Transfer Learning Based Approach for Semantic Person Retrieval

Takuya Yaguchi[⋆†] Mark S. Nixon[⋆]
[⋆]University of Southampton, Southampton, United Kingdom
[†]So-net Media Networks Corp., Tokyo, Japan
{ty1y17, msn}@soton.ac.uk

Abstract

Many algorithms for semantic person retrieval suffer from a lack of training data often due to the difficulties in constructing a large dataset. We therefore propose a transfer learning based approach for semantic person identification and semantic person search. We apply the fine-tuned Mask R-CNN and DenseNet-161 for detection and attribute classification. The networks were pre-trained on the MS COCO and ILSVRC 2012 datasets. Our proposed approach achieves the highest recognition rate at each rank of CMC curve for semantic person identification and the highest average localization precision for semantic person search on our validation dataset.

1. Introduction

As the number of surveillance cameras continues to increase, automatic analysis of surveillance video data has been increasingly important. Semantic person retrieval is one of the methods of automatic person identification that identifies the person who matches a textual query such as Red T-shirt and Blue Jeans. Semantic person retrieval contains two challenges: semantic person identification and semantic person search. Semantic person identification aims to identify a person from the gallery of subject images while semantic person search is the computer vision technique that locates the person from video and image data.

Semantic person retrieval is a difficult problem. One of the problems is that most surveillance images contain too few subjects, which impedes a supervised learning approach by occurence of over-fitting due to the lack of training data. For this problem, Zheng et. al. suggested the use of transfer learning model [12]. For person identification, [6] applied fine-tuned ResNet-151 model pre-trained on ImageNet Large Scale Visual Recognition Competition (ILSVRC) 2012 dataset [10]. Motivated by these approaches, we develop the transfer learning based approach for semantic person identification and search. The paper introduces our proposed approach and the result of analysis on the AVSS Challenge 2 training dataset.[1].

2. Related Works

Increasingly, deep learning is used for semantic person identification [6, 7, 8], by virtue of performance. [8] used the hand-crafted features and Extra Tree Classification (ETC) algorithm for identification while [7] used a Convolution Neural Network based approach named Semantic Retrieval Convolution Neural Network (SRCNN). These algorithms can identify 20.1% and 46.4% correct recognition rate at rank 1 for multi-shot identification respectively. However, these approaches suffered from the small number of training data, hence 3.9 % at rank 1 on zero-shot identification. To overcome this problem, [6] applied fine-tuned ResNet-152 which had been pre-trained on ILSVRC 2012 dataset [10] for attribute classification. There might still be room for improvement by using a newer classification algorithm.

For the semantic person search, there are two types of approach: detection based and tracking based algorithms. The early works were mainly based on detection based approach [11]. However, the detection-based approaches offer poor performance in a crowded or occluded situation. On the other hand, Denman et al. [1] suggested tracking based approaches, which combine detection and classification, might be able to reduce detection error. The CR based Avatar search proposed by [1] achieved the state of the art result for the dataset of [2]. However, since the detection based approach in the existing works uses handcrafted features, the performance might be improved by using state of the art algorithms for pedestrian detection motivated by DNN.

3. AVSS Challenge 2 dataset

We will use the training dataset of AVSS Challenge 2. The dataset consisted of two parts. The Task 1 dataset con-

