, all the foreground region pixels are traversed to obtain the largest set of results as the result of the final optimization function. At the beginning of the calculation, the three parameters are equal to the mean value, and the optimal result of the parameter group is (0.41; 0.18; 0.39).

5.6 Maximum a posteriori allocation

In this paper, the maximum a posteriori (MAP) probability allocation model is used to iteratively smooth the pixel-level prediction label results, i.e., some redundant label results are eliminated in the predicted label candidate set, and the most likely label result is retained. Label prediction and the process of parsing an image to get the correct label assignment can be seen as a solution to multi-pixel-labeling. All pixel values and possible labels in the image are used as nodes, and labels are assigned to a pixel. The consumption and the consumption caused by different label assignments between adjacent pixels form a connection between nodes, called edges, which constitute an undirected graph. Boykov et al. [26] proposed the concept of Graph cuts algorithm, which introduced a max-flow algorithm to cut this undirected graph and obtained a min-cut result, which is the minimum consumption. The correctness of pixel-label allocation at this time is optimal. This process of finding the minimum cost is solved in terms of probability equivalent to the MAP of a Markov random field (MRF). In this experiment, when solving the MAP, we consider that the probability of a label itself is the greatest, and the probability that the label is assigned to the corresponding pixels is the greatest. Generally, the probability that the final label prediction results are correct is the largest. This concept is consistent with the above notion that a min-cut is found in graph cuts to minimize the consumption of label-pixel links, i.e., the correctness of the pixel-label allocation is optimal. Thus, the MAP allocation model is considered.

6 Experiments and result analysis

6.1 Experimental datasets

6.1.1 Basic dataset for nearest neighbor sample retrieval

The Paper Doll dataset [27] acts on nearest neighbor sample retrieval before recognition, not for training and evaluation of the model. Real images from a photo-based social networking site with metadata tags are collected, which includes clothing categories and scenes in images. The main categories used here are labeled clothing garments. Initially, they collected more than 1 million samples, but in order to exclude some negative samples, they asked for metadata labels of the samples. At least one category of clothing was included in the samples. In this paper, we ensured that the samples included a human body, rather than simple background images. The number of samples contained in the dataset is 339,797.

6.1.2 Dataset of contrast experiment

The dataset of human clothing attribute recognition used in this paper is the Fashionista dataset (FS) [28]. The Fashionista dataset is a dataset collected and annotated from a public fashion social networking website. It contains 685 samples, of which 456 are training sets, and the remaining 229 are test sets. The whole dataset contains 56 clothing label categories. Its annotation information is for the pixel level; each

pixel has its own label marking information with the clothing garment category.

6.2 Experimental evaluation method

The main evaluation indicators used in experiment are accuracy, foreground (f.g.) accuracy, avg. precision, avg. recall and avg.$F-1$. Assuming that there are two categories of classification objectives—positive and negative—the types of result obtained are as follows: (1) True Positive (TP): the number of samples correctly classified as positive; (2) False Positive (FP): the number of samples that are incorrectly classified as positive; (3) False Negative (FN): the number of samples wrongly classified as negative; (4) True Negative (TN): the number of samples correctly classified as negative.

Accuracy refers to the degree of consistency between the measured mean and the true value under certain experimental conditions. It is expressed by deviation and used to indicate the magnitude of system deviation. Its function formula is as follows:

$$\text{Avg.accuracy} = \text{mean}\left(\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}\right) \tag{7}$$

F.g.accuracy refers specifically to the accuracy in the foreground area.

Avg.precision refers to the average accuracy of the non-null pixels after the prediction, and the accuracy is the ratio of the actual positive samples in the positive samples. Its function formula is as follows:

$$\text{Avg.precision} = \text{mean}\left(\frac{\text{TP}}{\text{TP}+\text{FP}}\right) \tag{8}$$

Avg.recall (the recall rate) refers to the ratio of the actual positive sample in the correctly predicted sample. The average recall rate takes the average value, and its function formula is as follows:

$$\text{Avg.recall} = \text{mean}\left(\frac{\text{TP}}{\text{TP}+\text{FN}}\right) \tag{9}$$

Avg.$F-1$ is the harmonic average of accuracy and recall rate. Its function formula is:

$$\text{Avg.}F-1 = \text{mean}\left(\frac{2\times\text{TP}}{2\times\text{TP}+\text{FN}+\text{FP}}\right) \tag{10}$$

