Visual Attribute Classification Using Feature Selection and Convolutional Neural Network

Rongqiang Qian*, Yong Yue*, Frans Coenen† and Bailing Zhang*
*Department of Computer Science and Software Engineering
Xi'an Jiaotong-Liverpool University
Suzhou, P.R.China, 215123
†Department of Computer Science
University of Liverpool
Liverpool, L69 3BX, United Kingdom

***Abstract*—Visual attribute classification has been widely discussed due to its impact on lots of applications, such as face recognition, action recognition and scene representation. Recently, Convolutional Neural Networks (CNNs) have demonstrated promising performance in image recognition, object detection and many other computer vision areas. Such networks are able to automatically learn a hierarchy of discriminate features that richly describe image content. However, dimensions of features of CNNs are usually very large. In this paper, we propose a visual attribute classification system based on feature selection and CNNs. Extensive experiments have been conducted using the Berkeley Attributes of People dataset. The best overall mean average precision (mAP) is about 89.2%.**

Keywords: deep learning, convolutional neural networks, feature selection, visual attribute classification.

I. Introduction

Visual attributes are human-nameable properties (e.g., $is\ male$, $wear\ hat$ and $wear\ glasses$) that are discernible in images or videos, and they are crucially important to solving various vision problems. For example, scene representation proposed in [1] characterizes target scene by a series of attributes rather than a single label which is too restrictive to describe the properties of a scene. In [2], the face verification problem is reformulated as the recognition of the presence or absence of describable aspects of visual appearance.

For computer vision tasks, feature expression is a critical factor that affects system performance. The problem of extracting discriminative and representative features has been profoundly researched in the past decades. Due to the powerful representational learning capabilities, CNNs have been widely applied [3], [4], [5], [6], [7]. However, dimensions of CNN features are usually very large with many components irrelevant to the final tasks [7], [8]. Therefore, feature section could be exploited to remove the irrelevant or redundant features, meanwhile, improving classification performance.

The main motivation of this paper is to implement a visual attribute classification system. Inspired by R*CNN [7] that aims to improve classification accuracy by introducing secondary regions, we focus on exploring the relationships between different portions of features with regard to visual attribute tasks. The main contributions include:

- A CNN model to learn discriminative feature representation.
- A novel feature selection method to remove irrelevant or redundant features.
- A novel feature selection method to improve classification performance by reducing over-fitting.

The rest of this paper is organized as follow: section 2 focuses on related works about convolutional neural network, feature selection and visual attribute. Section 3 gives a detailed introduction of the proposed visual attribute classification system. Experimental results will be introduced in Section 4, followed with conclusion in Section 5.

II. Related Works

A. Attribute Learning

In recent years, attribute learning has become a mainstream research issue which could provide bridge between low-level features and semantic vision tasks, such as face recognition [2], Zero-Shot Learning (ZSL) [9], action recognition [10] and scene representation [1]. Different formulations of attributes have been proposed, including binary attributes and relative attributes [11]. Binary attributes are used to indicate the presence of certain properties in images. Relative attributes describe the relative strength of each attribute in images, which are closer to the ways that human describe and compare objects in real world.

B. Convolutional Neural Network

Deep learning have achieved huge popularity in recent years. In particular, the success of Krizhevsky et al. [3] on the ILSVRC-2012 image classification benchmark, led a new way of applying CNNs to tasks like image recognition and object detection. The works in [3], [4] show that CNNs are able to automatically learn discriminative feature representations. With the help of region proposals, CNN based object detection [5], [6] is significantly developed. CNNs are also adopted in action recognition [7] and scene representation [1].

C. Feature Selection

Feature selection is an important issue in many computer vision tasks. The motivations include: (i) reducing the

dimension of features by removing irrelevant or redundant ones; (ii) improving classification performance; (iii) trying to reduce computational time in both training and testing stages. Feature selection methods can be divided into two categories, which are supervised feature selection method and unsupervised feature selection method. Supervised feature selection methods [12], [13] reduce feature dimensions based on correlation information between features and labels. On the other hand, unsupervised feature selection methods [14], [15] select features mainly based on similarity preserving or clustering. However, due to the lack of label information, feature selection should be performed without guide of classification accuracy, leading particular challenge in unsupervised feature selection.

