Tracking by Detection of Multiple Faces using SSD and CNN Features

Do Nhu Tai*, Soo-Hyung Kim*, Guee-Sang Lee*, Hyung-Jeong Yang*,
In-Seop Na**, A-Ran Oh*

Abstract

Multi-tracking of general objects and specific faces is an important topic in the field of computer vision applicable to many branches of industry such as biometrics, security, etc. The rapid development of deep neural networks has resulted in a dramatic improvement in face recognition and object detection problems, which helps improve the multiple-face tracking techniques exploiting the tracking-by-detection method. Our proposed method uses face detection trained with a head dataset to resolve the face deformation problem in the tracking process. Further, we use robust face features extracted from the deep face recognition network to match the tracklets with tracking faces using Hungarian matching method. We achieved promising results regarding the usage of deep face features and head detection in a face tracking benchmark.

■ keywords : Image Processing; Human face tracking; Active Appearance Model

I. INTRODUCTION

Visual multiple object tracking (VMOT) in general, and multiple face tracking (MFT) in particular, are the fundamental challenges in the field of computer vision such as face recognition system, face expression, and eye-gaze tracking, etc. It traces face positions to support for the specific application to process face information. However, up to now, the problem meet the challenges in the wild such as different light conditions, motion blur, view changes, background blending, and occlusion.

Single face tracking method often uses face observation model such as the stable skin locus model with illumination [1], face shape and textures [2] for keeping track of the face. As the single face tracking, multiple face tracking shares common challenges with face observation model in appearance, motion, and other elements. However, since it traces many faces at the same time, it has some unique challenges in tracking process such as identification for the specific face among many ones, occlusion by the crowd, initialization for new faces and termination for disappearing objects [3]. Tracking-by-detection [4] has emerged as an important tracking method to solve these challenges due to the development of object detection [5, 6]. In the first step, the object detectors will provide potential object locations in frame-by-frame, as well as represent the observation model of object candidates. The second step has the role to link deformable observation model to correspondent targets.

Due to the success of deep neural networks in image classification, object detection, face recognition and so on, the state-of-the-art methods approach to deep learning to solve the tracking problems in three main issues: prediction, feature extraction, data association [7]. These approaches [8, 9] use VGG-Net [10] pre-trained on ImageNet as the feature extraction for observation model, and data association using traditional methods such as correlation filter, the error minimization between deformable object features in the successive frames. MDNet [11] has analyzed the layers in the VGG-Net for features that are sensitive to placement and classification. From there, it builds online binary classification to update the change of object.

* Member, Department of Electronics and Computer Engineering, Chonnam National University
** Member, Software Convergence Education Institute, Chosun University

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2017R1A4A1015559, NRF-2018R1D1A3A03000947)

Manuscript : 2018. 09. 14.
Revised : 2018. 09. 14.
Confirmation of Publication:2018. 10. 12.
Corresponding Author : Soo-Hyung Kim,
e-mail : shkim@jnu.ac.kr

In this paper, we focus on multiple face tracking and solve two issues. In the first issue, we integrated face detection and deformable face model by training face detection network based on head dataset for tracing face deformation. The head dataset [12] contains the context in the wild such as school, class, daily activity, and so on. The second issue on data association solves by robust face feature extraction based on deep face recognition system. After that, we use the deep face features and the Hungarian matching method [13] to associate object candidates in the current frame to the previous frame.

The main contribution of the paper is three parts. Firstly, we trained face detection using Single Shot Multi-Box Detection (SSD) [5] on head dataset for face detection and deformable trace. Secondly, we apply deep face features in the deep face recognition system VGG-Face [14] to extract robust face features. From there, we use Hungary matching method to link face candidates at current frame to all previous frames. Finally, we experimented and evaluated the method with the good results.

In this paper, we organized the rest consisting of four parts. In the second part, we describe face detection using SSD as well as the head dataset. Then, the third part shows face feature extraction and data association to link faces between current frame and previous frames. Next, the four part shows the experiment and the results. Finally, the conclusions are outlined in the fifth part.

