Deep Multi Class-wise Clothing Attributes Recognition for the Elderly Care Robot Environment

Chankyu Park, Minsu Jang, Jaeyeon Lee and Jaehong Kim

***Abstract*— In this study, we propose a multi-attribute-based deep neural network classification model that recognizes not only the basic clothing type but also the various sub-attributes of clothing in order to analyze clothing used as the most important factor when recognizing human appearance. This multi-attribute recognition model improves recognition performance by considering the association between multiple attribute values in order to improve classification performance with existing binary attributes. We use this technology for services where robots interact with the elderly in a home where robots care for the elderly. In particular, in order to reflect the characteristics of elderly people's dressing, images were collected from TV content that appeared a lot of elderly people and shops selling elderly people's clothes. Multiple attributes were defined with 13 top and bottom attributes and were used to train the multi class-wise model.**

I. INTRODUCTION

In this study, we introduce a recognition method using a multi class-wise attribute-based deep neural network learning technique to recognize clothing information that is important for expressing a person's figure.[1][2] We present the process by which clothing attributes can be used as a service for elderly care robot environment.

From the point of view of a person's appearance, clothing and accessories reflect various characteristics according to cultural environment, age, social status, lifestyle, and gender, and can be said to be a significant tool for expressing people and also have a great social meaning.[1][2][3] In particular, as online shopping becomes a daily routine, various latest deep learning techniques have been developed and applied in applications such as clothes search, clothes recommendation, clothes classification, and fashion trend analysis using various types and properties of clothes.[3]

This study is being developed under the premise of a service that assists the daily life of the elderly while sharing the indoor life with the elderly and the robot. The service recognizes clothes information worn by the elderly and then combines them with surrounding context information to estimate the current state. We believe that many interactions can occur between the elderly and robots through these services. Figure 1. shows in the movie "Robots and Franks" that the robot observes the main character all the time and talks to him and helps a lot. The robot needs to know what clothes to do in order to do something related to the style of clothes, but it is also necessary to understand the various detailed attributes of clothes.

For example, if the robot recognizes what clothes an elderly person is wearing, it is possible to understand whether the elderly person is properly dressed from external weather or changes in seasons. If it is not appropriate, the robot can interact with the proposal. Also, when the elderly prepares to go out, they may be able to make comments or suggestions while looking at the style of the elderly's clothes. These services enhance the interaction between the robot and the elderly, thereby improving the quality of life for the elderly and making the robot feel like a companion.

Figure 1. A scene from the movie “Robot and Frank”

In the existing deep neural network object classification and detection studies, only the object included in the image (e.g.: puppy, person, desk, TV, etc.) also focused on locating the object. A new end-to-end multi-attribute-based deep network learning paradigm is needed to understand the different types of clothes and various sub-detailed clothes properties.[6]

In the case of clothes, they can be classified into more detailed categories such as shirts, jumpers, coats, and dresses. Additionally, various attributes (e.g.: color, gender, pattern, clothing style, sleeve length, season) that can express the unique properties of the object can be defined and extracted. [1] Multi-attribute-based classification technology is widely used in clothing recognition, pedestrian character recognition, human recognition, and fine-grained image recognition.[3][6]

We designed and trained a clothes recognition model using multi-attributes based end-to-end deep convolutional neural network. The dataset used for training was gathered for clothes worn by middle-aged and elderly people. We tagged ground -truth values to reflect the various clothing attributes required for robot interaction services for the elderly. If there are multiple people in the photo, first find the ROI of the person and predict the type of clothes and various attributes for the top and bottom for each person's ROI area.[4]

Chankyu Park is with the Human Robot Interaction Research Department, Electronics Telecommunication Research Institute, Daejeon, 34129, South Korea (corresponding author to provide phone: 82-860-6708; fax: 82-860-6796; e-mail: parkck@etri.re.kr). Minsu Jang is (e-mail: minsu@etri.re.kr). Jaeyeon Lee is (e-mail: leejy@etri.re.kr). Jaehong Kim is (e-mail: jhkim504@etri.re.kr).

II. CLOTHING MULTI-ATTRIBUTES DATASET

For the robot to interact with the elderly, clothes and accessories are real facts that can be observed with the robot's camera. It can also be seen that it has a clue to estimate various situations that can occur in daily life. For example, if you are wearing outerwear or carrying a bag, you might consider going out. To predict this situation, we mainly tried to collect videos from two meaningful groups containing clothes worn by the elderly.

First, in order to reflect the real life of actually wearing clothes, we gathered images through TV dramas and contents where many elderly people appeared. Second, since the basic shape, design, and types of clothes must be reflected in various ways, photos of elderly models were gathered from the elderly clothes mall. The final clothes dataset consisted of 30% of TV content and 70% of shopping mall photos, resulting in a total of 30,000 images. In Figure 2. , (a) and (d) are pictures of the elderly model obtained from the clothes shopping mall, and (b) and (c) show pictures extracted from actual TV content.

