SPECIALIZED GAZE ESTIMATION FOR CHILDREN BY CONVOLUTIONAL NEURAL NETWORK AND DOMAIN ADAPTATION

Wen Cui   *Jinshi Cui*   *Hongbin Zha*

Key Laboratory of Machine Perception(MOE), Peking University,
Beijing 100871, China

ABSTRACT

Children's social gaze behavior modeling and evaluation has obtained increasing attentions in various research areas. In psychology research, eye gaze behavior is very important to developmental disorders diagnosis and assessment. In robotics area, gaze interaction between children and robots also draws more and more attention. However, there exists no specific gaze estimator for children in social interaction context. Current approaches usually use models trained with adults' data to estimate children's gaze. Since gaze behaviors and eye appearances of children are different from those of adults, the current approaches, especially those with free-calibration assumptions which are utilized in usual human-robot interaction systems, will result in big errors. Note that children data is difficult to collect and label, so directly learning from children data is hard to achieve. We propose a new system to solve this problem, which combines a CNN feature extractor trained from adult data and a domain adaptation unit using geodesic flow kernel to adapt the source domain (adults) classifier to the target domain (children). Our system performs well in children's gaze estimation.

Index Terms— Gaze Tracking, Neural Network, Semi-Supervised Learning, Computer Vision for Automation

1. INTRODUCTION

Gaze behavior is a powerful social signal [1]. It is an effective and vital way of interaction for infants with mother or other care-givers. Abnormal social gaze pattern often indicates development disorder. Specifically, psychologists found that impaired gaze behavior is one of the diagnostic features of autism [2]. Motivated by psychological studies of children's social gaze in human-human interactions, computational modeling of human gaze behavior draws more and more attention in many areas, including human-robot interaction (HRI), e.g. a good computational gaze model enables a robot to perceive information about the state of the children in the interaction and then adjust its behavior accordingly [3]. In this background, appropriate gaze behavior is critical for establishing natural human-robot interaction which enables human to engage in social interactions.

Current gaze estimation methods are model-based or appearance-based [4]. Model-based methods need a high-resolution observation of eyes, and appearance-based methods need more person-specific training data. Currently a convolutional neural network [5] is proposed to cope with the low-resolution eye images, cross-person data, and variety of appearance and illumination. It experimentally outperforms previous methods. Motivated by their work, we propose AlexNet [6] structure of convolutional neural network to address the issues on gaze estimation.

When dealing with childrens gaze estimation problem, lack of children data is another vital problem. A large amount of samples and labels are required for a designed learning-based model. However data collection from children is very difficult. Previous methods are mostly trained from adult data and used for adults' and children's estimation, they perform poorly on children cases. To address the problem, we introduce a domain adaptation (DA) method which is rapidly developed.

In our system, we firstly apply facial analysis to obtain the head pose and eye images. A convolutional neural network is then utilized to extract the features of eye images and their corresponding head poses and detect gaze directions. Since we get the model, a domain adaptation process developed from Gong et al. [7] is executed. It is added to the network after fully-connected layer, using adult data as source domain and children data as target domain to improve following detection for child data.

Our contributions are two-fold. Firstly, we train a gaze detecting system which is person-independent, pose-independent and free-calibration. It is used for processing near-frontal and less-occlusion faces, and suitable for most situations in children user human-robot interaction. Secondly, to address the lack of child data, we are the first to use domain adaptation in gaze estimation. And our method can significantly reduce the error rate of the estimation for child data.

2. RELATED WORKS

Many researchers have focused on gaze estimation. Asteriadis et al. use screen system as visual plane to describe gaze [8], it is not suitable for natural social scenes. Noris et al. use remote wearable system [9] to estimate gaze, but it is invasive for children. Mora and Odobez [10] propose a geometric generative model estimating gaze from high-resolution eye images. Sugano et al. [11] propose a appearance-based model and use a large amount of cross-subject training data to train a 3D gaze estimator. And their model also concerns about the influence of head poses. Recently Zhang et al. [5] propose a CNN-based method to train a gaze estimator. A convolutional neural network is experimental confirmed to perform much better in free-calibration, person-independent and pose-independent situations, which are most of the human-robot interaction situations.