II. PROPOSED FACE DETECTION ALGORITHM

The first problem in this paper is to solve that at current processing frame, it is necessary to determine the face candidate locations. Moreover, the method traces the changes of the specific face during its lifetime. Therefore, we decide to use object detector based on deep learning and train on the head dataset suitable with changes over time.

Common object detection architectures based on convolutional neural network consists of two part. The first part is the pre-trained network in image classification for feature extraction. Next, the second part is the meta-architecture with feature extractor and classifier module for outputting object classes and localizations as Fig.1.

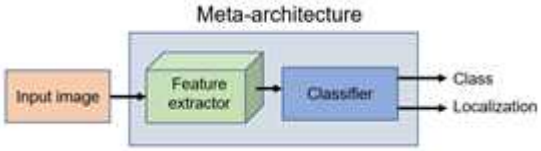

Fig.1. Object detection architecture

We chose SSD [5] due to the advantage of being one of the strongest algorithms for object detection (as depicted in Fig.2) next to Faster R-CNN [15], Yolo3 [6]. SSD directly identifies layers and anchor positions without the need to perform the classification of proposed regions as in Faster R-CNN. In addition, we chose the SSD-300 to meet the accuracy requirements, as well as the success rate of object detection reasonably in practice.

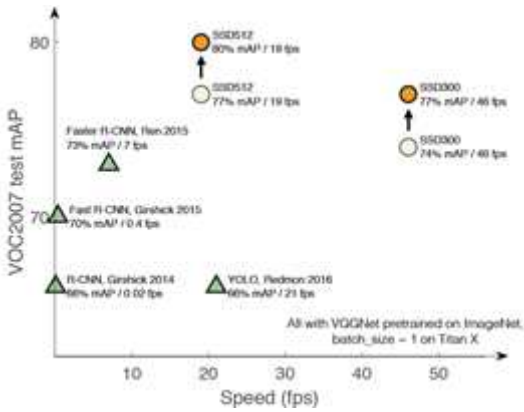

Fig.2. Speed and accuracy between object detection algorithms

In particular, the network architecture of the SSD is firstly to derive Conv5_3 layer from VGG-16 [10] for extracting object features from the specific subclasses. Next, the network adds a few layers to the end of the base network to predict the movements of the original selections with different sizes and scales as well as their reliability. Accuracy in object location detection is due to predicting initial object selections derived from layered features with different scales and sizes comparing with corresponding ground-truth positions. Furthermore, SSD has the advantage in end-to-end solution in training as well as the ability to switch the speed and accuracy with many options such as SSD-7, SSD-300, and SSD-500.