The types of clothes were defined as 7 types of tops (shirt, jumper, jacket, vest, parka, coat, dress) and 2 types of bottoms (pants, skirt). We defined that the top or bottom can have their own color, pattern, gender, season, sleeve length, pants length, and leg posture. Thus, two clothing type classes and 11 sub-attribute classes were combined to form 13 multi-class attributes. When expressed in binary attributes, it can be regarded as composed of 69 binary attributes.

(a) (b) (c) (d)

Figure 2. Ground Truth (a) pink no pattern woman spring jacket long sleeves, brown no pattern woman spring long pants (b) red no pattern woman winter jumper long sleeves, black woman no pattern winter long pants (c) grey no pattern woman autumn coat long sleeves, brown no pattern woman autumn long pants (d) white floral woman spring shirt long sleeves, red no pattern woman spring long pants

TABLE I. defines 13 attribute classes and their multiple attributes. It consists of 6 tops, 6 bottoms, and 1 leg posture. An attribute value has the meaning of binary-class attribute (e.g.: white, checker, man, spring, shirt, short) and the attribute class itself (e.g.: color, pattern, gender, season, sleeves, class) has multi-class attributes. Figure 2. also shows the ground-truths of the top-bottom clothing attributes of the four middle-aged women included in the clothing dataset. These ground-truths were tagged by workers in accordance with the attribute definition in TABLE I.

TABLE I. 13 MULTI-ATTRIBUTES DEFINITION OF TOPS & BOTTOMS CLOTHING

Attributes	Values
color_t[1]	white, black, gray, pink, red, green, blue, brown, navy, beige, yellow, purple, orange, mixed color
pattern_t	no_pattern, checker, dotted, floral, striped, custom pattern
gender_t	man, woman
season_t	spring, summer, autumn, winter
class_t	shirt, jumper, jacket, vest, parka, coat, dress
sleeves_t	short sleeves, long sleeves, no sleeves
color_b[2]	white, black, gray, pink, red, green, blue, brown, navy, beige, yellow, purple, orange, mixed color
pattern_b	no_pattern, checker, dotted, floral, striped, custom pattern
gender_b	man, woman
season_b	spring, summer, autumn, winter
class_b	pants, skirt
sleeves_b	short, long
Leg_pose	standing, sitting, lying

1._t: tops attribute, 2. _b: bottoms attribute

III. DEEP CLASS-WISE LEARNING MODEL FOR CLOTHING MULTI-ATTRIBUTES CLASSIFICATION

Figure 3. shows the structure of the proposed deep class-wise learning model for classifying multiple attributes of clothes. This model is largely composed of two steps: a detection part that finds a person's ROI in a photo and a classifier that predicts clothes attributes by inputting features from the ROI area. The yolo-v3 model is used to detect human location.[4] After the input image passes through the yolo-v3 model, a human ROI is obtained and the yolo-v3 2D feature vector region represented by the ROI coordinates is extracted and pooled. The roi pooled feature vectors are then passed to the input of the attribute classifier.[5]

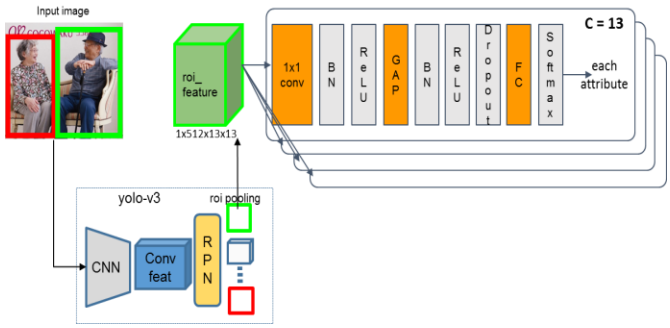

Figure 3. Deep Class-wise learning model for multi-attributes classification

The proposed multi- attribute classifier basically follows the basic concept of DeepMar's multi-attribute classifier.[6] This method is designed to maximize the relationship between multiple attributes without defining the attribute as a binary attribute classification problem. It uses the sigmoid cross entropy-based loss function defined in (1) and (2). DeepMar's classifier is a simple structure using two fc layers, but the proposed classifier consists of an optimized 1D neural network by adding several layers.[7] Each multi-attribute classifier performs additional operations such as 1x1 conv , batch normalization, relu activation, and global average pooling

before going through the last fc layer. The final layer consisted of multi-attribute classification outputs of cross-entropy softmax layer.

$$Loss_a = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}\log(\hat{p}_{ij}) + (1-y_{ij})\log(1-\hat{p}_{ij})) \quad (1)$$

$$\hat{p}_{ij} = \frac{1}{1+e^{(-x_{ij})}} \quad (2)$$

$\hat{p}_{ij}$ is the output probability for the *j*'th attribute of instance x_i. y_{ij} is the ground-truth label which represents whether the instance x_i has the *l*'th attribute. '*a*' index of $Loss_a$ means an element of $\{A_i\}$.