III. Approach

The proposed visual attribute classification system mainly includes three modules: feature extraction, feature selection and classification, which will be detailed in the following section.

A. Feature Extraction

A CNN is applied for feature extraction, and the adopted model is similar to the networks in Fast R-CNN [6] and R*CNN [7]. These networks are built based on the 16-layer architecture (VGG16) from [4], which have demonstrated outstanding performance in image classification and object detection. Since we only use the features before full connection layers, the last layer of our network is the region of interest (RoI) pooling layer [6]. The detail network configurations are illustrated in Table I.

TABLE I
ConvNet configurations

ConvNet Configurations		
Weight layer	VGG16	Ours
Input		
1	Convolution 3×3×64	Convolution 3×3×64
2	Convolution 3×3×64	Convolution 3×3×64
Max pooling		
3	Convolution 3×3×128	Convolution 3×3×128
4	Convolution 3×3×128	Convolution 3×3×128
Max pooling		
5	Convolution 3×3×256	Convolution 3×3×256
6	Convolution 3×3×256	Convolution 3×3×256
7	Convolution 3×3×256	Convolution 3×3×256
Max pooling		
8	Convolution 3×3×512	Convolution 3×3×512
9	Convolution 3×3×512	Convolution 3×3×512
10	Convolution 3×3×512	Convolution 3×3×512
Max pooling		
11	Convolution 3×3×512	Convolution 3×3×512
12	Convolution 3×3×512	Convolution 3×3×512
13	Convolution 3×3×512	Convolution 3×3×512
	Max pooling	RoI pooling
14	FC 4096	
15	FC 4096	
16	FC 1000	
	Soft-max	

B. Feature Selection

In feature selection stage, the features from the RoI pooling layer are collected for further refinement. The RoI pooling layer is a kind of adaptive max pooling layer, the size (7×7 in our system) of its output feature maps are fixed whatever the size of inputs. Therefore, the size of extracted features for each sample is 7×7×512 (25088 in total). Then, feature selection is performed using proposed method. For each visual attribute classifier, the details can be described as follows.

Step 1. Data collection. All the available samples in training set are divided into two classes (positive and negative) based on their labels of current attribute.

Step 2. Data processing. In order to measure the similarities of features belonging to the same class, all of the features are transformed into binary sequences using a threshold value. Since the activation function of the last convolutional layer (weight layer 13) is ReLU [3], only values larger than 0 are able to pass to next layer. This means feature positions can be considered to be activated if their values are larger than 0, thus, the threshold value used here is 0. Then, all of the sequences from same class are accumulated together and normalized by dividing by the number of samples. In this manner we got a series of sequences that indicate the probability of appearance for each feature position. Therefore, two probability sequences are achieved, namely, $p_{positive}$ and $p_{negative}$.

Step 3. Feature selection. In this step, feature selection is performed by comparing the magnitude of the probability of each position in $p_{positive}$ and $p_{negative}$. Firstly, a distance matrix can be computed based on $|p_{positive} - p_{negative}|$. Secondly, the matrix is sorted according to its magnitude. Finally, given a desired dimension n, the original 25088 feature can be reduced to n by simply select the positions that contain top n largest values in matrix.

C. Classification

In classification stage, linear SVMs are introduced [16]. Classification of SVMs are performed by constructing a hyperplane or set of hyper-planes in a high-dimensional space. With the selected features extracted from the previous stage, linear SVMs are trained to discriminate between presence or absence for each attribute.

TABLE II
Number of positive and negative labels for Berkeley Attributes of People Dataset.

	Positive	Negative
Is Male	3395	2365
Has Long Hair	1456	3361
Has Glasses	1238	4083
Has Hat	1096	5532
Has T-Shirt	1019	3350
Has Long Sleeves	3045	3099
Has Shorts	477	2020
Has Jeans	771	1612
Has Long Pants	2020	760

TABLE III
AP ON THE BERKELEY ATTRIBUTES OF PEOPLE TEST SET.