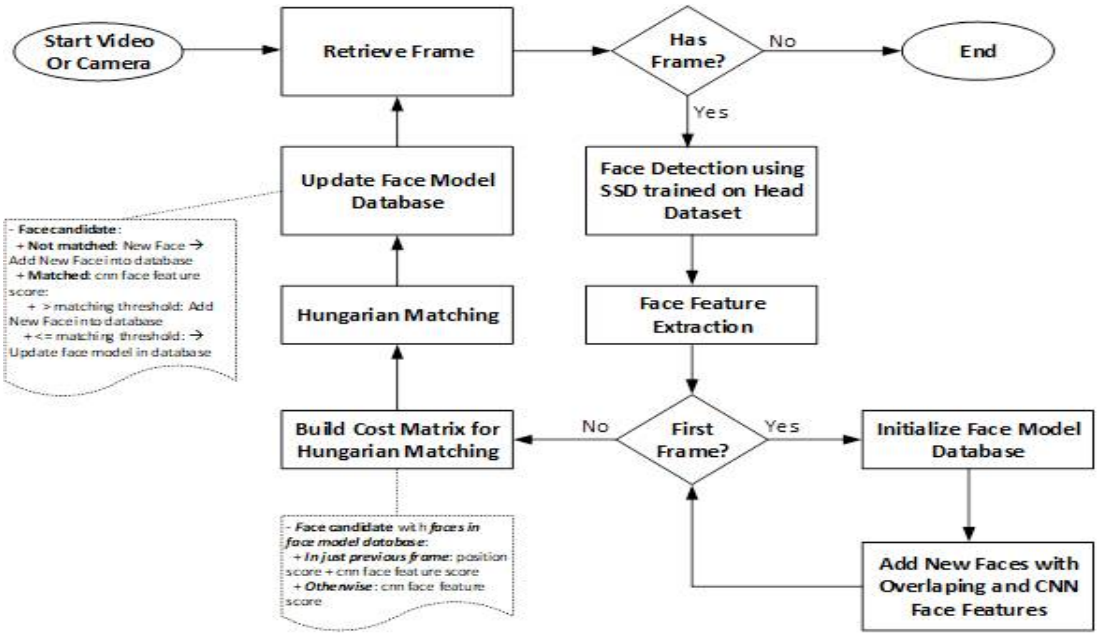

Fig.3. Our proposed tracking method

The head dataset as Fig.4 in the wild will help for face detector SSD to trace the face changes over time during the tracking process. It contains many context for daily activates. Moreover, it contains many faces with many different poses, occlusion, and crowd.

Fig.4. Head dataset in the wild in SCUT-HEAD Dataset A [12]

III. DATA ASSOCIATION

In data association, we use VGG-Face in DeepFace Recognition [14] as Fig. 5 for extracting the robust face features. After that, we proposed our method as Fig.3 for associating between the current face candidates with previous faces.

1. Face feature extraction

Next step in tracking algorithm also needs to propose the robust observation model for faces to link the corresponding faces in data association. Up to now, face features from deep face recognition show the success story in matching faces in face recognition task. In this paper, we choose the VGG-Face in DeepFace [14] from Oxford Research Group to extract robust face feature.

Fig.5. Base network architecture based on VGG-16 for face feature extraction [14]

The based-line network architecture as Fig.5 consists of four main blocks with feature map size 128, 256, 512, 512. Each block consists of convolution, max-pooling layers. The network trained on the face dataset includes 2622 identifiers with over two million face images. The network has two options of soft-max loss and embedding loss to produce feature size 4096 and 1024 respectively with corresponding accuracies 97.2% and 99.13%.

In this paper, we use the deep face feature extracted from the VGG-Face base-line network of DeepFace with the pre-trained weights of the network, which is bypassing the final fully connected layer as well as adding the global average pooling. It will output a 512-dimensional vector of the face model.

2. Data Association

Our proposed data association in tracking algorithm based on learning-by-detection use deep face feature and Hungarian matching described as Fig. 3.

Firstly, the tracking algorithm uses SSD to detect the face candidates at every frame. This detector is robust with the changes in the candidates due to training on the head dataset. Next, the deep face extractor based on VGG-Face will extract the face feature.

In first frame, the algorithm will initialize the face model database and add position information and CNN face features of the current faces detected by SSD. From second frame, the algorithm will build the cost matrix between current faces and the faces in face model database. The weight of face candidate and the faces in just previous frame will sum of overlapping score and CNN face feature score using Euclidean distance metric. Otherwise, the weight only uses CNN face feature score.

Next, the data association will use Hungarian matching algorithm to associate between the face candidates and the faces in face model database. The face candidates, which do not match with any faces model, are new faces to add into the database.

IV. EXPERIMENTAL RESULTS

Our program is developed in Python and Jupiter Lab environment. The computer's configuration is Core i5-3470, 16GB RAM, Windows 10, and graphics card GTX 1080 8GB RAM.

1. Face Detection

We performed experiments on SCUT_HEAD Dataset A, B [12] including two dataset sets. The SCUT_HEAD Dataset A with 2000 images and 67,321 people is taken in the context of a university classroom as Fig.6. There are 1,500 images for training and 500 images for testing. The SCUT_HEAD Dataset B with 2405 images and 43,940 people is in the context in the with from the Internet as Fig.4. In particular, the SCUT_HEAD Dataset B has 1905 images for training and 500 images for testing.

