Convolutional Neural Networks for Clothes Categories

Zhi Li, Yubao Sun, Feng Wang, and Qingshan Liu(✉)

B-DAT Lab, School of Information & Control, Nanjing University of Information
Science and Technology, Nanjing 210044, China
qsliu@nuist.edu.cn
http://bdat.nuist.edu.cn/

Abstract. Clothes classification is a promising research topic. Due to the manually-designed features' limitation, the existing algorithms have a problem of low accuracy in attributes classification. In this paper, we propose a new method to utilize convolutional deep learning for clothes classification. We firstly set up a new database by downloading the images of each category from Internet via related software and manual work, which divides clothes into 16 categories according to the common clothing style in the market. Then, the paper designs convolutional neural networks(CNNs) architecture and adaptively learns the feature representation of clothes from our constructed dataset. The experiment adopts Bag of Words (BOW), Histogram of Oriented Gradient (HOG)+ Support Vector Machine(SVM)and HSV (Hue, Saturation, Value)+SVM to test the new database and compares these methods with our CNNs model. The results demonstrate the superiority of our CNNs to the other algorithms.

Keywords: Clothes categories · Deep learning · Convolutional neural networks

1 Introduction

The 2013 annual Chinese apparel e-commerce operation report showed that variety of goods in the online market were exponentially expanding to meet customer's increasing demands, especially clothes and footwear products. In 2013, clothes and footwear products took up the highest market share in the online market, purchasing rate reached up to 76.2%. Thus, clothes and footwear products have become the most promising goods in the online market. Nowadays, most commercial image retrieval systems mainly rely on the key words search, such as TaoBao, JingDong and SuNing e-commerce. However, these systems have two weaknesses: first, every original image needs to be marked with key word. With the widely spread of smartphones, numerous images are updated everyday. It costs a large amount of human resource and materials to mark images. Second, because of cognition subjectivity, people may have different understandings of the same image, which will result in subjectivity and inaccuracy when the images are marked by different key words.

H. Zha et al. (Eds.): CCCV 2015, Part II, CCIS 547, pp. 120–129, 2015.
DOI: 10.1007/978-3-662-48570-5_12

Many researchers have devoted to designing automatic classification of clothing. Pan et al. [1] proposed a BP neural network to recognize woven fabric. Ben et al. [2] recognized woven fabric based on the texture features and SVM classifier. Yamaguchi et al. [3] described clothes by labeling superpixels, which were obtained from image segmentation making use of a Conditional Random Field model. Liu et al. [4] had a proposal for describing clothes based on pose estimation and using the features like color, SIFT and HOG and classified clothes into 23 categories. Bourdev et. al. [5] proposed a system describe the appearance of people by using 9 binary attributes such as male/female with T-shirt and long hair. For clothes segmentation, Manfred et al. [6] presented an approach for segmenting garments in fashion stores databases. Hu et al. [7] proposed a new clothing segmentation method using foreground and background estimation based on the constrained delaunay triangulation (CDT), without any pre-defined clothing model. Weber et al. [8] introduced a novel approach to get the mask of the clothing starting from a set of trained pose detectors, in order to deal with occlusions and different poses inherent to humans. Also, the clothes can be classified by the attributes, such as color, pattern, neck type, sleeve and others. Chen et. al [9] proposed a system that is capable of generating a list of nameable attributes for clothes in unconstrained images. Lorenzo-Navarro et al. [10] presented an experimental study about the capability of the LBP, HOG descriptors and color for clothing attribute classification.

However, the previous clothing categorization algorithms have been trapped in two limitations. First, the traditional features can't achieve satisfactory results, especially for similar classes. Second, there has not a public clothing database yet to evaluate the algorithms in fair. Therefore, this paper contributes to proposing the clothes classification algorithm based on deep learning and setting up a new large clothing database. We design convolutional neural networks(CNNs) architecture which adaptively learns the feature of clothes representation. In additional, we set up a new clothing database by downloading the images of each category from Internet via related software and manual work, which divides clothes into 16 categories according to the common clothing style in the market. Comparing with some traditional manually-designed features methods, our algorithm obtains a better performance.

