Human Face Tracking and Modeling using Active Appearance Model with Motion Estimation

Hong Tai Tran, In Seop Na, Young Chul Kim, Soo Hyung Kim*

Abstract

Images and Videos that include the human face contain a lot of information. Therefore, accurately extracting human face is a very important issue in the field of computer vision. However, in real life, human faces have various shapes and textures. To adapt to these variations, A model-based approach is one of the best ways in which unknown data can be represented by the model in which it is built. However, the model-based approach has its weaknesses when the motion between two frames is big, it can be either a sudden change of pose or moving with fast speed. In this paper, we propose an enhanced human face-tracking model. This approach included human face detection and motion estimation using Cascaded Convolutional Neural Networks, and continuous human face tracking and modeling correction steps using the Active Appearance Model. A proposed system detects human face in the first input frame and initializes the models. On later frames, Cascaded CNN face detection is used to estimate the target motion such as location or pose before applying the old model and fit new target.

Keywords: Image Processing|Human face tracking|Active Appearance Model

I. INTRODUCTION

With the recent development of science, and technology, extracting information from human face is an important necessity for many applications such as face security system, facial expression recognition, automatic improving a taken human photo from camera, or automatic robot interaction with human, etc. Therefore, a system to accurately extract the human face information automatically from video is a very urgent requirement for many applications[12]. Following that necessity, there are many approaches proposed with various techniques and their combinations for human face tracking[1-3,12]. Among those approaches, we decide to use the Active Appearance Model (AAM)[1] for the face tracking task. Because the AAM is not only suitable for adapting to the change of a human face due to expression or illumination – but the landmarks from AAM can also be used as features for further development or other processing. Our system consist of AAM as the based method for tracking and modeling. However as model based method has weakness to fast motion and need a good initial location, we also use the cascaded CNN(Convolutional Neural Networks) face detector to solve these problem.

In this paper, we will describe an approach to human face tracking using the AAM through the following sections. Section 2 will briefly describe the background researches, which used in our paper. After that, Section 3 will describe our proposed method and some experimental result can be seen in Section 4.

* This work was partly supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No.2017-0-00383, 스마트 회의실: 빅 스크린을 활용한 지능형 회의 솔루션 개발). This research was also supported by Ministry of Regional Innovation, creative business training Research Foundation of Korea, 2014 (NRF-2014H1C1A1066771).

*Member, Department of Electronics and Computer Engineering, Chonnam National University

Manuscript : 2017 07 17
Revised : 2017.08.29
Confirmation of Publication: 2017.09.07

Corresponding Author : Soo Hyung Kim, e-mail : shkim@chonnam.ac.kr

II. BACKGROUND

1. *Active Appearance Model*

Active Appearance Model (AAM), which proposed by Cootes [1,8,12], is a very popular method in the field Computer Vision. The advantages of AAM is that the models can define the varying shape and appearance of an object based on trained model from a representative training set. AMM is a combination of shape and texture models is a combination of shape and texture (sometimes called appearance) models[12]. For the shape model, we require a fixed set of landmarks over the sets of training images which will define the shape of the learning object. Then, the highly similar shapes will be removed using Procrustes Analysis. Finally, we apply PCA to obtain the final shape model which defined by a mean shape (s_o) and the eigen-vectors set (S). When fitting to a new sample, the estimate model $\hat{s}$ can be calculate as:

$$\hat{s} = s_o + Sp$$

where S_p is the estimated shape weights/parameters to fit $\hat{s}$ to the sample.

Similarly, the texture model can be learned by taking the texture from marked regions of learning objects and warp them into the mean shape. Then, we apply PCA to obtain the texture model t. As mention above, in this model, the shape is all warp into the mean shape so while the color changing the shape is remaining the same when changing texture model parameter. When apply to a new target, the model can be estimate as:

$$\hat{t} = t_0 + Tq$$

where T_q is the estimated texture weights/parameters to fit t_0 to the sample.

After that, when input a new target, the model will be fit onto the target by minimizing the difference between the target texture and shape with the model follow either fitting them separately or after combined them into one model. Both methods have their own strong and weak points [2].

2. *Cascaded CNN Face Detector*

Cascade face detector was first proposed by Viola [5], using Haar-like features to train cascaded classifiers which can do the task of face detection in real-time with notable efficiency in frontal face detection. However, the Viola's cascades is tended to fail in cases of unexpected lightning or faces poses[12]. We will use K.Zhang's cascaded CNN[6] which follow the approach of Viola cascade classifiers but the cascade is trained using the deep learning – CNN network.

