PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Fashion pose machine for fashion landmark detection

Ying Hu, Liqiang Xiao, Yongkun Wang, Yaohui Jin

Ying Hu, Liqiang Xiao, Yongkun Wang, Yaohui Jin, "Fashion pose machine for fashion landmark detection," Proc. SPIE 10836, 2018 International Conference on Image and Video Processing, and Artificial Intelligence, 108360Y (29 October 2018); doi: 10.1117/12.2515278

SPIE.

Event: 2018 International Conference on Image, Video Processing and Artificial Intelligence, 2018, Shanghai, China

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 06 May 2019 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Fashion Pose Machine for Fashion Landmark Detection

Ying Hu[a,b], Liqiang Xiao[a,b], Yongkun Wang[c], and Yaohui Jin[a,b]

[a]State Key Lab of Advanced Optical Communication System and Network, Shanghai Jiao Tong University, Shanghai, China
[b]Artificial Intelligence Institute, Shanghai Jiao Tong University, Shanghai, China
[c]Network and Information Center, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Combined with deep learning technologies, fashion landmark detection is an efficient method for visual fashion analysis. Existing works mainly focus on eliminating the effect of scale and background, and require prior knowledge of body structure. In this paper, we propose a fashion pose machine which is based on the location method of the landmark for human posture estimation. To increase the accuracy of fashion detection, we utilize convolutional neural network to learn the spatial structure among fashion landmarks in sequential prediction framework, which can eliminate the effect of the clothing placement and model posture on fashion landmark in the image. Our method does not require any prior knowledge of human body structure to learn the dependencies between different landmarks. We evaluated our model on the dataset of FashionAI, and the result showed that our model is 25% better than the state-of-the-art alternative.

Keywords: Landmark detection, Pose Machine, convolutional neural network

1. INTRODUCTION

Using deep learning technologies to assist the fashion industry to conduct visual fashion analysis has drawn a lot of attention in industry and academia. The related research includes fashion trends discovery[1,2] attribute prediction,[3] clothing retrieval[4]and recommendation.[5] However, it might be problematic to directly utilize deep learning techniques in these research works, because the scale and shape are seriously influenced by the various factor, such as photography angle, clothing placement and model posture. In addition, a mess background can also seriously affect the results of the fashion analysis.

To solve these problems, the fashion landmarks[6] was proposed, which refers to functional key points defined on clothes. At the same time, a branch of neural networks FashionNET[6] and Deep Fashion Alignment(DFA)[7] model that has three deep neural networks were proposed to gradually improve the estimation of fashion landmark. The Deep LAndmark Network(DLAN)[8] was designed to handle the problems of unconstrained fashion marker detection without additional bounding boxes in both training and test as well as scale variance and messy background. These methods all focus on eliminating the effect of scale and background on fashion landmark detection. Nevertheless, there are lots of fashion images presented in the form of clothes on models, the posture of the model still has a significant influence on fashion landmarks detection. Recently, a fashion grammar model called Bidirectional Convolutional Recurrent Neural Network (BCRNN)[9] combines the learning ability of neural network and domain feature grammar to grasp the kinematic and symmetric relations between fashion landmarks. However, BCRNN is a tree-structure model which needs prior knowledge of body structure.

Inspired by Convolutional Pose Machine,[10] we propose the **Fashion Pose Machine (FPM)**, which is used to estimate human posture. FPM is based on pose machine framework,[11] offering a sequential prediction framework for learning rich implicit spatial model which use the convolutional network as calculation module for prediction and image featuring. Especially, it does not require prior knowledge of body structure. FPM can efficiently locate the fashion landmark, even when clothing in the image is influenced by model's posture.

Our contributions are as following:

- We propose FPM model based on human posture evaluation method which can better cope with the influence caused by model posture diversity on key point positioning;

2018 International Conference on Image and Video Processing, and Artificial Intelligence,
edited by Ruidan Su, Proc. of SPIE Vol. 10836, 108360Y · © 2018 SPIE
CCC code: 0277-786X/18/$18 · doi: 10.1117/12.2515278

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 06 May 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

