A Cross-modality and Progressive Person Search System

Xiaodong Chen[1†], Wu Liu[2*], Xinchen Liu[2], Yongdong Zhang[1], Tao Mei[2]
[1] University of Science and Technology of China, Hefei, China
[2] AI Research of JD.com, Beijing, China
cxd1230@mail.ustc.edu.cn,{liuwu1,liuxinchen1,tmei}@jd.com,zyd73@ustc.edu.cn

ABSTRACT

This demonstration presents an instant and progressive cross-modality person search system, called "CMPS". Through the system, users can instantly find the lost children or elderly persons by simply describing their appearance through speech. Unlike most existing person search applications which have to cost much time to find the probe images, CMPS will save more valuable time in the early stage of losing. The proposed CMPS is one of the first attempts towards instant and progressive person search leveraging the audio, text, and visual modalities together. In detail, the system first takes the speech that describes the appearance of a person as the input to obtain a textual description by speech-to-text conversion. Then the cross-modal search is performed by matching the textual embedding with the visual representations of images in the learned latent space. The searched images can be used as candidates for query expansion. If the candidates are not right, the user can quickly adjust their description through speech. Once a right image is found, the user can directly click it as a new query. Finally the system will give the complete track of the lost person by once-click. On the built CUHK-PEDES-AUDIOS dataset, the system can achieve 82.46% rank-1 accuracy in real-time speed. Our code of CMPS is available at https://github.com/SheldongChen/Search-People-With-Audio.

ACM Reference Format:
Xiaodong Chen, Wu Liu, Xinchen Liu, Yongdong Zhang, Tao Mei. 2020. A Cross-modality and Progressive Person Search System. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3394171.3414455

1 INTRODUCTION

Person search or re-identification (Re-ID) is an important and challenging task in the multimedia and computer vision communities. With wide real-world applications such as intelligent video surveillance, smart retailing, etc. [17, 24], this task aims at searching for the same person captured by multiple non-overlapping cameras. It has achieved excellent results due to the deep learning-based model and large-scale labeled data [10, 17–21].

Figure 1: The workflow of our cross-modality system. 1) the system first takes the speech or text that describes the appearance of a person as the input; 2) the searched images can be used as candidates for query expansion of speech and text; 3) If the candidates are not right, the user can quickly adjust their query; 4) once right images are found, the user can directly click them as new query; 5) the system will give the complete track of the lost person by once-click.

However, existing person search or Re-ID methods usually use images of a specific person as the probe [7, 14], which has limitations in real-world urgent scenarios. For example, when a mom lost her child in the mall, she knows clearly what her child is wearing today, but she cannot provide a photo. The time is very urgent in the early stage of losing. The fastest method is directly describing the child through speech or text description, then find more candidate images as query expansion. This is a common situation in real-life, while traditional Re-ID methods often neglect this situation. Although there have been some methods [4, 11, 12] for text-based person search in recent years, they only consider text as input. Speech is faster and more convenient than inputting text. Therefore, in this paper, we develop a simple, convenient, and real-time person search system based on multi-modality interaction with speech, text, and image as input.

Due to the difficulty of fine-grained cross-modal matching, cross-modal person search faces several challenges, such as understanding of long sentences, accurate speech recognition, compatibility with diverse inputs, and so on. Fortunately, with the gradual maturity of natural language processing and speech recognition, and the remarkable development of computer vision[3, 6, 15, 22], we can design a progressive cross-modality system, which is easy-to-use and effective for this task. To this end, we design a cross-modal person search system of which the workflow is shown in Figure 1. This system has several featured properties: 1) It provides a convenient input and interaction mode, which takes the audio of speech as the input to search for a target person captured by cameras. By this

*Wu Liu is the corresponding author.
†This work is done when Xiaodong Chen is an intern at JD AI Research.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10.
https://doi.org/10.1145/3394171.3414455

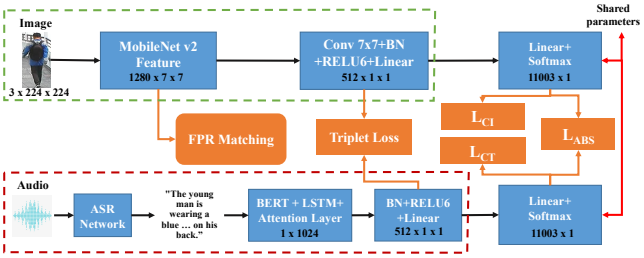

Figure 2: Architecture of the proposed network.

means, users will not only interact with the system efficiently by speech instead of typewriting, but also provide more fine-grained and detailed descriptions of the target person. 2) This system performs person search in a progressive manner to guarantee both the accuracy and speed, users can interactively input new queries and query expansion, which makes it able to find more accurate results with less time consumption[13].