[1] https://semanticsbsearch.wordpress.com/

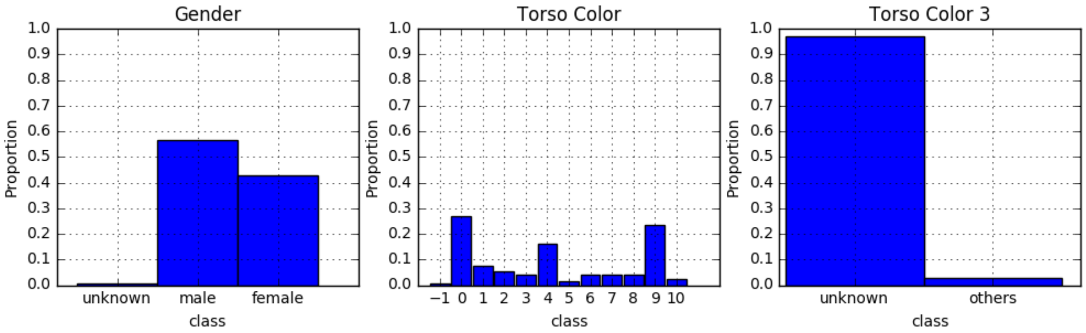


Figure 1. Example of distribution of class in attributes from Task1. The class in Torso Color is follows:-1:unknown, 0:black, 1:blue, 2:brown, 3:green, 4:grey, 5:orange, 6:pink, 7:purple, 8:red, 9:white, 10:yellow

tains pedestrian images of subjects cropped from surveillance cameras and 13 types annotations of attributes for each subject. The annotations are gender, pose, luggage and type, color and texture of torso and leg clothing. Color attributes has three annotations primary, secondary and third color for Task 1. Most of the annotation of attributes are clothing attributes. Training dataset has 520 subjects and semantic parsings of them. The training dataset of Task 2 are provided in [2] which has 110 sequences with annotation for ground truth. It contains 21-290 frames of video which include target subjects and others with 16 types of annotations for the target subject. The annotations are age, gender, build, height, skin, hair, luggage and color and texture of torso and leg clothing. There is a additional attributes from Task 1 and some class of attributes are modified. For example, the class of leg clothing type are unknown, long pants, dress, skirt, long shorts and short shorts for Task2 while it is only unknown, short and long. Therefore, transferring the data for different tasks cannot be directly conducted.

A notable fact is that some of features are imbalanced. Figure 1 shows the distribution of some attributes in Task 1. As it shows, most of third torso clothing color of subjects are annotated as -1 (unknown). Such attributes might be less effective for classification due to the imbalanced class distribution.

4. Proposed model

4.1. The model for Task 1

Our proposed model firstly predicts the attributes of subject images, and using these predicted features, we calculate the loss as the match score with the query. The attribute prediction model is motivated by [6]. The modifications from their approach are that DenseNet model is used because DenseNet is compatible with ResNet using fewer parameters and might perform better with optimal parameter tuning [4]. We applied DenseNet-161 with growth rate, $k = 48$, since this model achieved the best top-1 error rate in [4]. Our DenseNet-161 model are pre-trained on ILSVRC 2012 dataset [10] as same as [6]. After that, it is fine-tuned with the Task 1 dataset. The model uses a 224 × 224 image, hence subject images are resized to fit it. For fine-tuning, we employed stochastic gradient decent optimiser with learning rate of 10^{-3}, weight decay of 10^{-6} and a Nesterov momentum of 10^{-9}. The model predicts each attribute one by one. Thus, the last layer is replaced to new soft max layer with the class size of each attribute. We applied categorical cross-entropy as loss function. We trained models over three epochs as the accuracy of training data stabilized at the point. To increase the volume of training data, we also augment the data by applying horizontal flipping, rotation, shearing, horizontal translation and illumination shift.

For training and prediction of the classification model, global and local prediction are applied. Figure 2 illustrates the training strategy of the attribute classification method of