6.3 Analysis of experimental results

Figure 3 shows the $F-1$ score results for a selection of the tags. As can be seen from the data in the figure, the $F-1$ scores of null, skin, and hair are higher than other label categories. The standard we set in this paper is that each analysis result always has null, skin, and hair, because these three tags are certain to appear in every sample. For clothing labels, dresses, jeans, and shorts often have higher $F-1$ scores,

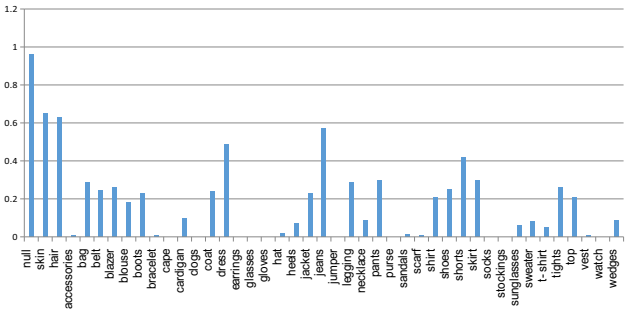

Fig. 3 $F-1$ scores of different clothing items

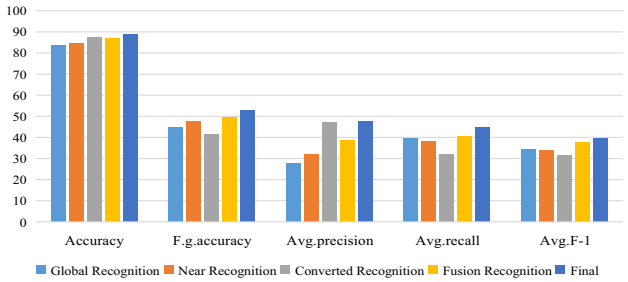

Fig. 4 Histogram of the evaluation data at different stages of the experiment

because these clothing label categories tend to appear in most samples, and the occupied pixel area is large. Clothing labels such as glasses, earrings, and other small jewelry types have a lower $F-1$ score because the ratio of the appearance of these categories is not high, and the occupied pixel area is smaller, which leads to relatively poor resolution effect.

Table 1 summarizes the results of the six evaluation indicators at different stages of analysis. The six evaluation indicators are Accuracy, F.g.accuracy (Foreground Accuracy), Avg.precision (Average Precision), Avg.recall (Average Recall), Avg.$F-1$ (Average $F-1$ score). To compare performance evaluations at different stages, different feature descriptors which are fused by different simple features are used in different stages. Global recognition and near recognition only use their respective training sets to train logistic regression classification, so they have low accuracy. From converted recognition to fusion analysis, after iterative smoothing to the final result, the local recognition results of different situations are transferred to the whole one, so the accuracy is increased. The histogram is shown in Fig. 4.

Table 2 compares the method presented in this paper with other experimental methods using FS dataset. Similarly, the measurement indicators are six evaluation indicators. In terms of accuracy, our method achieves 88.91%, while Yamaguchi [28] achieves 77.45%, with a 14.8% improvement in performance. Compared with Paperdoll [27] method, our method also improves 4.2%. In terms of average recall rate, compared with ATR (PCA) [29] method, our method improves 1.5%. The histogram is shown in Fig. 5.

Table 1 Comparisons of evaluation data at different stages of the experiment

Method	Accuracy	F.g. acc	Avg.pre	Avg.recall	Avg.$F-1$
Global	83.61	44.85	27.89	39.47	34.39
Near	84.77	47.73	32.18	38.29	34.03
Converted	87.21	41.5	47.21	31.82	31.40
Fusion	87.16	49.44	38.76	40.38	37.68
Final	88.91	52.86	47.72	44.92	39.42