2 Construction of Clothing Database

So far, there is no a public clothing database, and also previous works often evaluated the method in a small database. In this paper, we build a new large database. We divide the clothes into 16 categories (8 categories of menswear and 8 categories of womenswear) according to the common clothing styles in the market, and we download the images from the Internet with human labeling. As the show of table 1, the new database contains 33965 samples, and we randomly select 27565 images as the training samples(14142 menswear samples, 13425 womenswear samples) and the rest 6400 images as the validation samples(3200 menswear samples, 3200 womenswear samples). The clothes are categorized into 16 clothing categories:

Jacket, Mens shirts, Men's windbreaker , Men's suits, Ski-wear, Men's knitwear, Men's down jacket, Men's T-shirts, Cheongsam, Women's shirt, Women's Windbreaker, Women's suits, Dress, Women's fleece, Women's down jacket, Women's T-shirt. The number of each categories samples is shown in the table 2 and table 3. The figure 1 shows us samples from our database.

Table 1. Train and validation total samples

samples	Men	Women	Total
Train	14142	13423	27565
Validation	3200	3200	6400
Total	17342	16623	33965

Table 2. Men's train and validation samples

	Train samples	Validation samples	Total samples
Jacket	1587	400	1987
Men's shirts	1582	400	1982
Men's windbreaker	2204	400	2604
Men's suits	1854	400	2254
Ski-wear	1652	400	2052
Men's knitwear	1873	400	2273
Men's down jacket	1636	400	2036
Men's T-shirts	1754	400	2154
Total	14142	3200	17342

Table 3. Men's train and validation samples

	Train samples	Validation samples	Total samples
Cheongsam	1610	400	2010
Women's shirt	1662	400	2062
Women's windbreaker	1603	400	2003
Women's suits	1662	400	2062
Dress	2017	400	2417
Women's fleece	1637	400	2037
Women's down jacket	1688	400	2288
Women's T-shirt	1544	400	1944
Total	13423	3200	16623

3 CNN Based Feature Learning

Deep learning model [11] is a class of machines that can learn a hierarchy of features by building high-level features from low-level ones. Such learning machines can be trained using either supervised or unsupervised approaches, and widely used in the field of computer vision such as object detection [12],

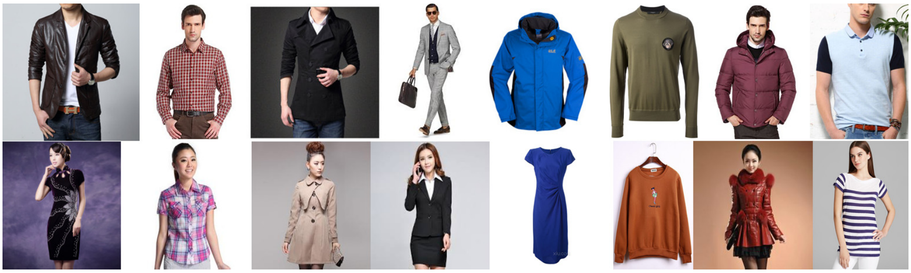

Fig. 1. Samples from the database

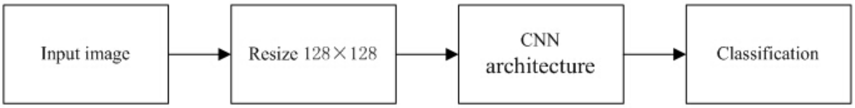

Fig. 2. Image processing for CNN

image classification [13] and image segmentation [14]. The convolutional neural network(CNN) [15] is a popular deep model in which trainable filters and local neighborhood pooling operations are applied alternatingly on the raw input images. CNN has been incorporated into a number of visual recognition systems in a wide variety of domains. CNN is previously proposed to contain many hidden layer of multilayer perceptron. By combining low-level features and discovering distributed characteristic presentation of data, deep learning forms more high-level characteristics stand by attribute categorisation and assembling. CNN attracted much attention in recent years, after obtaining much success in digit recognition [16], OCR [17] and object recognition tasks [18]. Due to the complex pattern of clothes, the common manually-designed features have the limitation of low accuracy in attributes classification. CNN can adaptively learn the high-level semantic features by the multiple layer architecture, which has the capacity to improve the performance of clothes classification. Thus, in this paper, we propose a new method to utilize convolutional deep learning for clothes classification.

From the figure 2, it can briefly show how to process image using CNN. We resize the image in the size of 128×128 for different image sizes from database. Then, the images are fed into CNN to learn network parameters. In order to improve the clothing recognition accuracy, the core is to design the effective network architecture which can learn appropriate features to represent the complex clothing appearance.