In our experimental work we have tested based on method which are included 3 stages, each stage uses a cascaded built using a full CNN:

Stage 1: Using the Proposal Network(P-Net) which obtains every facial window candidate. The candidates are going through bounding box regression and non-maximum suppression (NMS) to merge the highly overlapped candidates.

Stage 2: After merging the all the candidate, then going into use the Refine Network (R-Net), it rejects the face false candidates and leads the true candidates to another NMS and bounding box regression.

Stage 3: The detected face region from the second stage is fed to the O-Net and also is going through NMS and bounding box regression to obtain the five facial landmarks.

Fig. 1. Cascaded CNN example output

III. PROPOSED SYSTEM

On the first frame of input video or direct camera frames, the system detects the human face and build the initial model. Then, to detect the target new location and estimate its roll pose, a small region search will be applied for every new frame. In case of without face finding, the proposed system let the model fit to do the tracking job itself. Util the video/camera terminate, the process is work repeatedly. Fig. 2. shows our system block diagram and Fig. 3. illustrates the system process for face tracking.

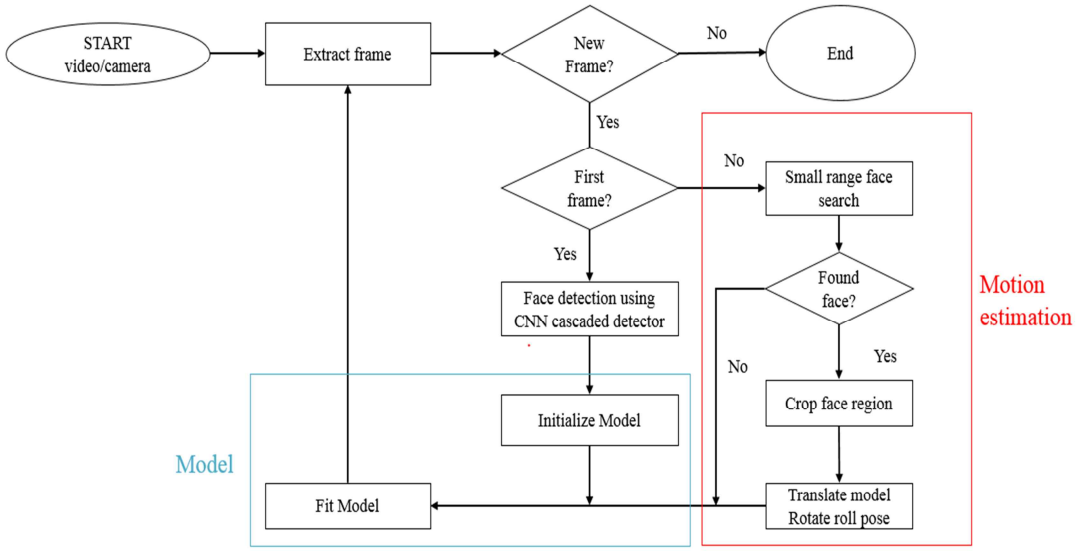

Fig. 2. System block diagram

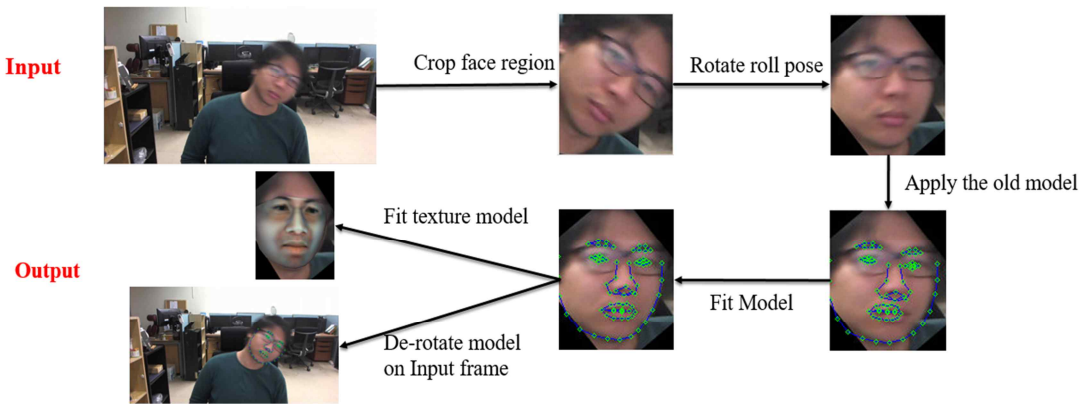

Fig. 3. Tracking process illustration

1. Active Appearance Model

As mentioned above, our system uses AAM as bas e method for modeling and tracking. To build the m odel, we used the MUCT(Milborrow / University of Cape Town) face database which consist of 3755 fa ces with 76 landmarks. Some exa-mples of MUCT database can be seen in Fig. 4.