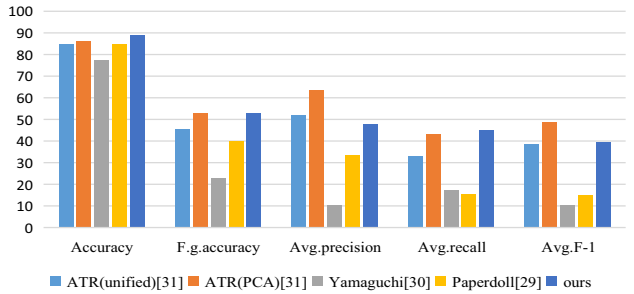

Fig. 5 The histogram of results on the FS dataset of different methods

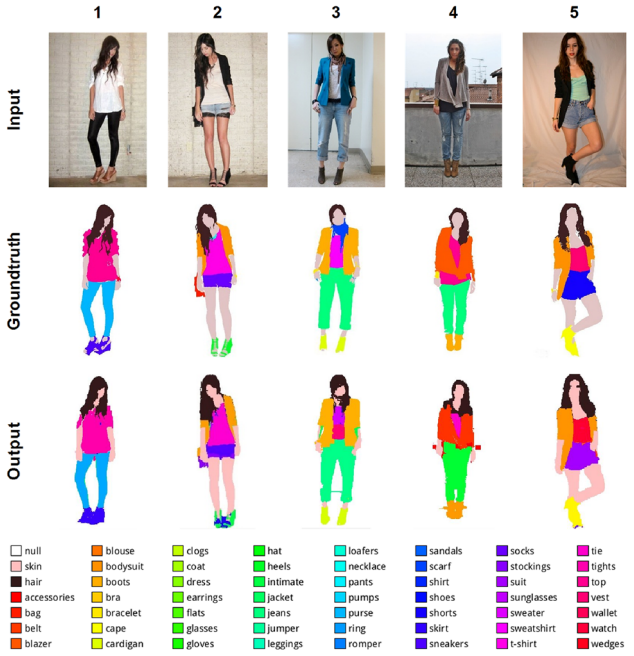

Fig. 6 Experiment results on the FS dataset

6.4 Experimental results

Figure 6 shows some of the results of the experiments in this paper, including the original input image, its groundtruth annotation and the results obtained through the recognition method in this article. The figure includes the good results and poor results due to interference factors. It can be seen that the results are basically the same as the groundtruth annotation. The result is better when the distinction between different clothes on human is obvious in the image. Moreover, when the background of the image is simple and the foreground and background area are easily defined and segmented, the recognition result is also better.

Table 2 Results on the FS dataset of different methods

Method	Accuracy	F.g.acc	Avg.pre	Avg.recall	Avg.$F-1$
ATR (unified) [29]	84.95	45.65	51.90	33.07	38.62
ATR (PCA) [29]	86.43	52.83	**63.50**	43.39	**48.87**
Yamaguchi [28]	77.45	23.11	10.53	17.20	10.35
Paperdoll [27]	84.68	40.20	33.34	15.35	14.87
Ours	**88.91**	**52.86**	47.72	**44.92**	39.42

The definition for the significance of bold is the best result in the column

7 Conclusion

This paper proposes a human clothing attribute recognition method based on multiple-feature fusion, combined with human detection and pose estimation based on deep learning, which effectively enhances the detection of a human foreground area. Before attribute recognition, the method improves the overall image quality by partly eliminating motion blurring. In addition, the problem of color distortion and image blurring caused by photographic interference factors is solved to some extent by means of image enhancement. During the process of clothing attribute recognition, an appearance style retrieval is carried out through a huge dataset, and the obtained results act as a candidate set to help the recognition. The human clothing attributes are then recognized by fusing three recognition methods and combining the results of the candidate labels retrieved. Generally speaking, the experiment of the model on the benchmark datasets shows that the model has performs and generalizes well.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 61972097, Grant 61672159, and Grant 61672158, in part by the Technology Guidance Project of Fujian Province under Grant 2017H0015, in part by the Natural Science Foundation of Fujian Province under Grant 2018J1798.