IV. EXPERIMENTS

We have created a dataset of 40,000 human instances from 30,000 images that must contain more than one person. This dataset was configured to use 60% for training and 10% for validation and 30% for testing. One human instance was composed of areas containing one person wearing tops and bottoms. The hyperparameters we used for training were 50 epochs, a learning rate of 0.001, and an optimizer of Adam. The evaluation metric used in this experiment is multi-class accuracy that extends the binary classification accuracy to the A_i each attribute dimension, and *i* denotes the index of the total number of attribute categories *i=1~13*.[6][7]

Figure 4. shows the experimental results for the classification performance for each clothes attribute. In particular, we found that the classification accuracy of attributes for color and season was lower than in other cases. Let's assume that the color that is obscure to the human eye comes in as input. Colors can be predicted with other adjacent color labels because ground-truth values are tagged with category labels rather than numerical values. We realized that the seasonal attributes are not clear for spring, autumn, and winter except for summer, so they can also be predicted by other adjacent seasonal labels.

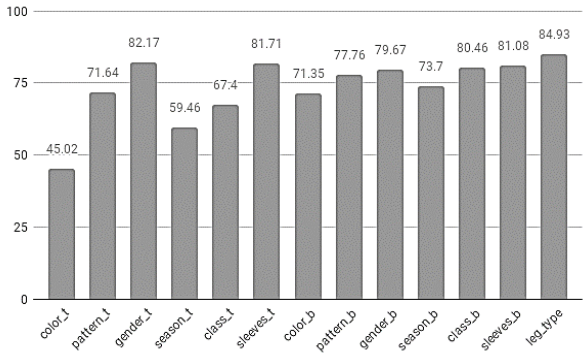

Figure 4. Experimental results for multi clothing attributes : average accuracy metric

As another method of evaluating classification accuracy for multiple attributes, if we convert 13 multiple attributes into captions and convert them into one caption prediction problem, we can apply evaluation metrics for captions such as BLEU, METEOR, ROUGE. This evaluation provides an opportunity to interpret from the point of view of a single sentence that includes all of the attributes rather than the evaluation of individual attribute.[8] TABLE II. compares the famous caption generation method using LSTM as a decoder and the multi class-wise attributes-based caption generation method. The proposed method shows more accurate caption generation than the existing LSTM method as "show and tell: A neural image caption generator".[8]

TABLE II. RESULTS OF CAPTION GENEGERATION ACCURACY METRIC

	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE_L
Caption (LSTM)	61.4	50.4	40.4	33.2	35	64.4
ours	67.2	57.2	45.5	34.2	45.2	69.1

Figure 5. Results of clothing multi-attributes prediction of two images including women

To evaluate the experimental results qualitatively, two result images are shown in Figure 5. . The left side of each result image shows the human ROI area found from the input image, and the right side shows the actual predicted multi-attribute labels. In the first image, the type of clothes was tagged as a ground-truth jumper, but predicted as a parka. This is a wrong prediction, but it can be considered as a parka, so we can see that these problems often occur in multi-attribute recognition. We can see that other multi-attributes are predicted quite accurately.

V. CONCLUSION

We propose an end-to-end multi-class-wise deep learning framework for clothing multi-attributes recognition, which can handle attributes categories that have various multiple labels. In addition, we consider the clothes worn by the elderly, which can be used as important clues or elements in the care robot environment for the elderly. Experimental results on clothing dataset for the elderly have shown that our proposed methods are effective in predicting multiple attributes on clothes accurately. In the future, we will explore new attention functions to reflect attribute's local coherent property and apply our multi-class-wise attribute learning.[9]

ACKNOWLEDGMENT

This work was supported by the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2017-0-00162, Development of Human-care Robot Technology for Aging Society).

REFERENCES

[1] Bossard, Dantone, Leistner, Wengert, Quack, and V. Gool. “Apparel classification with style”, *computer vision ACCV 2012*.

[2] H. Chen, A. Gallagher, and B. Girod. “Describing clothing by semantic attributes”, *In Proceedings of the 12th European Conference on Computer Vision – Volume Part III, ECCV'12*, pages 609–623, Berlin, Heidelberg, 2012. Springer-Verlag.

[3] Xiao Wang and Shaofei Zheng and Rui Yang and Bin Luo and Jin Tang, “Pedestrian Attribute Recognition: A Survey,” *arXiv*, 1901.07474, 2019.

[4] Joseph Redmon and Ali Farhadi, “YOLOv3: An Incremental Improvement,” *arXiv*, 1804.02767, 2018.

[5] Girshick, Ross. “Fast r-cnn.” *Proceedings of the IEEE International Conference on Computer Vision*. 2015

[6] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. “Multi attribute learning for pedestrian attribute recognition in surveillance scenarios”, *In Proceedings of the IAPR Asian Conference on Pattern Recognition*, pages 111–115, 2015

[7] Y. Deng, P. Luo, C. C. Loy, and X. Tang. “Learning to recognize pedestrian attribute”, *arXiv*:1501.00901, 2015

[8] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 3156-3164, doi: 10.1109/CVPR.2015.7298935

[9] Han, Kai and Guo, Jianyuan and Zhang, Chao and Zhu, Mingjian, “Attribute-Aware Attention Model for Fine-grained Representation Learning”, *Proceedings of the 26th ACM international conference on Multimedia*, 2040-2048, 2018.