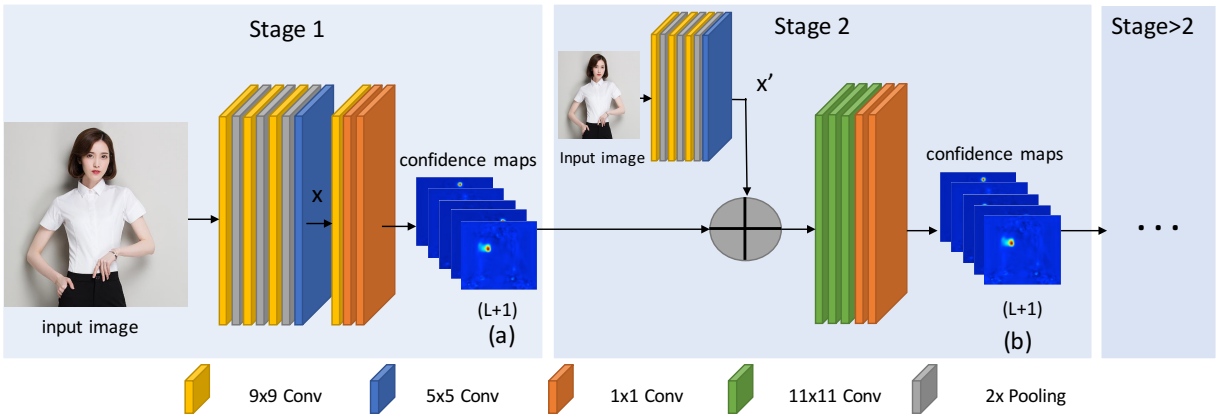

Figure 1. Architecture of Fashion Pose Machine

(a) In the first stage, only image feature x calculated by the convolutional network is taken as input, and the initial confidence maps are obtained and provided to the next stage. (b) In the subsequent stages, image feature x' and the spatial context features ψ calculated by the confidence maps from the previous stage are used as input information. In our convolutional network, the receiving context is used to represent spatial context features and new confidence maps are output which will be passed to the next stage.

- We show that FPM is a sequence inference structure that learns the interrelation of parts in implicit space, without using hand design prior.
- We evaluate FPM on the key point positioning dataset of the recent FashionAI contest held by Alibaba and compared it with the traditional FashionNET method to show the effectiveness.

2. METHOD

FPM is a model based on sequential prediction framework. By predicting the model's pose in image, pose machine helps to predict the clothing landmark. FPM is consisted of a series of convolutional networks, which creates confidence maps for every landmark at each stage and put the maps together with image features as the input for the next stage. At the same time, FPM learns rich image dependence spatial models of the relationships between landmarks through the confidential maps. Figure 1 shows the architecture of FPM. In this part, we first introduce our fashion landmark detection framework. Then, the description of each stage and the training method are presented in detail.

2.1 Framework

Considering a sequential prediction framework has a series of multiple classifiers, our method uses convolutional network as multiple classifier to predict the location of every landmark l^{th}. The location of each cloth landmark in the image is denoted as $Y_l \in Z \subset \mathbb{R}^2$, where Z is the set of all (u, v) of the location in the image. (u, v) is the coordinate of a certain location. The goal of the sequential prediction is to predict the image locations $Y = (Y_1, ..., Y_L)$ for all L cloth landmarks. In each stage $t \in \{1...T\}$ has a multiple classifier g_t. The input of multiple classifier in the first stage only contains the feature of the image data $x_z \in \mathbb{R}^d$. In the subsequent stages, the feature of image data $x_z \in \mathbb{R}^d$ and the contextual feature $\psi(\cdot)$ of the landmark former stage output are the input information of the classifiers. The multiple classifier outputs confidence in accordance with every landmark to predict the location of landmark.

The classifier in the first stage $t = 1$ produces the confidence values as following:

$$g_1(x_z) \rightarrow \{c_1^l(Y_l = z)\}_{l \in \{0,...,L\}} \tag{1}$$

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 06 May 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

where $c_1^l(Y_l = z)$ is the score predicted by the classifier g_1 for assigning the l^{th} landmark in the first stage at image location z. We represent all the confidences of landmark l evaluated at every location $z = (u, v)^T$ in the image as $c_t^l \in \mathbb{R}^{w \times h}$ as following:

$$c_t^l[u, v] = c_t^l(Y_l = z) \tag{2}$$

where w and h are the width and height of the image, respectively. For convenience, we denote the collection of confidence maps for all the landmarks as $c_t \in \mathbb{R}^{w \times h \times (L+1)}$ (l landmarks plus one for background).

For each stage t of the sequence, the confidence for the assignment $Y_l = z$ is computed and denoted by:

$$g_t(x_z^{'}, \psi_{t>1}(z, c_{t-1})) \rightarrow \{c_t^l(Y_l = z)\}_{l \in \{0,\ldots,L\}} \tag{3}$$

where $\psi_{t>1}(\cdot)$ is a mapping from the confidences c_{t-1} to context features. In each stage, the computed confidences provide an increasingly refined estimation for the location of each part. There have no specific calculation formula for contextual feature in a convolutional network framework. Thus we take the receptive field output from the convolutional network layer as the spatial context feature, and the receptive field is calculated by the confidence maps from the previous stage.