Fig.6. Head dataset in the wild in SCUT-HEAD Dataset B [12]

The distribution histogram shows the number of heads per images in two datasets as Fig.7 and Fig.8 below:

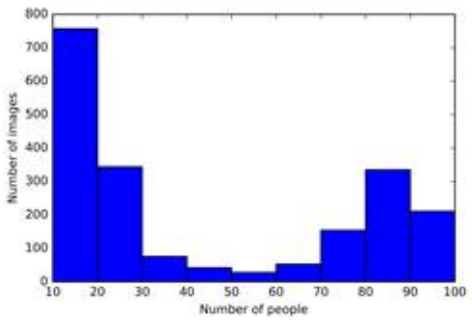

Fig. 7. Histogram of people per image in SCUT-HEAD Dataset A

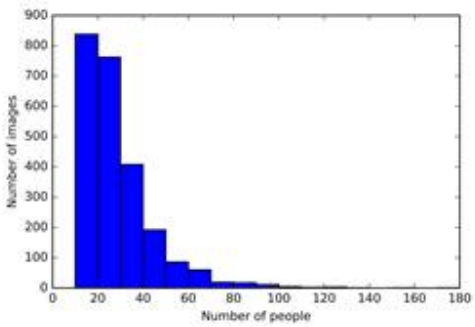

Fig. 8. Histogram of people per image in SCUT-HEAD Dataset B

We train the SSD network with SSD-300 due to the harmony between the runtime and accuracy of the head detection.

Training diagram after 128 epochs with an error of 2.8274 in loss as Fig.9.

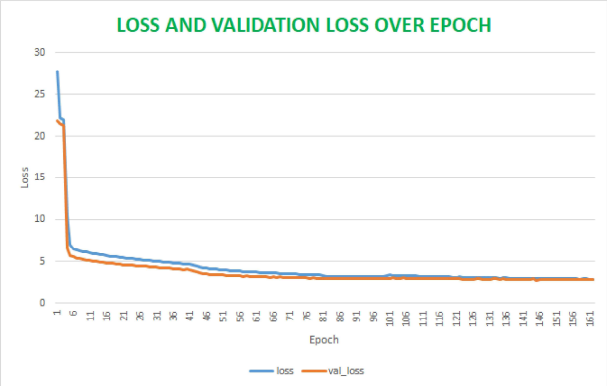

Fig. 9. Train and validation loss over epoch

We conducted to test on the validation data of dataset A and B. The mAP results are 0.513 on the SCUT_HEAD Dataset A and 0.817 on the SCUT_HEAD Dataset B. The low result on dataset A is due to a large number of head people with small size. The high result in dataset B shows the good performance for the context in the wild.

The result images from face detector is shown in Fig.10 and Fig.11.

Fig.10. The result image in SCUT-HEAD Testing Dataset A

Fig.11. The result image in SCUT-HEAD Testing Dataset B

2. Experiment 1: Face Detection

In experiment 1, we use the single face tracking videos in [16] to evaluate the performance face detection methods. With these data, we show the performance of the face detections with face changes during tracking process.

We use Intersection over Union (IoU) metric for the corresponding bounding boxes on every frame. The IoU is metric ratio between the intersection and the union of the predicted boxes and the ground truth boxes as Fig.12.

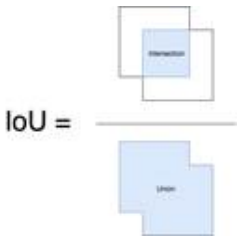

Fig.12. Intersection over Union (IoU) metric

The IoU on a video is average of IoU on every frame with best matching between ground-truth and prediction bounding boxes. Table 1 shows the performance of the face detection in paper (SSD Head) compared with MTCNN [17] and Tiny Face [18] in IoU:

Table 1. Comparison between SSD Head and face detecion methods for Intersection over Union metric