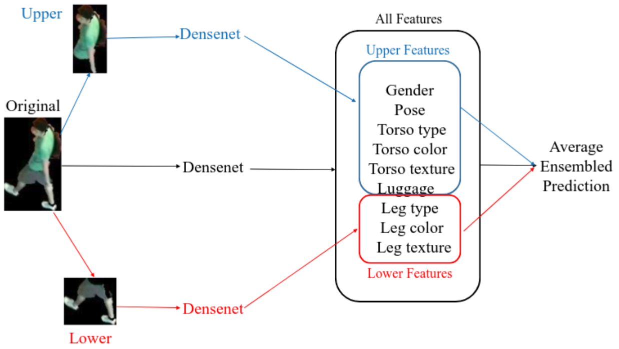
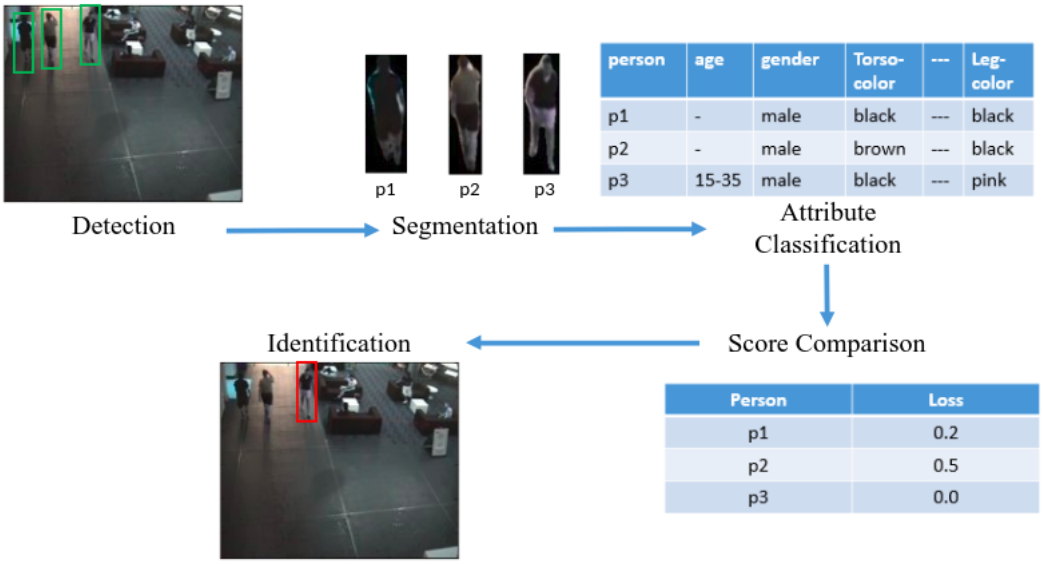


Figure 2. Overview of the attribute classification method of proposed model for Task 1. The method uses the original subjects (global) and cropped body parts (local) for training the model. We crop subjects into upper and lower parts. The upper features are gender, pose, luggage and clothing attributes of torso, and the lower features are clothing attributes of legs. After prediction, we combine global and local predictions by averaging.



person	age	gender	Torso-color	---	Leg-color
p1	-	male	black	---	black
p2	-	male	brown	---	black
p3	15-35	male	black	---	pink

Person	Loss
p1	0.2
p2	0.5
p3	0.0

Figure 3. Overview of the proposed model for Task 2

proposed model. For the local part prediction, the model divides the original figure into an upper and lower part by using provided parsing. Using these part images, we train DenseNet-161 model independently to classify upper and lower features. These upper and lower features are described in Figure 2. Finally, we ensemble the result of local and global prediction by averaging the prediction. In addition, we apply U-net [9] for segmenting subject from background.

After classification, we calculate the match score:L as follows;

$$L = \sum f(y_i^m, a^m) : m \in M \quad (1)$$

where y_i^m is prediction of the attribute m of subject i in the gallery and a^m is the attribute m of the probe. $f(y_i^m, a^m)$ calculates hamming loss as follows;

$$f(y_i^m, a^m) = \begin{cases} 1 & y_i^m = a^m \\ 0 & (otherwise) \end{cases} \quad (2)$$

We consider the Hamming loss as a match score where 0 describes a perfect match with a probe.

4.2. The model for Task 2

Our proposed model for semantic person search is based on the detection based approach. Figure 3 describes our proposed model. The pipeline of the approach is person detection, segmentation, attribute classification, score comparison and finally identification. For the person detection and segmentation, Mask R-CNN is applied since this approach provides detection and segmentation syntactically [3]. Since the parsing is not provided for the Task 2 dataset, our Mask R-CNN is trained using MS COCO dataset [5] and we only consider the prediction of person. We resize the original image of Task2 dataset into 1024 × 1024 since the image in MS COCO is this size. For the attribute classification, we will apply fine-tuned DenseNet-161. Since the limited computation facility, we only apply global feature prediction for this task. For calculating match score of the textual query, hamming loss is calculated by the same manner for Task 1. After that, we will identify the target subject by assuming the subject who has the minimum loss is the target subject.