References

1. Wang, J.Y, Zhu, X.T., Gong, S.G., et al.: Attribute recognition by joint recurrent learning of context and correlation. In: IEEE International Conference on Computer Vision, pp. 531–540 (2017)
2. Ke, X., Li, J., Guo, W.: Dense small face detection based on regional cascade multi-scale method. IET Image Process. **13**(14), 2796–2804 (2019)
3. Ke, X., Zhou, M., Niu, Y., et al.: Data equilibrium based automatic image annotation by fusing deep model and semantic propagation. Pattern Recognit. **71**, 60–77 (2017)

4. Niu, Y., Lin, W., Ke, X.: CF-based optimisation for saliency detection. IET Comput. Vis. **12**(4), 365–376 (2018)
5. Kapuriya, B.R., Pradhan, D., Sharma, R.: Detection and restoration of multi-directional motion blurred objects. SIViP **13**(5), 1001–1010 (2019)
6. Li, D.W., Chen, X.T., Huang, K.Q.: Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In: Asian Conference on Pattern Recognition, pp. 111–115 (2015)
7. Wang, J.Y., Zhu, X.T., Gong, S.G.: Discovering visual concept structure with sparse and incomplete tags. Artif. Intell. **250**, 16–36 (2017)
8. Mliki, H., Zaafouri, R., Hammami, M.: Human action recognition based on discriminant body regions selection. SIViP **12**(5), 845–852 (2018)
9. Khan, M.H., Farid, M.S., Grzegorzek, M.: Spatiotemporal features of human motion for gait recognition. SIViP **13**(2), 369–377 (2019)
10. Gong, K., Liang, X.D., Zhang, D.Y., et al.: Look into person: self-supervised structure-sensitive learning and a new benchmark for human parsing. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6757–6765 (2017)
11. Ke, X., Zou, J., Niu, Y.: End-to-end automatic image annotation based on deep CNN and multi-label data augmentation. IEEE Trans. Multimed. **21**(8), 2093–2106 (2019)
12. Wang, J., Yang, Y., Mao, J.H., et al.: CNN–RNN: a unified framework for multi-label image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285–2294 (2016)
13. Li, Y., Lin, G.S., Zhuang, B.H., et al.: Sequential person recognition in photo albums with a recurrent network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5660–5668 (2017)
14. Liu, X.H., Zhao, H.Y., Tian, M.Q., et al.: HydraPlus-Net: attentive deep features for pedestrian analysis. In: IEEE International Conference on Computer Vision, pp. 350–359 (2017)
15. Zhang, S., Li, X., Zong, M., et al.: Efficient kNN Classification with different numbers of nearest neighbors. IEEE Trans. Neural Netw. Learn. Syst. **29**(5), 1774–1785 (2018)
16. Chanop, S., Richard, H.: Optimised KD-trees for fast image descriptor matching. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
17. Tao, F.U., Yang, X.M., Wu, W., et al.: Retinex-based image enhancement framework by using region covariance filter. Soft. Comput. **22**(5), 1399–1420 (2018)
18. Cai, B.L., Xu, X.M., Guo, K.L., et al.: A joint intrinsic–extrinsic prior model for Retinex. In: IEEE International Conference on Computer Vision, pp. 4020–4029 (2017)
19. Jin, M.G., Roth, S., Favaro, P.: Normalized blind deconvolution. In: European Conference on Computer Vision, pp. 694–711 (2018)
20. Liu, W., Anguelov, D., Erhan, D., et al.: SSD: single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37 (2015)
21. Qiu, J.T., Wang, J., Yao, S., et al.: Going deeper with embedded FPGA platform for convolutional neural network. In: ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, pp. 26–35 (2016)
22. Chen, X., Yuille, A.: Parsing occluded people by flexible compositions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3945–3954 (2015)
23. Chu, X., Ouyang, W.L., Li, H.S., et al.: Structured feature learning for pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4715–4723 (2016)
24. Zhong, S., Chen, T., He, F., et al.: Fast Gaussian kernel learning for classification tasks based on specially structured global optimization. Neural Netw. **57**, 51–62 (2014)
25. Chen, X.J., Yuille, A.L.: Articulated pose estimation by a graphical model with image dependent pairwise relations. In: Annual Conference on Neural Information Processing Systems, pp. 1736–1744 (2014)
26. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. **23**(11), 1222–1239 (2002)
27. Yamaguchi, K., Kiapour, M.H., Berg, T.L.: Paper doll parsing: retrieving similar styles to parse clothing items. In: IEEE International Conference on Computer Vision, pp. 3519–3526 (2013)
28. Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., et al.: Parsing clothing in fashion photographs. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3570–3577 (2012)
29. Liang, X., Liu, S., Shen, X., et al.: Deep human parsing with active template regression. IEEE Trans. Pattern Anal. Mach. Intell. **37**(12), 2402 (2015)