Approaches above are trained and tested on datasets composed of images of adults' faces. But we concentrate on children's gaze estimation. The distribution of the features of children are different from that of adults. Thus the model should be correspondingly revised to adopt on children data. Domain adaptation, which is well used in object recognition, aims to build classifiers to overcome the biases between training data and testing data and be robust to mismatched distributions. Gong et al. propose a geodesic flow kernel [7] to reveal the underlying commonness and difference between the domains, which works well in object recognition. Motivated by their work, we propose to utilize geodesic flow kernel to address the difference from adult data to child data.

3. APPROACH

Fig. 1 presents an overview of our method. There are three steps in our model.

3.1. Feature Extraction and Normalization

In this process, matrices characterizing the eye images are calculated from the adult and child data obtained from videos and photographs, and the vectors representing the head poses are computed correspondingly.

Head-Track package of Visage SDK [1] is used to track faces in the videos, and a high-performance, accurate 3D head pose can be obtained for each frame. In the detected face region, eye regions I are detected by using the left and right corners of the eyes, aligned by 3D head pose, and resized into a fixed resolution of 36×60. Aligned images are obtained using head pose θ by bilinear interpolation. The alignment normalizes eye images' rotation to be strictly frontal, it separates the contribution of eye motion and head motion in gaze.

The eye images are then histogram-equalized. And the head poses are transformed into 3D rotation vectors by Rodrigues transformation. Considering the symmetry of the facial features of human beings, the right eye images are mirror reflected from right to left. Correspondingly, their vectors of 3D head pose are mirror changed as well. It can sharply reduce the complexity of the subsequent training.

3.2. Multimodal Convolutional Neural Network

In this process, we learn a regression from input data, including eye images and 3D head pose, to the continuous 2D gaze angles. And the features learned from the training, which determines the effects of regression, are recorded for subsequent domain adaptation process.

Zhang et al. [5] recently propose to use CNN to train a gaze estimator and their model use LeNet network architecture. Motivated by the well-performance of CNN in vision researches, we use AlexNet [6] network architecture, which consists of three convolutional modules (including one convolutional layer, one batch normalization layer and one max-pooling layer), two fully connected layer and a regression layer, as shown in Fig. 1.

Head poses have a big influence on the gaze estimation. Some previous researchers use geometric model to calculate the real gaze direction considering the head pose. Recently, some researchers directly add head pose into their learning-based models as inputs. Mora and Odobez [10] use head pose as a hidden variable combined with other hidden variables characterizing shape of eyes in their generative probabilistic graphic model and Zhang et al. [5] propose to concatenate the head pose vector with output of fully connected layer. Following their idea, in our network, the second fully connected layer combines the output of the first fully connected layer and the head pose vector as inputs and feed its output to the regression layer.

After training, the regression layer is chopped off and the rest of the network is used as a feature extractor of the eye images.

3.3. Domain Adaptation

In this process, adult data and child data are regarded as source domain and target domain respectively, and a semi-supervised domain adaptation process is executed by geodesic flow kernel method.

Domain adaptation is aimed to describe the differences of the distribution. Geodesic flow kernel method focuses on deriving a new feature representation to facilitate domain shift. It embeds source and target datasets in a Grassmann manifold and then constructs a geodesic flow between the two points in the manifold and integrates subspaces in the flow. The raw features from CNN are projected into these subspaces and form a new infinite-dimensional feature vector. More details and proofs about the geodesic flow kernel model can be found in [7].

In our work, child data are insufficient due to the difficulty of collecting and labelling, while adult data are sufficient. Feature vectors are extracted from above CNN. And then features from both domain are collected to form a cross-domain subspace and each feature vector is aligned to this space. Finally the new derived representations are used to construct a 1-nearest-neighborhood classifier to estimate the