2 METHODOLOGY

For solving this task, we design a new pseudo-siamese network which learns visual features and vocal features by two parallel subparts, as shown in Figure 2.

The first subpart is a Convolutional Neural Network (CNN), as shown in the green dotted box in Figure 2. We use MobileNet V2 [16] as the backbone of this subpart for extracting visual features due to its efficiency. After the feature part of MobileNet, we replace the pooling layer with the conv layer. To search persons with the image as the probe, we match images with the Foreground-aware Pyramid Reconstruction (FPR) [8], which can alleviate the mismatching caused by occlusion.

The second subpart adopts a Recurrent Neural Network (RNN), as shown in the red dotted box in Figure 2. We first use DeepSpeech2 [1] for speech recognition and use BERT [5] to get word embedding of the description. Then we feed the word embedding into Long Short-Term Memory [9] and the attention layer. Instead of softmax, we use sigmoid as a part of our attention layer, because it can solve the matching error better caused by sentences of different lengths.

Based on the above pseudo-siamese networks, we design a special loss function to optimize the whole network.

$$L = L_{CT} + L_{CI} + L_{ABS} + L_{triplet} + L_{FPR}, \tag{1}$$

$$L_{ABS} = \mathbf{E}(\sum_i (P_T \cdot |P_T - P_I|) + \sum_i (P_I \cdot |P_I - P_T|)), \tag{2}$$

where L_{CT} and L_{CI} are the cross-entropy loss functions for images and text, respectively. L_{ABS} is used to balance the optimize of the RNN subpart and CNN subpart. P_T is ID label of text and P_I is ID label of images. L_{FPR} [8] is FPR matching loss and $L_{triplet}$ [23] is triplet loss to learn the embedding space of speech and images.

3 THE SYSTEM AND EVALUATIONS

The interface of our cross-modal system is shown in Figure 3. In our system, all models are trained on the CUHK-PEDES-AUDIOS dataset. In this dataset, we generate an audio file by Text to Speech for every text in CUHK-PEDES [12] and crop the images to 224×224 for the system.

Our model is trained on one TITAN RTX GPU for 100 epochs. The ASR network is fine-tuned on the CUHK-PEDES-AUDIOS dataset

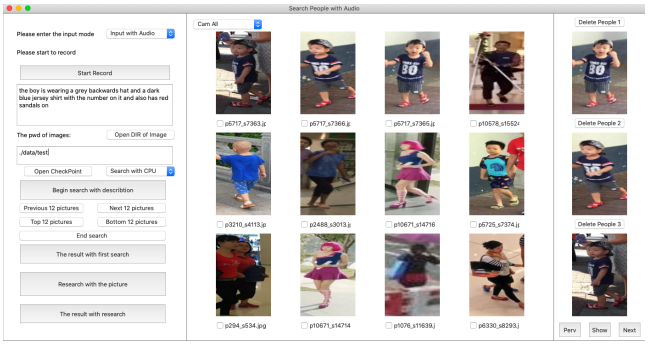

Figure 3: The interface of our cross-modal system. On the left of the interface is the input area, and when beginning a search, we can use different forms of input as probe. In the middle of the interface is the search result, and on the right is the candidate characters to show the tracks.

Table 1: The experimental results.

Method	Rank1	Rank10	mAP
deeper LSTM Q+norm [2]	17.19	57.82	-
GNA-RNN [12]	19.05	53.64	-
Latent Co-attention [11]	25.94	60.48	-
PWM-ATH [4]	27.14	61.02	-
CMPS-Audio-Fast*	35.03	70.14	34.61
CMPS-Audio	40.00	73.91	38.11
CMPS-Audio-Image**	82.46	95.93	76.12

* CMPS-Audio-Fast uses Word2Vec as word embedding.
** CMPS-Audio-Image uses image as probe.

for 10 epochs. The experimental results are listed in table 1. In the dataset, each picture has several descriptive sentences, so when calculating our Rank and mAP, we take the average value of this sentences. We referred to the previous papers and all the evaluation indicators. From the results, we can observe that our system outperforms existing real-time methods.