2.2 The First Stage

The initial estimation of the confidences for the location of each landmark can be obtained in the first stage of FPM. Based on the given explanation of the landmark location, the convolutional network of the first stage can get the patch through its neighbor image location. The image patch can be used to predict the confidence of each clothing landmark output. FPM utilizes five concatenated convolution and two 1×1 convolution to build a full convolutional network. We use this network to slip through the whole picture and output a vector, which represents the confidence scores of all landmarks, with the size of $L+1$ by the recurrence the local patch of every landmark.

2.3 Subsequent Stages

In the subsequent stages of FPM, the input information of convolutional network is the feature vector of images and the spatial context feature of each landmark that the former stage output, which is the receptive field. The definition of the receptive field is the mapped region in the original picture, which is produced by the pixels of feature map from the output of each stage in a convolutional network. FPM requires the receptive field to be large enough to learn the potential complexity and long-distance relevance of different landmarks.

Landmarks of relatively fixed areas like shoulder and nick are located more easily and accurately. Whereas, the areas like sleeves that relied on the human pose, are more difficult to be located. The location of sleeves can be improved by enlarging receptive field and building the special connection between landmarks. According to the confidence map built by the neighboring area of the coordinate of the landmark location, there is a lot of noise, which can provide profound and useful information. For instance, when checking the landmark of the right sleeve of a T-shirt, the right shoulder and armpit in the receptive field can provide with helpful implication. We choose to use multiple convolutional layers to achieve large receptive field on the $8\times$ downscaled heatmaps without the number of parameters.

The convolution network of subsequent stages allows multiple classifier to freely group contextual information through choosing the most predictable feature.

2.4 Training

At each stage, we define the distance between the prediction confidence map and the ideal confidence map of each part as the loss function, and the ideal confidence map of landmark l is written as $c_*^l(Y_l = z)$, which passes through each body. A Gaussian peak is created at the ground truth position of the landmark l. So the cost function for each stage is as follows:

$$loss_t = \sum_{l=1}^{L+1} \sum_{z \in Z} \|c_t^l(z) - c_*^l(z)\|_2^2 \tag{4}$$

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 06 May 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Table 1. Quantitative results for clothing landmark detection on the FashionAI datasets with normalized error(NE). Smaller values are better. L.R refers to left and right. I and O refer to in and out.

Methods	Neck L.R	C.Front	Shoulder L.R	Armpit L.R	W.line L.R	Cuff.I L.R	Cuff.O L.R
FashionNET	0.0672	0.0773	0.0676	0.1267	0.1145	0.2609	0.2713
FPM	0.0394	0.0483	0.0398	0.0981	0.0766	0.2043	0.2146

Methods	Top Hem L.R	W.band L.R	Hem L.R	Crotch	Bottom.I L.R	Bottom.O L.R	Ave
FashionNET	0.1927	0.1682	0.1145	0.2551	0.1401	0.1416	0.1636
FPM	0.1648	0.1332	0.0771	0.1791	0.1025	0.1028	0.1226

If the gradient directly drops across the entire network, the error of the output layer can be greatly reduced by multiple layers backward propagation. Therefore, the loss function is calculated at the end of each stage in this proposed scheme. And the result is calculated until the loss of each stage is calculated and stored. Consequently, the gradient is calculated with the loss of each stage as follows:

$$F = \sum_{t=1}^{T} loss_t \tag{5}$$

We use standard stochastic gradient descend to jointly train all the T stages in the network. In order to make all subsequent stage obtain the image feature $x^{'}$, we share the weights of corresponding convolutional layers across stages $t > 2$.

3. EXPERIMENTS

3.1 Dataset

The dataset we used is from the collective database of clothing key point test track of the 2018 FashionAI Global Challenge which is jointly organized by image and beauty team of Alibaba and textile and clothing department of Hong Kong Polytechnic University. The holder defined the key point of clothing set based on clothing design according to the definition of the 6 aspects of female dress (blouses, coats, trousers, skirts, dresses, jumpsuit). The 24 key points are defined as neckline left/right, center front , shoulder left/right, armpit left/right, waistline left/right, cuff in left/right, cuff out left/right, top hem left/right, waistband left/right, hemline left/right, crotch, bottom in left/right, bottom out left/right.