	Is Male	Has Long Hair	Has Glasses	Has Hat	Has T-Shirt	Has Long Sleeves	Has Shorts	Has Jeans	Has Long Pants	mAP
Fast R-CNN	91.8	88.9	81.0	90.4	73.1	90.4	88.6	88.9	97.6	87.8
Ours (500)	91.8	86.9	87.9	93.0	66.2	91.0	87.2	86.4	98.1	87.6
Ours (1500)	92.7	88.2	88.2	93.5	67.5	91.7	89.0	88.1	98.3	88.6
Ours (2500)	92.9	88.2	88.4	93.6	68.3	91.6	89.9	88.3	98.4	88.8
Ours (3500)	93.3	88.4	88.6	93.8	68.5	91.7	89.9	87.8	98.4	88.9
Ours (4500)	93.4	88.3	88.5	93.8	68.2	91.6	90.2	87.9	98.4	88.9
Ours (highest)	**93.4**	88.7	**88.7**	**94.0**	68.9	91.9	90.5	88.5	98.4	89.2
Poselet [17]	82.4	72.5	55.6	60.1	51.2	74.2	45.5	54.7	90.3	65.2
PANDA [7]	91.7	82.7	70.0	74.2	49.8	86.0	79.1	81.0	96.4	79.0
Gkioxari et al.[7]	92.9	**90.1**	77.7	93.6	72.6	**93.2**	**93.9**	**92.1**	**98.8**	**89.5**
R*CNN [7]	92.8	88.9	82.4	92.2	**74.8**	91.2	92.9	89.4	97.9	89.2

IV. EXPERIMENT

In this section, the implementation details of our system will be introduced. Then the Berkeley Attributes of People Dataset will be presented. Finally, the experimental results of proposed method will be illustrated and analyzed followed with the performance comparison. The details will be introduced with the corresponding experiments in the following.

A. Implementation Details

To implement our system, a computer with Xeon E3-1231 V3 CPU, 32GB memory and 6GB memory 970m GPU was employed. The program runs on a 64-bit Open-source Linux operating system with CUDA 7.5, Python 2.7.3, Matlab 2014b and Caffe deep learning platform installed.

B. Berkeley Attributes of People Dataset

The Berkeley Attributes of People dataset [17] contains 8035 images with at least a full body of a person included. 9 attributes are provided, and the detail distribution of labels are illustrated in Table II. Some examples from the dataset also have been shown in Fig. 1.