5. Experimental results

5.1. Task1

Using AVSS Challenge 2 Task 1 training dataset, we compare the semantic identification algorithm in related works with our proposed approach. The algorithms include Extra Tree Classifier (ETC) [8], SRCNN [7], fine-tuned ResNet [6] and Channel Representational (CR) Avatar Search [1]. For CR Avatar Search, we use the equation to calculate the similarity of the foreground to the template in [1] as match score as follows;

$$S_f = \sum_n^N \sum_x^X \sum_y^Y |C_n(x,y) - I(x,y)| \quad (3)$$

where X and Y are height and width of the image and N is the total number of channels. $C_n(x,y)$ is channel representation and $I(x,y)$ is the channel representation of likelihood of the pixel being n_{th} channel representation.

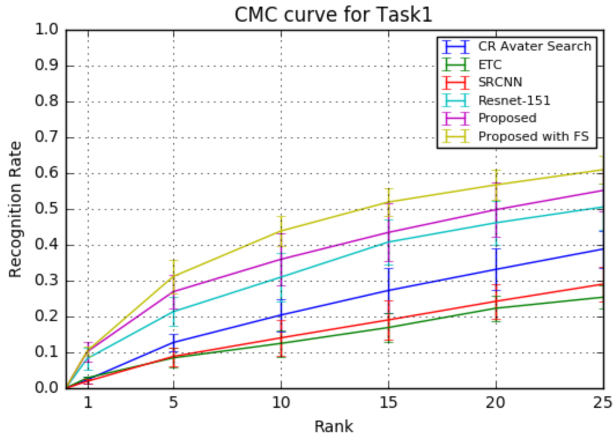


Figure 4. CMC curve for Task1

To emulate the test dataset, we apply 3-fold cross-validation on training dataset and segment training dataset by U-net algorithm instead of using provided semantic parsing. By applying 3-fold cross-validation, 350 subjects are used for training and 170 subjects are for validation. Since some attributes are considered to affect worse for identification, we apply greedy forward selection for feature selection for our proposed approach. By greedy forward selection, we only used seven features: gender, pose, torso type, torso texture, torso primary color, leg type and leg primary color.

Figure 4 shows the CMC curve of each method and Table 1 shows the recognition rate at each rank. This shows our proposed approach achieves a higher recognition rate among the methods in related works and our proposed approach with feature selection (FS) achieves the highest recognition rate among the others. Also, our proposed method achieves a minimum equal error rate (EER) as Table 2 shows.

Method	r=1	r = 5	r = 10	r = 20	r = 50
CR Avatar [1]	2.3	12.7	20.3	33.1	58.6
ETC [8]	2.9	8.5	12.5	22.3	42.3
SRCNN [7]	1.9	8.9	14.0	24.2	46.3
ResNet-152 [6]	8.3	21.3	31.0	46.1	66.9
Proposed	10.2	26.9	35.9	49.8	73.0
Proposed with FS	**10.4**	**31.1**	**43.8**	**56.7**	**77.9**

Table 1. Recognition rate at each rank of different methods for Task1

Method	EER
CR Avatar [1]	36.3
ETC [8]	43.3
SRCNN [7]	40.2
ResNet-152 [6]	30.9
Proposed	26.9
Proposed with FS	**24.6**