Fig. 4. MUCT face database examples

For fitting the AAM, we use the separated scheme where the shape model is fitted first for a target face followed by the texture model on the shape region This approach helps us stabilize the tracking result compared to the

combined approach where both model is fitting at the same time.

2. Motion Estimation

In this section, we describe the process of motion estimate to improve the tracking power of the AAM (Fig. 5). First, since model based method is weak to large location motion, we apply a small range search using CNN cascaded detector within the window of

$$\left(\left[x_0-\frac{1}{2}w_0, x_0+\frac{3}{2}w_0\right], \left[y_0-\frac{1}{2}h_0, y_0+\frac{3}{2}h_0\right]\right),$$

where x_0, y_0, w_0, h_0 are the coordinates, width, and height of the target respectively.

Second, we want to estimate the pose of the target since sometimes the target changing its pose rapidly also cause the model to be failed. In this case, using the output of face features from small range search, we estimate the face roll pose and straighten it before applying the model.

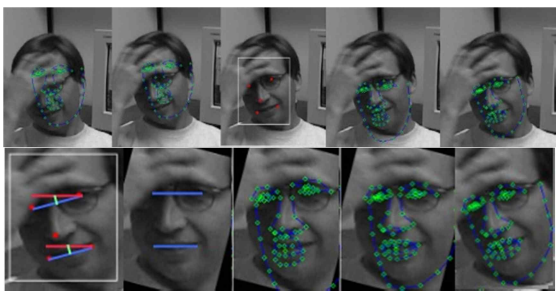

Fig. 5. Location (upper) and roll pose (lower) estimation.

IV. EXPERIMENTAL RESULTS

Our system has been implemented in Python 2.7, on a system configured with an Intel Core i5-3470 CPU, NVIDIA GeForce GTX 660 GPU, 8 GB RAM, and Ubuntu 16.04. Each page from the PDF files of the dataset was converted into a JPG image before applying our system. The dataset used for experiments and evaluation is sequences with human face ground truth in TB-100 Sequences[11]. We tested all samples with face in TB-100. The example result is shown in Fig. 6, Fig. 7, and Fig. 8.

To proof the power of our tracking method, we also compare with tracking method from [10] with the error rate calculated by different between output and ground truth center normalize by size of target from ground truth. The result is shown in Table 1.

The error rate is calculated by normalizing the Euclidian distance between the center of ground truth and output by the size of target from the ground truth:

$$e = d_E(c_0, c)/p_0$$

where:

e is the error

$d_E(c_o, c)$ is the Euclidian distance between c_0 and c

c_0 and c are centers of the ground truth window and output window respectively

p_0 is the size of the ground truth window

Table 1. Comparison between our method and L1_APG

No.	Input Name	Error	
		Our method	L1_APG
1	David2	0.0202	0.0017
2	BlurFace	0.0012	0.0101
3	Boy	0.0321	0.0434
4	Dragon Baby	0.0105	0.022
5	Dudek	0.0015	0.0011
6	FaceOcc1	0.0007	0.0008
7	FaceOcc2	0.0016	0.0021
8	FleetFace	0.0021	0.0034
9	Freeman1	0.0092	0.0221
10	Girl	0.0149	0.0061
11	Jumping	0.0385	0.0396
12	Man	0.0046	0.0014
13	Mhyang	0.0014	0.0015
14	Trellis	0.0026	0.0082
	Average	0.0100	0.0107

The result of evaluation (Table 4.1) shows that our face tracking accuracy is higher than L1_APG[10] which is real time robust L1 tracker using accelerated proximal gradient approach in most of comparison sequences. Moreover, while we initialize our system location using the face detection in the system, the L1_APG was initialize using the ground truth on the first frame.

V. CONCLUSION

In this paper, we proposed a system using a combination of AAM fitting and cascaded CNN detector to track and model the human face. To avoid the weaknesses from most model-based tracking