In Fig. 3, we design the architecture for our CNNs model. The architecture consists of 4 convolutions layers. We consider the image of size 128×128 as inputs to the CNN model. Then, we apply convolutions with a kernel of size 7×7, stride of 1, pad of 2 and C1 layer consists of 16 feature maps. We set pad as 2 in each convolutions in our architecture. In the subsequent subsampling layer S2, we apply 2×2 subsampling on each of the maps in the C1 layer. The next convolution layer C3 is obtained by applying convolution with a kernel

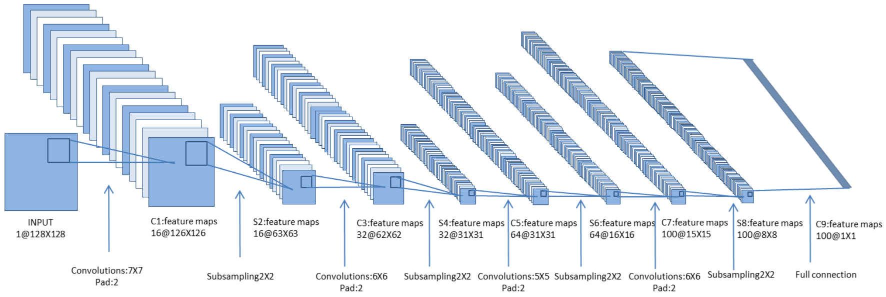

Fig. 3. The convolutional neural network architecture

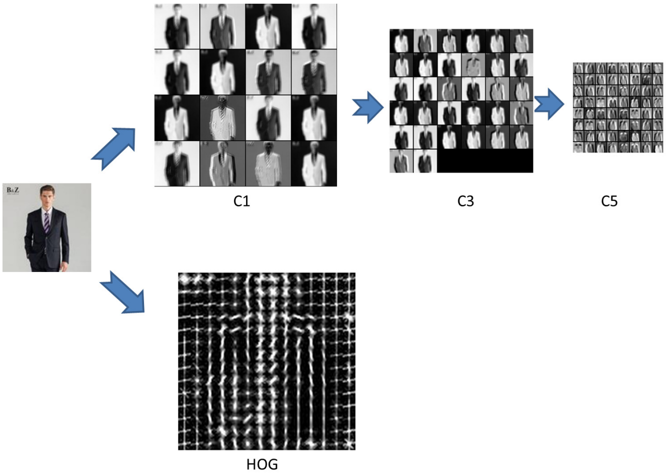

Fig. 4. Visual features learned by CNNs and HOG

of size 6×6 on each of feature maps separately. In the subsequent subsampling layer S4, we also apply 2×2 subsampling on each of the maps in the C3 layer. We set the convolution with a kernel of size 5×5 in the third layer and 6×6 in the fourth layer. The full connection layer consists of 100 feature maps of size 1×1. The size of image will be made 126×126, 62×62, 31×31, 15×15 after each convolution. Through the layers of convolution, the deep model can obtain the better features from shallow to deep.

Fig. 4 demonstrates the learned visual features by our designed CNNs. The output of C1, C3, C5 layer are displayed in the first column. The extracted HOG feature is also listed to compared with our learned features. It can be seen that HOG only represent the edge characteristic of clothes, lacking of the global pattern description and HOG features are sensitive to the noise result from HOG descriptors gradient operation. Different from the HOG features, our CNNs has the ability to abstract the features layer by layer. The features output form C5 can extract the global pattern of various clothes, not like the low-level edge information. Thus, the features leaned by our CNNs model can effectively represent the high-level semantic characteristic of clothes, which is more useful for clothes classification.

4 Experiments

4.1 Model

In order to evaluate the performance of our CNNs model, we adopt the classification accuracy as the measure criteria. Our model will also be compared with three baseline method, including BOW, HOG+SVM and HSV+SVM

BOW model is a common document representation method in image retrieval field. It consists of three steps. First, we extract visual vectors from different images by using the SIFT descriptor [19]. Second, we gather all feature points vectors together, and merge vectors with similar meaning through K-means algorithm [20]. Third, we compute the frequency that these words show up in images. Thus, these images are transformed into K-dimensional vectors. We put the feature vectors and labels into the SVM to train the classifier.

HOG initially proposes a descriptor which can implement human object detection. This method abstracts shape characteristic and movement information. In our work, we make use of a cell size of 8×8 pixels and the block is 32×32 cells.

HSV is a model that consists of three parameter: hue(H), saturation (S), value(V). The hue is measured in angle, and the range of hue is 0°∼360°. Hue counts from red counterclockwise. In this way, red represents 0°, green represents 120° and blue represents 240°. Saturation (S) varies from 0.0 to 1.0. The bigger the value is, the more saturated the color is. The range of value (V) is 0(black)∼255(white). In additional, HSV is a six pyramid model. We quantify hue into 64 intervals, and quantify saturation into 12 intervals, while value is not quantified. So we will establish a 768-order histogram.