Table 2. EER of different methods for Task1

5.2. Task 2

We train our proposed model with different strategies. The first strategy is using Task1 dataset. The second strategy is using Task 2 dataset. To use the Task 2 dataset to train our model, we crop the ground truth randomly 20 times from frames of provided sequences and construct training dataset. Since feature selection performs better for Task1, we also apply greedy forward selection of features. By this process, 4 features, torso primary color, leg primary color, leg type and torso texture, from Task 1 and 8 features, torso primary color, leg type, gender, torso secondary color, leg primary color, height, torso type and hair, from Task2 was selected. To emulate the test dataset, we apply 3-fold cross validation same as Task1. We tried a fusion approach which merges the prediction of DenseNet-161 trained from Task 1 and Task 2. For fusion, we normalize the match score between 0 to 1 and average the summed loss. Later we also fused DenseNet-161 and CR avatar. For this fusion, we also normalize the match score, S_f of CR avatar between 0 to 1 and averaged the summed loss.

The performance of algorithms is measured by

1. Equal Error Rate (EER) of Receiver Operator Characteristic (ROC) curve.

2. average precision of localization.

To obtain EER, we use the cropped ground truth of the subjects which is not used for training the classifier. The EER of CR avatar search is obtained in the same method used in Task1.The localization precision is calculated using the same method as [1], which enables to compare with the result of [1].

Figure 5 shows the ROC curve for Task 2 and shows that CR avatar search performs better for classifying subjects than our proposed approach unlike the result of Task 1. CR avatar search achieves a minimum EER as shown in Table 3. This is due to two reasons: i) the Task 2 dataset contains a too small number of the subject for training DenseNet-161. By 3-fold cross-validation, we only use around 70 subjects; and Ii) not all of the features of Task1 can be transferable to Task 2. This result indicates our approach might not have

Method	EER
CR Avatar	**28.7**
Proposed using Task1 data	31.9
Proposed using Task2 data	30.4

Table 3. EER of different methods for Task2

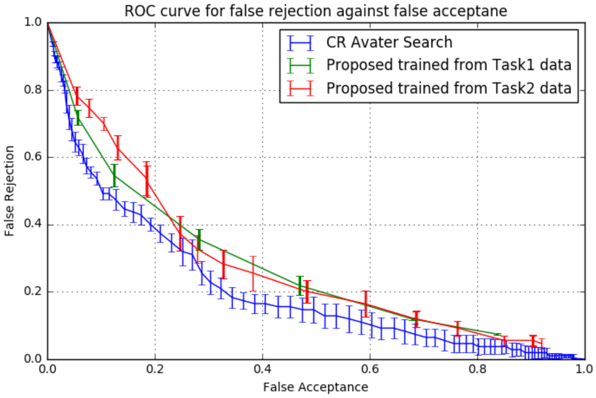


Figure 5. ROC curve of each method for Task2

better ability to locate a person. However, this result suggests that the fusion of fine-tuned DenseNet-161 and CR avatar might improve the result rather than using these independently.

Table 4 shows the overall IoU of all 110 sequences in the training dataset. This shows our propose approached are compatible with the baseline (the result of [1]) if they are used independently. The fusion of our proposed approach achieved higher average IoU than the baseline by 2%. Detection based CR avatar using Mask R-CNN also achieves a higher result than baseline. Finally, the fusion approach of fine-tuned DenseNet-161 and Mask R-CNN based CR avatar achieves the highest average IoU among the others.

Figure 6 illustrates the percentage of sequences above a given accuracy level. This shows the 55% of sequences of our proposed approach achieves 40% IoU which is compatible with the default approach of [1] and 70% of sequences of the fusion of DenseNet-161 and CR avatar achieves more than 40% IoU which is compatible with the best approach of [1]. In addition to this, 20% of the sequences of our proposed approach achieves more than 70% IoU while only few sequences in the methods in [1] achieves more than 70% IoU. This means our approach can track target subject longer than the approach of [1], which contributes to better overall IoU than [1].