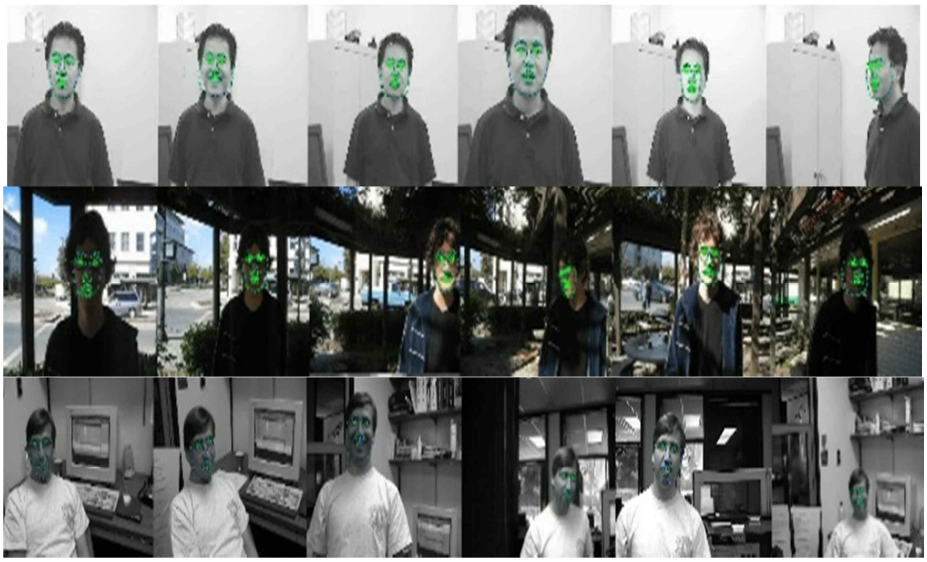

Fig. 6. Sample output from Mhyang, Trellis, and Dudek.

Fig. 7. Texture Sample from Mhyang, Trellis, and Dudek.

target which is the initialization and fast motion, the system try to utilize the model from AAM to get more information which come from the tracking target. Although there is still a limit in how much the system can improve for the model-based method, the system should be able to track the human face in most of normal life activity. To improve the model fitting speed and accuracy are still a problem.

Fig. 8. Other example outputs from TB-100seq which has human face ground truth.

REFERENCES

[1] T. Cootes, G. Edwards, and C. Taylor, "Active Appearance Models," Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 484-498, 2001.

[2] I. Matthews, and S.Baker, "Active appearance models revisited," International journal of computer vision, vol. 60, no. 2, pp. 135-164, 2004.

[3] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," IEEE Conference on Computer Vision and

Pattern Recognition, pp. 5325-5334, 2015.
[4] S. Milborrow, J. Morkel, and F. Nicolls, "The MUCT Landmarked Face Database," Pattern Recognition Association of South Africa, 2010.
[5] P. Viola, and M. J. Jones, "Robust real-time face detection," International journal of computer vision, vol. 57, no. 2, pp. 137-154, 2004.
[6] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Neural Networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, 2016.
[7] E.N. Arcoverdo Neto, R.M. Duarte, R.M. Barreto, J.P. Magalhaes, C.M. Bastos, T.I. Ren and G.D.C. Cavalcanti, "Enhanced real-time head pose estimation system for mobile device," Integrated Computer Aided Engineering, vol. 21, no. 3, pp. 281-293, 2014.
[8] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models-their training and application," Computer Vision and Image Understanding, vol. 61, no. 1, pp. 38-59, 1995.
[9] C. Zhang, and Z. Zhang, "Improving multiview face detection with multi-task deep convolutional neural networks," IEEE Winter Conference on Applications of Computer Vision, pp. 1036-1041, 2014.
[10] C. Bao, Y. Wu, H. Ling, and H. Yi, "Real Time Robust L1 Tracker Using Accelerated Proximal Gradient Approach", Conference on Computer Vision and Pattern Recognition, pp. 1830-1837, 2012
[11] http://cvlab.hanyang.ac.kr/tracker_benchmark/datasets.html
[12] H. T. Tran, "Enhanced Model Based Human Face Tracking Method using CNN Cascade Face Detector",Masters Thesis, Chonnam National University, Korea, 2017.

Authors

Hong-Tai Tran

He received his B.S. degree in Mathematics and Computer Science in Ho Chi Minh City University of Science, in 2014. Since 2015, he has been a Master student at Electronics and Computer Engineering, Chonnam National University, South Korea. His research interests are pattern recognition, machine learning, and medical image processing.

In-Seop Na

He received his B.S., M.S. and Ph.D. degree in Computer Science from Chonnam National University, Korea in 1997, 1999 and 2008, respectively. Since 2012, he has been a research professor in Department of Computer Science, Chonnam National University, Korea. His research interests are image processing, pattern recognition, character recognition and digital library.

Young-Chul Kim

He received his PhD from Michigan State University, USA, the MS from the University of Detroit, USA, and BS in electronics engineering from Hanyang University, Korea. In 1993, he joined the Department of Electronics Engineering at Chonnam National University (CNU) where he is currently a professor. From 2000 to 2004, he was a director of IDEC at CNU. From 2004 to 2005, he was a Vice Dean of the College of Engineering in this university. From 2004 to 2014, he was the chief of the LG Innotek R&D center at CNU. His research interests are SoC design and smart interface and Natural User Interface (NUI) and low power design.

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, and ubiquitous computing