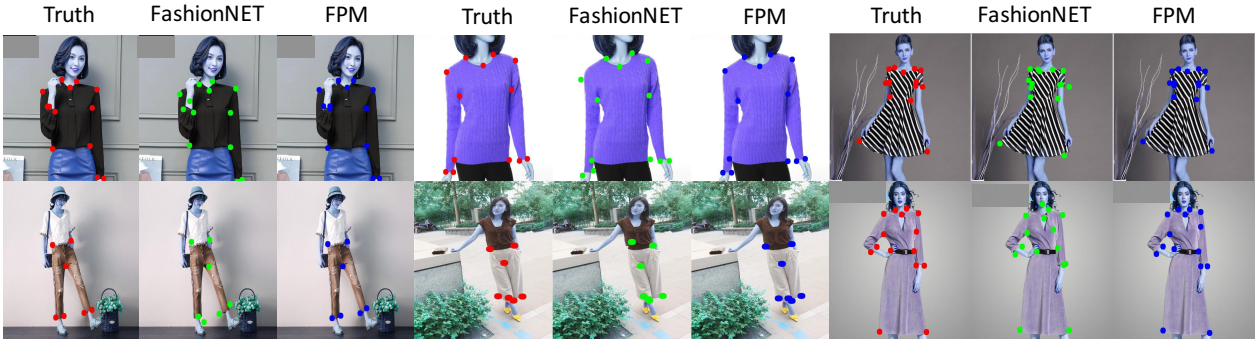

Figure 2. Qualitative results of our method on the fashonAI dataset. From left to right are true landmarks, FashionNET and FPM

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 06 May 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

3.2 Experimental Setting

In this experiment, we mainly use the data about blouses, trousers and dresses in FashionAI to test out method. Landmarks of blouse, trousers and dress can cover all 24 landmarks. 18420 fashion images are used for model training, 8107 images for validation and 5076 images for testing. We employ normalized error(NE)[7] to evaluate fashion landmark detection. NE is defined as the l_2 distance between predicted landmarks and ground truth landmarks in the normalized coordinate space.

In this experiment, we uniformly set the input image size as 512×512 and generate 64×64 heatmaps for clothing landmarks. The model is designed with 6 stages for training and testing. Our model is implemented in Python with TensorFlow, and trained with Adam optimizer. In each training iteration, we use a mini-batch of 5 images, and our whole model is trained with six epochs with 0.0001 learning rate.

3.3 Performance Evaluation

We compare our model with FashionNET,[6] and show the comparison results on the FashionAI dataset with NE score in Table 1. Comparing with the result of FashionNET, the mean NE value of FPM is 0.1226, which is 25% better than the 0.1636 of FashionNET. With regard to the cuff and bottom whose locations seriously affected by the model pose, the NE value of FPM is much lower than that of FashionNET. It is demonstrated that the effect of the model postures on landmark detection can be eliminated by estimating model's pose. The samples of landmark detection results are presented in Figure 2. It is shown that the proposed method which combines with pose estimation can obtain more accurate detection results of cuff landmarks, when the landmarks of cuff are changed with human pose.

4. CONCLUSION

We propose FPM for clothing landmark detection, which is based on the sequential inference structure. FPM merges the human pose estimation and the clothing landmark detection. FPM is an end-to-end model without the need for graphic model inference, nor the prior knowledge of body structure. We evaluated FPM on FashionAI dataset to show that FPM improves 25% accuracy of landmark location by model pose prediction.

REFERENCES

1. Kiapour, M. H., Yamaguchi, K., Berg, A. C., and Berg, T. L., [*Hipster Wars: Discovering Elements of Fashion Styles*], Springer International Publishing (2014).
2. Chen, Q., Huang, J., Feris, R., Brown, L. M., Dong, J., and Yan, S., "Deep domain adaptation for describing people based on fine-grained clothing attributes," (2015).
3. Chen, H., Gallagher, A., and Girod, B., "Describing clothing by semantic attributes," in [*European Conference on Computer Vision*], 609–623 (2012).
4. Lin, K., Yang, H. F., Hsiao, J. H., and Chen, C. S., "Deep learning of binary hash codes for fast image retrieval," in [*Computer Vision and Pattern Recognition Workshops*], 27–35 (2015).
5. Simo-Serra, E., Fidler, S., Moreno-Noguer, F., and Urta-sun., R., "Neuroaesthetics in fashion: Modeling the perception of beauty," (2015).
6. Liu, Z., Yan, S., Luo, P., Wang, X., and Tang, X., "Fashion landmark detection in the wild," in [*European Conference on Computer Vision*], 229–245 (2016).
7. Liu, Z., Luo, P., Qiu, S., Wang, X., and Tang, X., "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in [*Computer Vision and Pattern Recognition*], 1096–1104 (2016).
8. Yan, S., Liu, Z., Luo, P., Qiu, S., Wang, X., and Tang, X., "Unconstrained fashion landmark detection via hierarchical recurrent transformer networks," 172–180 (2017).
9. Wang, W., Xu, Y., Shen, J., and Zhu, S.-C., "Attentive fashion grammar network for fashion landmark detection and clothing category classification," (2018).
10. Wei, S. E., Ramakrishna, V., Kanade, T., and Sheikh, Y., "Convolutional pose machines," 4724–4732 (2016).
11. Ramakrishna, V., Munoz, D., Hebert, M., Bagnell, J. A., and Sheikh, Y., "Pose machines: Articulated pose estimation via inference machines," in [*European Conference on Computer Vision*], 33–47 (2014).

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 06 May 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use