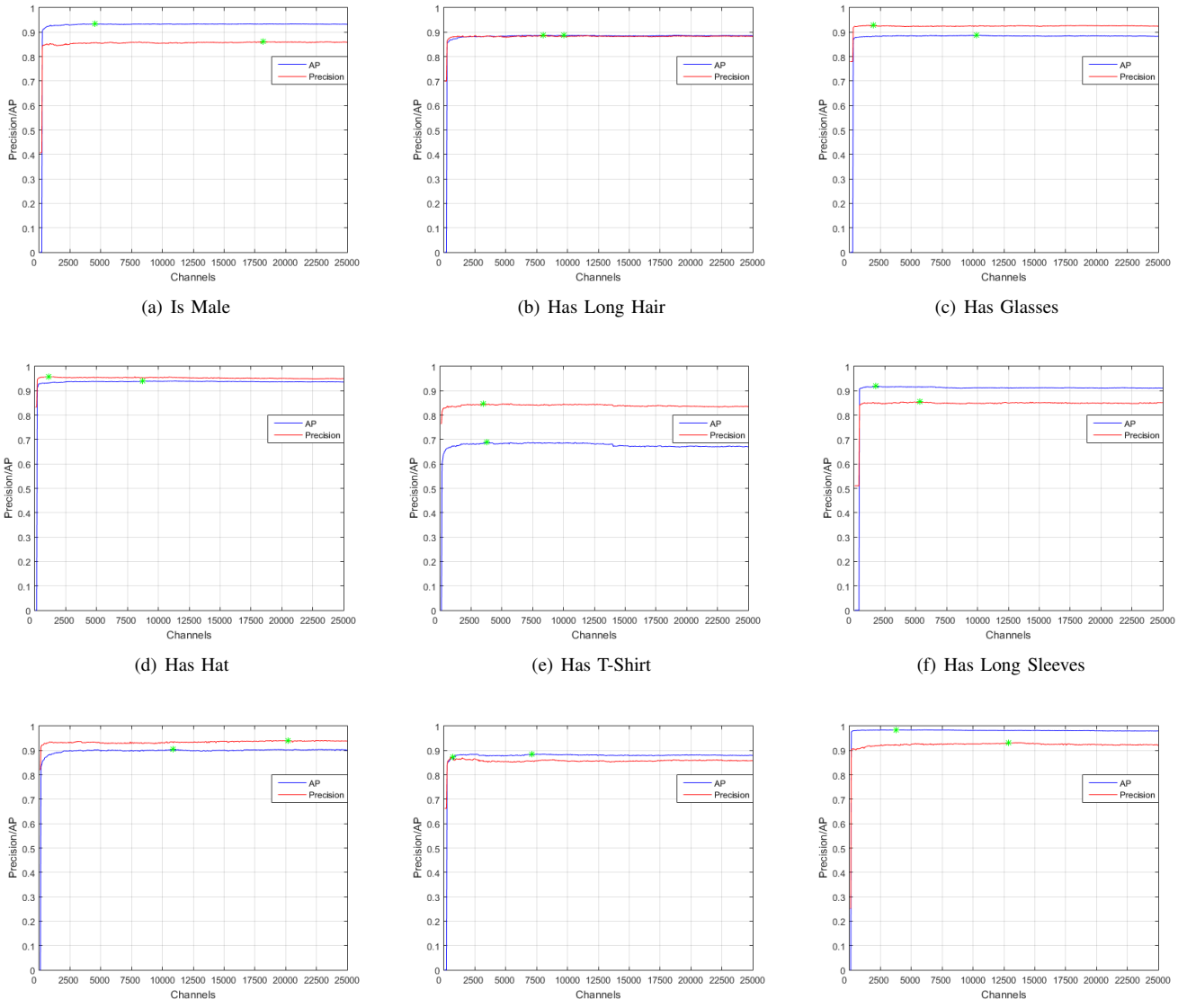
1) Experiment Setup: Following traditional training scheme, our CNN started from a model [7] initialized with discriminative pre-training for the Berkeley Attributes of People dataset, and fine-tuning was not performed for our CNN. For each sample in the dataset, only the information provided from the ground-truth region was used for the tasks of visual attribute classification.

2) Performance of Feature Selection: The objective of proposed feature selection method is to remove the redundant or irrelevant parts of the features. Thus, the curves of classifying performance versus the numbers of selected features (channels) are presented, as illustrated in Fig. 2 with all the attributes on the test set of the Berkeley Attributes of People dataset included. The sizes of selected features are set from 50 to 25088 with a step size of 50. Two measurement parameters are employed for evaluating the system performance, namely, average precision (AP) and precision. The highest values of AP and precision for each attribute task are highlighted with green stars. As Fig. 2 indicates, the classifying performance increases as the selected feature dimension (< 2500) increases,

Fig. 1. Examples from Berkeley Attributes of People Dataset. The persons in question are highlighted with a red box.

which means the discriminant part of the features have been selected. Subsequently, the classifying performance becomes nearly stable regardless of the increasing of feature dimension, which means the rest features are not such relevant to the corresponding tasks. Therefore, our experiments show that the proposed methods can reduced the feature dimension effectively.

3) Performance Comparison: Table III shows the comparative results of all the visual attributes in the Berkeley Attributes of People dataset. Since we used the pre-trained model from [7], the performance of Fast R-CNN is shown as the baseline, followed by the results of proposed method with feature selection dimensions (500, 1500, 2500, 3500 and 4500). Comparing with the baseline method, the proposed method achieves better performance for the most of tasks.