4.2 Evaluation

Fig. 5 gives the classification results of four methods. We can see our CNN performs better than other methods and achieve the accuracy of 61.22%. HOG+SVM achieves the accuracy of 60.36% ranking in the second position. The third position is BOW with the accuracy of 56.27% and HSV+SVM performs worst with the accuracy of 20.58%.

Fig. 6 plots the accuracy curve of various method for each category. We can see the curve of our CNN compared with other curve which is overall at the top of the figure. In addition, we find an interesting phenomenon, if the accuracy of certain category is higher in CNN compared with other class, the same is happened in other three methods. On the other hand, if the accuracy of category compared with other class is lower, the same is true in other algorithms.

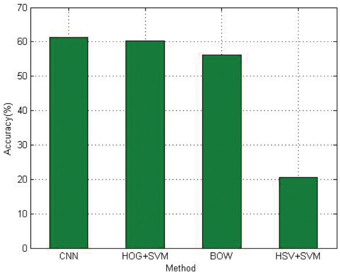

Fig. 5. The classification results of different methods

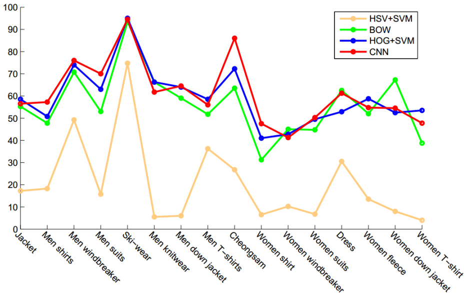

Fig. 6. Each category classification results for CNN, HOG+SVM, BOW and HSV+SVM in the database

Table 4. Detailed classification accuracy(%) for men's clothing

	CNN	HOG+SVM	BOW	HSV+SVM
Jacket	56.50	58.50	55.25	17.25
Men's shirts	57.25	50.75	47.75	18.25
Men's windbreaker	76.00	74.00	70.75	49.25
Men's suits	70.00	63.00	53.00	15.75
Ski-wear	94.50	95.00	93.50	74.75
Men's knitwear	61.75	66.25	66.25	5.50
Men's down jacket	64.50	64.00	59.00	6.00
Men's T-shirts	56.00	58.50	51.75	36.25

Detailed results of each category is shown in table 4 and table 5. We find that the accuracy of each category differs from each other. For examples, ski-wear and cheongsam accuracy are higher, and men's jackets, women's suits and women's shirts are relatively lower. We consider the main reason is that the ski-wears features of edge and color are relatively obvious and high degree of differentiation. However, the edge feature of jacket is confusing with windbreaker, suit and other kind of categories. In addition, their color features are not obvious which leads to the relatively lower accuracy.

Table 5. Detailed classification accuracy(%) for women's clothing

	CNN	HOG+SVM	BOW	HSV+SVM
Cheongsam	86.00	72.25	63.50	26.75
Women's shirt	47.50	41.00	31.25	6.50
Women's windbreaker	41.25	42.75	45.00	10.25
Women's suits	50.25	49.50	44.75	6.75
Dress	61.25	52.90	65.50	30.50
Women's fleece	54.75	58.75	52.00	13.50
Women's down jacket	54.50	52.50	67.25	8.00
Women's T-shirt	47.75	53.50	38.75	4.00

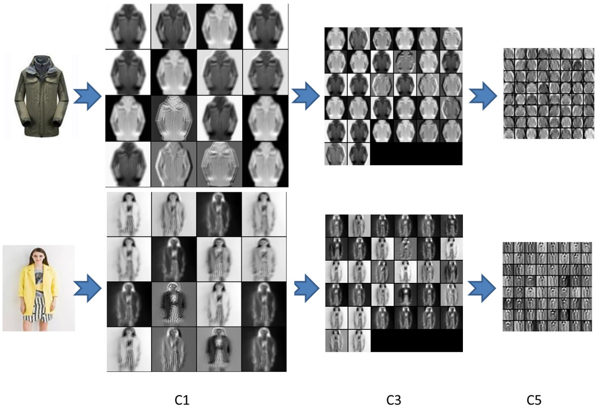

Fig. 7. Visual features of C1,C3,C5 for ski-swear and women's windbreaker

BOW is based on the regional block to extract feature, which can obtain more characteristics. However, compared with the HOG+SVM and CNN algorithm to extract the edge character, the accuracy of clothes recognition using BOW is slightly lower. But in some specific aspects such as Women's down jacket, dress and so on, it have certain advantages. The training of CNN model needs constantly iterative optimization. It can refer this iteration classification results to adjust the next iteration parameters. In addition, the convolution can capture good edge information of clothes and learn semantic feature. Therefore, the clothes with strong edge feature such as ski-wear, cheongsam and their accuracies are higher. However, for some similar clothes style, it's easily confused with each other on edge feature. Therefore, their accuracies are lower than other class. On the whole, our CNNs obtains a better performance.