No.	Input Name	MTCNN	Tiny Face	SSD Head
1	Freeman1	0.38	0.62	0.6
2	Trellis	0.59	0.68	0.55
3	Girl	0.37	0.62	0.73
4	Man	0.32	0.65	0.81
5	Boy	0.4	0.61	0.76
6	FaceOcc1	0.83	0.5	0.62
7	FleetFace	0.67	0.5	0.79
8	David2	0.26	0.76	0.39
9	Mhyang	0.51	0.84	0.55
10	DragonBaby	0.55	0.5	0.68
11	Dudek	0.75	0.65	0.8
12	FaceOcc2	0.56	0.75	0.47
13	BlurFace	0.49	0.68	0.38
14	Jumping	0.32	0.65	0.49
	Average	0.55	0.67	0.61

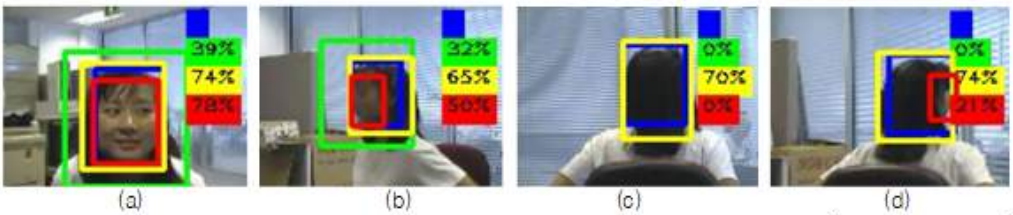

Fig.13. The face in the image changes with many poses (a) frontal face (b) 90^0 rotation (c) 180^0 rotation (d) 270^0 rotation. The colors of the bounding boxes show the face detection method as blue(ground-truth), red (tiny face), green (MTCNN), yellow (SSD Head)

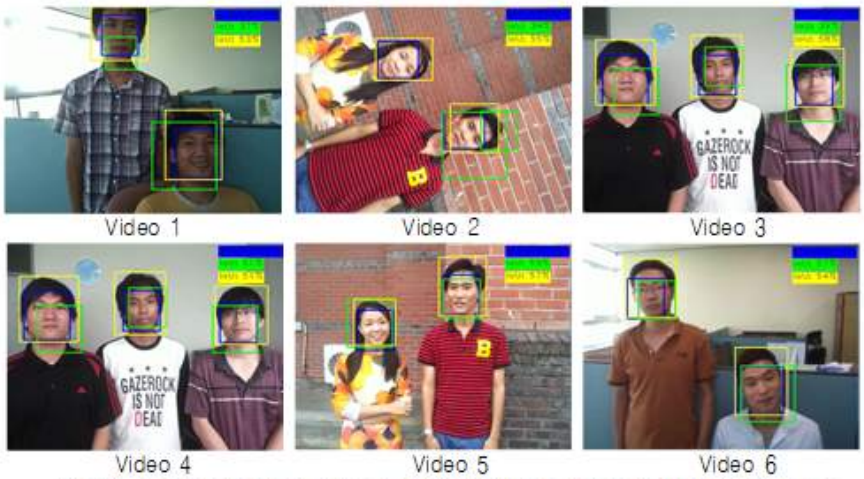

Fig.14. The benchmark videos in [19] used to experiment tracking methods

The face detection method use SSD [5] trained on SCUT-HEAD Dataset to show average IoU in the benchmark videos 0.61, better than MTCNN. Besides, SSD Head is better than other methods when the face is obscured by rotation as in Fig. 13.

Fig.13 shows the advantages of SSD Head when the face is rotation 1800. It also keeps tracking face with the big change of face. Besides, SSD Head also is better than Tiny Face and MTCNN in processing time with average frame per seconds 45.9 as Table 2.