Method	IoU
Baseline [1]	44.0
Proposed using Task1 data	44.9
Proposed using Task2 data	41.8
Mask R-CNN + CR	49.3
Proposed fusion(Task1 + Task2)	46.2
Proposed fusion(DenseNet + CR)	**51.1**

Table 4. Overall IoU of different methods for Task2

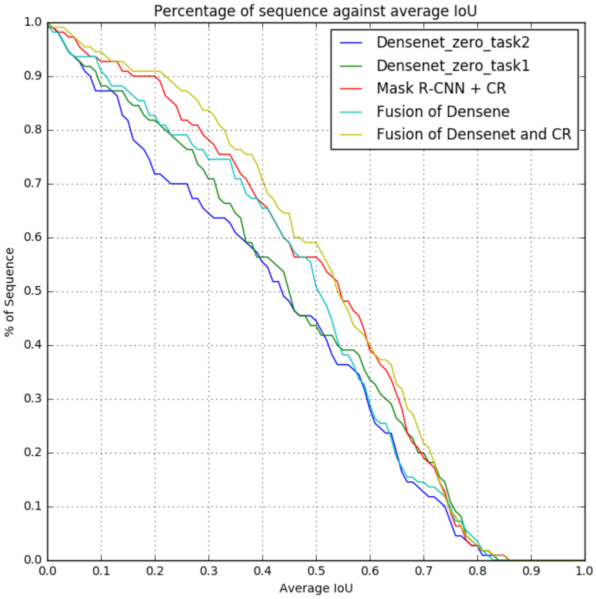


Figure 6. The percentage of sequences above a given accuracy level

6. Conclusions

We have proposed new approaches for semantic person identification and search. The results show that our proposed approach achieves a better performance on semantic person identification and search than existing approaches. For Task 2, CR Avatar achieves a better EER than fine-tuned DenseNet-161. However, for localization, our proposed approach is compatible with the state of the art algorithm, and using fusion achieves a better performance. This suggests that the fusion of transfer learning based DNN approach and hand-crafted based approach can perform better on the small dataset. Further, there are more sophisticated fusion approaches and these could be investigated in the future.

References

[1] S. Denman, M. Halstead, C. Fookes, and S. Sridharan. Searching for people using semantic soft biometric descriptions. *Pattern Recognition Letters*, 68:306–315, 2015.

[2] M. Halstead, S. Denman, S. Sridharan, and C. B. Fookes. Locating people in video from semantic descriptions: A new database and approach. In *Proceedings of the 22nd International Conference on Pattern Recognition*, pages 4501–4506. IEEE, 2014.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[4] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017*

Method	r=1	r = 5	r = 10	r = 20	r = 50
CR Avatar [1]	2.5	9.6	15.3	27.5	49.4
ETC [8]	1.0	10.2	14.7	25.5	49.4
SRCNN [7]	0.5	4.0	5.1	12.7	27.5
ResNet-152 [6]	11.7	30.1	45.4	56.1	78.5
Proposed with FS	**19.3**	**42.8**	**52.5**	**65.8**	**81.6**

Table 5. Recognition rate at each rank of different methods for Task1 test data

Method	IoU
Proposed fusion(DenseNet + CR)	**51.1**

Table 6. IoU of the proposed method for Task2 test data

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2261–2269. IEEE, 2017.

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[6] D. Martinho-Corbishley, M. Nixon, and J. Carter. Superfine attributes with crowd prototyping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14, 2018.

[7] D. Martinho-Corbishley, M. S. Nixon, and J. N. Carter. Retrieving relative soft biometrics for semantic identification. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 3067–3072. IEEE, 2016.

[8] D. Martinho-Corbishley, M. S. Nixon, and J. N. Carter. Soft biometric retrieval to describe and identify surveillance images. In *Identity, Security and Behavior Analysis (ISBA), 2016 IEEE International Conference on*, pages 1–6. IEEE, 2016.

[9] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[11] J. Thornton, J. Baran-Gale, D. Butler, M. Chan, and H. Zwahlen. Person attribute search for large-area video surveillance. In *Technologies for Homeland Security (HST), 2011 IEEE International Conference on*, pages 55–61. IEEE, 2011.

[12] L. Zheng, Y. Yang, and A. G. Hauptmann. Person reidentification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.