(a) Is Male (b) Has Long Hair (c) Has Glasses

(d) Has Hat (e) Has T-Shirt (f) Has Long Sleeves

(g) Has Shorts (h) Has Jeans (i) Has Long Pants

Fig. 2. Illustration of AP-channels and precision-channels curves of the attribute classifiers on the test set. The sizes of selected features are set from 50 to 25088 with a step size of 50, and the highest values of AP and precision for each attribute task are highlighted with green stars.

In particular, the tasks $Is\ Male$, $Has\ Glasses$, $Has\ Hat$ and $Has\ Long\ Sleeves$ perform obviously better. The mAP seems to stable at about 88.8% after the feature selection dimension larger than 2500, which means the proposed method can effectively reduce the feature domension to 2500 without loss of performance.

The proposed method is also compared to other approaches. As Table III indicates, we collected the highest result for each task under different feature selection dimensions, and our method obtains the best performance in the tasks $Is\ Male$, $Has\ Glasses$ and $Has\ Hat$. The maximum mAP is about 89.2%, which is the same as the result obtained by R*CNN.

4) Error Analysis: Although the proposed visual attribute classification system has demonstrated satisfactory performance on most attributes, some of the attributes are still hard to discriminate, such as $Has\ Shorts$ and $Has\ T-Shirt$. As Fig. 2(e) elaborated, the classification performance of the attribute $Has\ T-Shirt$ is especially low, the AP value is stabled at about 68.5%, which is much lower than the corresponding precision value 84.3%. Some of false examples are given in Fig. 3.

The main reasons cause false predictions include: (i) low resolution of images; (ii) partial occlusions; (iii) huge appearance variations; (iv) the representative ability of features extracted by CNN are not such discriminative. Even though the proposed feature selection method is able to improve classification performance by removing irrelevant or redundant features, the system performance will be significantly influ-

ence if the features are extracted using under-fitting models. Comparing to the baseline method, some of the attributes show decreased performance, which indicates that Multi-layer Perceptions could present better performance than Support Vector Machines for these tasks.

Fig. 3. Illustration of examples where the prediction failed for attribute $Has\ T-Shirt$.

V. Conclusion

This paper proposes a visual attribute classification system based on feature selection and CNN, with main contributions including: (i) a CNN model to learn discriminative feature representation; (ii) a novel feature selection method to remove irrelevant or redundant features; (iii) a novel feature selection method to improve classification performance by reducing over-fitting. By introducing the proposed method, the feature dimension can be significantly reduced. Moreover, the overall recall and precision rates of the system can be higher than the baseline approach. Extensive experiments have been conducted, yielding competitive results.

References

[1] J. Shao, K. Kang, C. C. Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, June 2015, pp. 4657–4666.

[2] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *2009 IEEE 12th International Conference on Computer Vision*, Sept 2009, pp. 365–372.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *2015 International Conference on Learning Representations (ICLR)*, 2015.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, June 2014, pp. 580–587.

[6] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1440–1448.

[7] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with r*cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1080–1088.

[8] V. Escorcia, J. C. Niebles, and B. Ghanem, "On the relationship between visual attributes and convolutional networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1256–1264.

[9] S. Antol, C. L. Zitnick, and D. Parikh, *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*. Cham: Springer International Publishing, 2014, ch. Zero-Shot Learning via Visual Abstraction, pp. 401–416.

[10] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, *Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, ch. Attribute Learning for Understanding Unstructured Social Activity, pp. 530–543.

[11] D. Parikh and K. Grauman, "Relative attributes," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 503–510.

[12] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint l2, 1-norms minimization," in *Advances in neural information processing systems*, 2010, pp. 1813–1821.

[13] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, "Trace ratio criterion for feature selection." in *AAAI*, vol. 2, 2008, pp. 671–676.

[14] S. Boutemedjet, D. Ziou, and N. Bouguila, "Unsupervised feature selection for accurate recommendation of high-dimensional image data." in *NIPS*, 2007, pp. 177–184.

[15] C. Maung and H. Schweitzer, "Pass-efficient unsupervised feature selection," in *Advances in Neural Information Processing Systems*, 2013, pp. 1628–1636.

[16] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008. [Online]. Available: http://dl.acm.org/citation.cfm?id=1390681.1442794

[17] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 1543–1550.