Table 2. Comparison between SSD Head and face detecion methods for processing time (frame per second)

No.	Input Name	MTCNN	Tiny Face	SSD Head
1	Freeman1	39.3	8.8	52.8
2	Trellis	32.8	9.0	32.0
3	Girl	71.2	17.9	37.3
4	Man	71.2	17.9	50.1
5	Boy	41.1	1.5	44.2
6	FaceOcc1	30.4	7.4	46.3
7	FleetFace	16.1	3.4	43.8
8	David2	38.6	9.5	52.4
9	Mhyang	41.2	9.7	53.6
10	DragonBaby	15.6	4.5	38.4
11	Dudek	14.8	3.5	44.2
12	FaceOcc2	41.3	9.8	48.6
13	BlurFace	15.7	3.9	39.7
14	Jumping	32.0	7.7	49.1
	Average	33.8	8.3	45.9

3. Experiment 2: Multiple Face Tracking

We uses 6 videos in dataset [19] for multiple face tracking benchmark as Fig. 14 and Table 3. The videos are selected with two, three people with rotation and translation movement. Besides, we use TLD using Tracking-learning-detection (TLD) [4] to compare with our proposed method.

Table 3. The videos in dataset [19]

No.	Name	Frames	Original Names
1	Video 1	478	multiple_camera1.mp4
2	Video 2	474	multiple_camera3.mp4
3	Video 3	486	multiple_head2.mp4
4	Video 4	488	multiple_head3.mp4
5	Video 5	470	multiple_head8.mp4
6	Video 6	448	multiple_head9.mp4

We also use IoU metric matching for every face in every frame. After that, we calculate average IoU for all frames in the video. We use TLD method [4] with the implementation from OpenCV. The result is shown in Table 4 with mean IoU 0.55 better than TLD method.

Table 4. Comparison between our proposed method and TLD method for for Intersection over Union metric

No.	Input Name	OpenCV TLD	Our Method
1	Video 1	0.49	0.57
2	Video 2	0.59	0.54
3	Video 3	0.43	0.55
4	Video 4	0.50	0.57
5	Video 5	0.55	0.58
6	Video 6	0.36	0.49
	Average	0.49	0.55

Besides, our method takes the real time performance with average 24.39 frame per second as Table 5.

Table 5. Comparison between our proposed method and TLD method for for processing time (frame per second)

No.	Input Name	OpenCV TLD	Our Proposed
1	Video 1	8.66	25.76
2	Video 2	10.99	24.49
3	Video 3	5.92	20.05
4	Video 4	10.08	25.99
5	Video 5	11.21	24.50
6	Video 6	10.25	25.68
	Average	9.49	24.39

Ⅴ. CONCLUSION

The paper has applied the deep face feature extracted from deep face recognition as a metric for matching the faces in current frame with previous frame. The data association uses Hungarian method with deep face metric for linking the faces. The change of face is improved by SSD object detector trained on the head dataset. We archive the good results in the experiments.

REFERENCES

[1] H.N. Bui, 김수형, 나인섭, "Illumination Invariant Face Tracking on Smart Phones using Skin Locus based CAMSHIFT," *스마트미디어저널*, 제2권, 제4호, 9-19쪽, 2013년 12월

[2] 트란 홍타이, 김수형, 김영철, 나인섭, "Human Face Tracking and Modeling using Active Appearance Model with Motion Estimation," *스마트미디어저널*, 제6권, 제3호, 49-56쪽, 2017년 9월

[3] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, X. Zhao, T. Kim, "Multiple Object Tracking: A Literature Review," *Computing Research Repository (CoRR)*, vol. abs/1409.7, pp. 1 - 18, 2017.

[4] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 7, pp. 1409 - 1422, 2012.

[5] W.Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu and A. Berg, "SSD: Single shot multibox detector," *Proc. of Computer Vision - 14th European Conference (ECCV)*, pp. 21 - 37, 2016.

[6] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *Computing Research Repository (CoRR)*, vol. abs/1804.0, 2018.

[7] S. Krebs, B. Duraisamy, and F. Flohr, "A survey on leveraging deep neural networks for object tracking," *Proc. of IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 411 - 418, 2017.