[1] http://visagetechnologies.com/

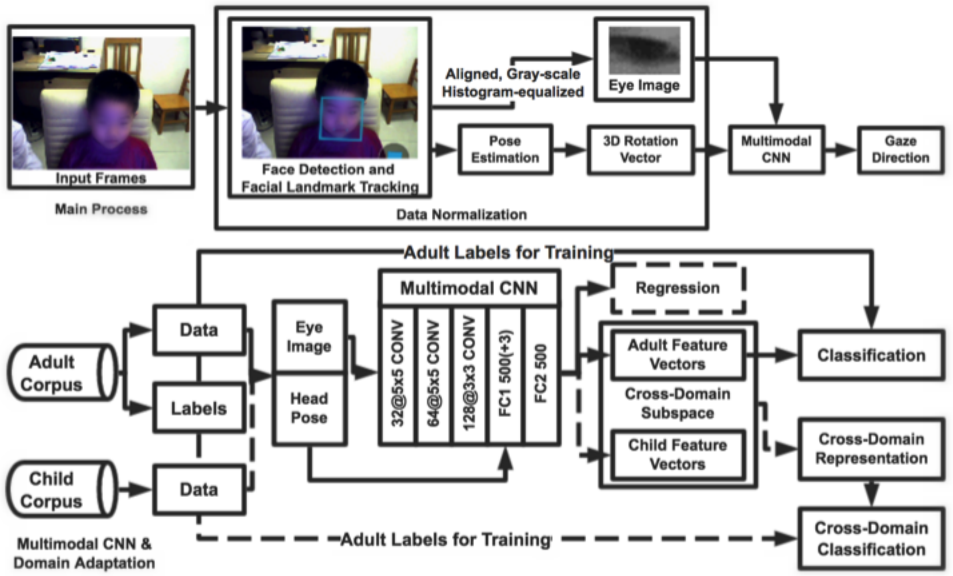

Fig. 1. The main process shows the processing flow of child data, and the flowchart of CNN and DA module gives more details about the classifier trained on adult data and child data.

gaze. For a better performance, a few labels from target domain is chosen for the semi-supervised learning process. To define the labels of gaze, the field of vision is divided into infinite number of cells in horizontal direction and vertical direction, and each cell has the same width (in our experiments, we define 6 degrees per cell).

4. EXPERIMENT

4.1. Data Collection

We use UT Multiview dataset [11] as our adult data to learn the CNN. The dataset contains dense gaze samples of 50 participants as well as 3D reconstructions of eye regions that can be used to synthesize images for arbitrary head poses. We used the synthesized eye images and their correspondingly head poses in the dataset.

As for children data, we collect data in a monitor-screen setting by using web camera to capture and Tobii TX-300 [2] eye tracker to label. We chose 7 children whose ages range from 3 to 5 to obtain data. In the task, we asked them to sit in front of a screen and watch a short video playing in the screen at a distance of 60 centimeters. The web camera was set at the top of the screen to capture each child at a resolution of 640 × 480 pixels and 25 frames per second. The eye tracker was set at the bottom of the screen to track eye gaze at a frequency of 300.

There was an five-point eye calibration before capturing data. The calibration was executed three to five times to make them reliable enough. After calibration, the eye tracker could estimate the coordinates of the gaze points and eye positions in the screen coordinate system. With the assumption that the gaze tracker provided accurate values, the ground-truth gaze direction could be calculated by these coordinates.

Then we mapped the video frames captured from the webcam to the calculated gaze by the timestamps. Because of distractions and tightness of children, data collection was very difficult. Many frames are abandoned for uncertain observations, e.g. bowing down head, shaking violently, and looking outside screen. We finally captured 1860 well-performed and labelled frames and 3720 eye images, and chose these samples to help execute and evaluate our experiments.

With data obtained above, we carried out our experiments in two steps.

4.2. Step I. Training Multimodal Convolutional Neural Network

In this step, the UT Multiview Gaze dataset [11] is used to train and test the CNN. Many previous gaze estimation methods are compared in this dataset in [5], and we mainly compare our method with [5]'s in the same settings. There are two settings. One is within-individual method, where we mixed all individuals' data and randomly selected training samples and testing samples. Thus each individual's data can be utilized in training. The other is cross-individual method, where we randomly left several individuals out as testing set, and other individuals' data were utilized in training.