We quantitatively evaluate the operating speed of our system on TITAN RTX GPU. Due to the lightweight network structure, our network achieves excellent performance. Here we only compute the algorithm time without the human operation. As the used DeepSpeech2 can real-time recognize user's speech, we do not count it for search time. For the once CMPS-Audio-Fast search which uses the word embedding instead of Bert, the search speed is 10 ms per query. For the progressive CMPS-Audio-Image search with Bert, the search speed is 333 ms per query.

4 CONCLUSIONS

We propose an instant and progressive person search system trough cross-modality retrieval. It can be used in the urgent situations where no probe image can be found. Users can directly find the target person with speed description in real-time. The system also supports the progressive search with searched images as query extension. In the future, we will further implement the system on more portable devices. We can also fuse more important information in the query, such as face image on the identification card.

ACKNOWLEDGMENTS

This research was supported by the National Natural Science Foundation of China (Grant No.61525206).

REFERENCES

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, and et al. 2016. Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin. In *ICML*, Vol. 48. 173–182.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *IEEE/CVF ICCV*. 2425–2433.

[3] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan. 2020. Generating Visually Aligned Sound from Videos. *CoRR* abs/2008.00820 (2020).

[4] Tianlang Chen, Chenliang Xu, and Jiebo Luo. 2018. Improving Text-Based Person Search by Spatial Matching and Adaptive Threshold. In *WACV*. 1879–1887.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. 4171–4186.

[6] Chuang Gan, Hang Zhao, Peihao Chen, David D. Cox, and Antonio Torralba. 2019. Self-Supervised Moving Vehicle Tracking With Stereo Sound. In *IEEE/CVF ICCV*. 7052–7061.

[7] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. 2018. Deep Spatial Feature Reconstruction for Partial Person Re-Identification: Alignment-Free Approach. In *IEEE/CVF CVPR*. 7073–7082.

[8] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. 2019. Foreground-Aware Pyramid Reconstruction for Alignment-Free Occluded Person Re-Identification. In *IEEE/CVF ICCV*. 8449–8458.

[9] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[10] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak, and Mubarak Shah. 2018. Human Semantic Parsing for Person Re-Identification. In *IEEE/CVF CVPR*. 1062–1071.

[11] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. 2017. Identity-Aware Textual-Visual Matching with Latent Co-attention. In *IEEE/CVF ICCV*. 1908–1917.

[12] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person Search with Natural Language Description. In *IEEE/CVF CVPR*. 5187–5196.

[13] Wu Liu, Tao Mei, Yongdong Zhang, Jintao Li, and Shipeng Li. 2013. Listen, look, and gotcha: instant video search with mobile phones by layered audio-video indexing. In *ACM MM*. 887–896.

[14] Lei Qi, Jing Huo, Lei Wang, Yinghuan Shi, and Yang Gao. 2018. MaskReID: A Mask Based Deep Ranking Neural Network for Person Re-identification. (2018).

[15] Weijian Ruan, Wu Liu, Qian Bao, Jun Chen, Yuhao Cheng, and Tao Mei. 2019. POINet: Pose-Guided Ovonic Insight Network for Multi-Person Pose Tracking. In *ACM MM*. 284–292.

[16] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *IEEE/CVF CVPR*. 4510–4520.

[17] M. Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. 2018. A Pose-Sensitive Embedding for Person Re-Identification With Expanded Cross Neighborhood Re-Ranking. In *IEEE/CVF CVPR*. 420–429.

[18] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. 2018. Mask-Guided Contrastive Attention Model for Person Re-Identification. In *IEEE/CVF CVPR*. 1179–1188.

[19] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. 2018. Part-Aligned Bilinear Representations for Person Re-identification. In *ECCV*. 418–437.

[20] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In *ECCV*. 501–518.

[21] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. In *ACM MM*. 274–282.

[22] Qi Wang, Xinchen Liu, Wu Liu, An-An Liu, Wenyin Liu, and Tao Mei. 2020. MetaSearch: Incremental Product Search via Deep Meta-Learning. *IEEE Trans. Image Process.* 29 (2020), 7549–7564.

[23] Yuyu Wang, Chunjuan Bo, Dong Wang, Shuang Wang, Yunwei Qi, and Huchuan Lu. 2019. Language Person Search with Mutually Connected Classification Loss. In *IEEE ICASSP*. 2057–2061.

[24] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. 2017. Scene Graph Generation by Iterative Message Passing. In *IEEE/CVF CVPR*. 3097–3106.