[8] C. Ma, J. Bin Huang, X. Yang, and M. H. Yang, "Hierarchical convolutional features for visual tracking," *Proc. of International Conference on*

Computer Vision (ICCV), pp. 3074 - 3082, 2015
[9] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," *Proc. of International Conference on Computer Vision (ICCV)*, pp. 3119 - 3127, 2015.
[10] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Proc. of International Conference on Learning Representations (ICLR)*, pp. 1 - 14, 2015.
[11] H. Nam and B. Han, "Learning Multi-Domain Convolutional Neural Networks for Visual Tracking," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4293 - 4302, 2016.
[12] D. Peng, Z. Sun, Z. Chen, Z. Cai, L. Xie, and L. Jin, "Detecting Heads using Feature Refine Net and Cascaded Multi-scale Architecture," *Computing Research Repository (CoRR)*, abs/1803.09256, 2018.
[13] H. W. Kuhn, "The Hungarian method for the assignment problem," *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*, Springer, pp. 29-47, 2010.
[14] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," *Proc. of the British Machine Vision Conference (BMVC)*, 2015.
[15] R. Girshick, "Fast R-CNN," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1440 - 1448, 2015.
[16] Y. Wu, J. Lim, and M. Yang. "Online object tracking: A benchmark," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2411-2418, 2013.
[17] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol.23, no.10, pp. 1499-1503, 2016.
[18] P. Hu and D. Ramanan, "Finding tiny faces," *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1522-1530, 2017.
[19] L.N. Do, H.J. Yang, S.H. Kim, G.S. Lee, I.S. Na, and S.H. Kim, "Construction of a Video Dataset for Face Tracking Benchmarking Using a Ground Truth Generation Tool," *International Journal of Contents*, vol.10, no.1, pp. 1-11, 2014.

저 자 소 개

Nhu-Tai Do

He received the B.S. in Information System major from HCM City University of Foreign Language - Information Technology, Vietnam in 2005 and his M.S. in Information System Management from International University, Vietnam National University at HCMC, Viet Nam in 2017. From 2005 to 2017, he was a lecturer in Faculty of Information Technology, HCM University of Foreign Languages and Information Technology, Vietnam. Since 2017, he is Ph.D candidate in the Department of Computer Science, Chonnam National University, Korea. His research interests are pattern recognition, deep learning, computer vision, and parallel programming.

Soo-Hyung Kim

He received his B.S. degree in Computer Engineering from Seoul National University in 1986, and his M.S. and Ph.D. degrees in Computer Science from Korea Advanced Institute of Science and Technology in 1988 and 1993, respectively. From 1990 to 1996, he was a senior member of research staff in Multimedia Research Center of Samsung Electronics Co., Korea. Since 1997, he has been a professor in the Department of Computer Science, Chonnam National University, Korea. His research interests are pattern recognition, document image processing, medical image processing, artificial intelligence, and deep learning..

Guee-Sang Lee

He received the B.S. degree in Electrical Engineering and the M.S. degree in Computer Engineering from Seoul National University, Korea in 1980 and 1982, respectively. He received the Ph.D. degree in Computer Science from Pennsylvania State University in 1991. He is currently a professor of the Department of Electronics and Computer Engineering in Chonnam National University, Korea. His research interests are mainly in the field of image processing, computer vision and video technology.

Hyung-Jeong Yang

She received her B.S., M.S. and Ph.D. from Chonbuk National University, Korea. She is currently a associate professor at the Department of Electronics and Computer Engineering, Chonnam National University, Gwangju, Korea. Her main research interests include multimedia data mining, pattern recognition, artificial intelligence, e-Learning, and e-Design.

In-Seop Na

He received his B.S., M.S. and Ph.D. degrees in Computer Science from Chonnam National University, Korea in 1997, 1999 and 2008, respectively. From 2012 to 2018, he was a research professor in Chonnam National University, Korea. Since 2018, he has been a visiting scholar in University of California (Merced), USA and an assistant professor in Software Convergence Education Institute, Chosun University, Korea. His research interests are image processing, pattern recognition, object detection, segmentation, recognition, and emotion recognition

A-Ran Oh

She received her B.S. degree in school of Computer Statistics from Chosun University, Korea in 2009, She is currently a research engineer at the Department of Computer Science, Chonnam National University, Korea.