4.3 Visual Analysis

Because of many clothes categories in the database, we wish to know what is the difference between these clothes categories by our CNNs. The figure 7 shows the visual features of ski-swear and women's windbreaker. We can see the original image and each features image after the C1, C3, C5. The size of original sample is 128×128. From the picture, we consider ski-wear is better than the women's windbreaker in features of edge. The profile of ski-wear's still clear even after C5 and these features can be learned easily by computer, which show better results. In contrast, the women's windbreaker doesn't show nice performance due to less strong edge features.

5 Conclusion

In this work, our convolutional neural networks obtains good results for clothes categories recognition. The experiments carry out with database which is set up by us. Our method learns the global information of image and semantic feature. The paper contributes to setting up a new clothing categories database and proposing the clothes classification algorithm based on CNN. We use the convolutional neural networks in deep learning, which can overcome the low accuracy in attributes classification. Comparing CNN with other traditional manually-designed features abstracted methods, our algorithm obtains a better performance. In future extensions of this work, we will optimize our deep networks architecture to improve the accuracy of database. In addition, database should be expanded with the increasing numbers of images furthermore, and we will publish our database in the right time.

Acknowledgments. This work was supported in part by the Natural Science Foundation of China under Grant 61272223, Grant 61300162, and in part by the Natural Science Foundation of Jiangsu Province, China, under Grant BK2012045 and Grant BK20131003.

References

1. Pan, R., Gao, W., Liu, J., Wang, H.: Automatic recognition of woven fabric pattern based on image processing and bp neural network. The Journal of the Textile Institute **102**(1), 19–30 (2011)
2. Salem, Y.B., Nasri, S.: Automatic recognition of woven fabrics based on texture and using svm. Signal, image and video processing **4**(4), 429–434 (2010)
3. Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., Berg, T.L.: Parsing clothing in fashion photographs. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3570–3577. IEEE (2012)
4. Liu, S., Feng, J., Domokos, C., Xu, H., Huang, J., Hu, Z., Yan, S.: Fashion parsing with weak color-category labels. IEEE Transactions on Multimedia **16**(1), 253–265 (2014)
5. Bourdev, L., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1543–1550. IEEE (2011)
6. Manfredi, M., Grana, C., Calderara, S., Cucchiara, R.: A complete system for garment segmentation and color classification. Machine Vision and Applications **25**(4), 955–969 (2014)

7. Hu, Z., Yan, H., Lin, X.: Clothing segmentation using foreground and background estimation based on the constrained delaunay triangulation. Pattern Recognition **41**(5), 1581–1592 (2008)
8. Weber, M., Bauml, M., Stiefelhagen, R.: Part-based clothing segmentation for person retrieval. In: Advanced Video and Signal-Based Surveillance (AVSS), pp. 361–366. IEEE (2011)
9. Chen, H., Gallagher, A., Girod, B.: Describing clothing by semantic attributes. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 609–623. Springer, Heidelberg (2012)
10. Lorenzo-Navarro, J., Castrillón, M., Ramón, E., Freire, D.: Evaluation of LBP and HOG descriptors for clothing attribute description. In: Distante, C., Battiato, S., Cavallaro, A. (eds.) VAAM 2014. LNCS, vol. 8811, pp. 53–65. Springer, Heidelberg (2014)
11. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
12. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587. IEEE (2014)
13. Ciresan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3642–3649. IEEE (2012)
14. Turaga, S.C., Murray, J.F., Jain, V., Roth, F., Helmstaedter, M., Briggman, K., Denk, W., Seung, H.S.: Convolutional networks can learn to generate affinity graphs for image segmentation. Neural Computation **22**(2), 511–538 (2010)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
16. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4), 41–551 (1989)
17. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
18. Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y.: What is the best multi-stage architecture for object recognition? In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2146–2153. IEEE (2009)
19. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
20. Sangalli, L.M., Secchi, P., Vantini, S., Vitelli, V.: K-mean alignment for curve clustering. Computational Statistics & Data Analysis **54**(5), 1219–1233 (2010)