We use Keras [12] and Theano [13] Backend to train the CNN of ours and [5]'s, the learning rate is 0.001 and the num-

[2] http://www.tobii.com/

Mean Error	Within-Individual Ours	Cross-Individual Ours	Within-Individual [5]	Cross-Individual [5]
Horizontal	**2.4879°**	3.0911°	2.7132°	**2.8293°**
Vertical	**2.5882°**	**3.5797°**	3.2022°	3.8784°
Geometric	**3.9609°**	**5.2876°**	4.3229°	5.5767°

Table 1. CNN regression evaluation in UT Multiview dataset

Methods	Horizontal Mean Error	Vertical Mean Error
Our CNN	11.8182°	24.4106°
CNN+DA	8.5498°	17.5080°

Table 2. Results for Child Data with and without DA

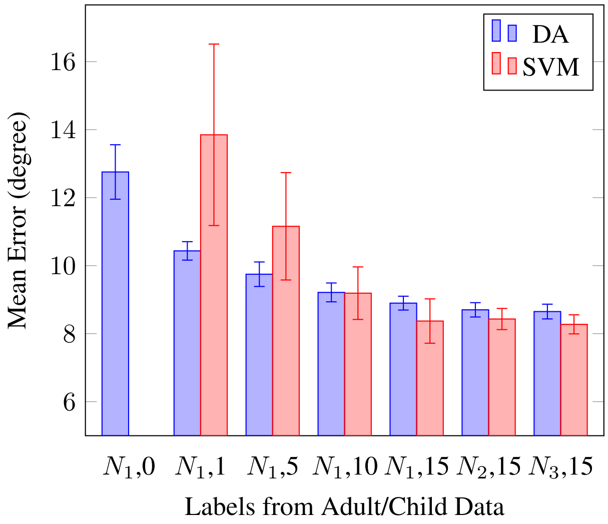

Fig. 2. Results of DA and SVM Methods in Estimating Child Data. The x axis represent the number of labels used in training from adult data and child data. $N_1 = 10^3, N_2 = 1.5 \times 10^3, N_3 = 2 \times 10^3$.

ber of Epoch is 50, Table 1 shows the results of ours and [1]'s. It shows that our network performs well enough as state-of-the-art methods.

After learning and confirming that our network performs well enough, the regression layer is chopped off and the outputs of fully connected layer are extracted to execute the following experiments. The network is used as a feature extractor to help calculate the feature vectors characterizing adult and child data. A 500-dimensional vector for each sample including an eye image and its corresponding head pose can be finally obtained after extracting process.

4.3. Step II. Domain Adaptation from Adult Data to Child Data

We evaluate the domain adaptation in two ways. First, we use the CNN trained from adult data in child data, and compare it with results after DA process. And after that, we evaluate the results using different amount of adult samples and child samples, and compare them with SVM method using the same data. As we train 1-nearest-neighbor classifier to estimate gaze, we separate labels by 6 degree per class. So in horizontal direction, there are 8 classes, and in vertical direction, there are 7 classes.

Table 2 shows that with-DA method certainly reduced the error compared with without-DA method. And the error in vertical is larger than in horizontal, because of features loss of blink or semi-open eyes. In Fig. 2, the params in the first column means the number of samples used from adult(the former) and child(the latter). Analyzing the results we find that, firstly, the more adult samples are used, the better DA performs, secondly, the more child samples are used, the better DA performs, thirdly, DA performs much better than SVM when the samples are very few, and lastly, DA has much smaller standard deviation than SVM method.

5. CONCLUSIONS

We learn from the well-performed works in gaze estimation and domain adaptation, and propose a fine combination to address the special problem in children's gaze estimation. And experimentally, we confirm that our method performs better than previous works. It provides a new methodology in dealing with children's gaze estimation with lack of labelled data. It is an effective model to extend to a wide use in human-robot interactions and psychological diagnosis for children. However there are still some bottlenecks in our work, e.g. the robustness to semi-open eyes which often occur in children data, the limitation of layers in the outmoded CNN structure, it can be deeper, introduced more modules and perform much better. In future works, we will improve our system, introduce the gaze estimator into the wearable camera, and study the gaze behavior in children's social interactions.

6. ACKNOWLEDGEMENT

This work was supported by NSFC Grant (No.61273253) and National High Technology Research and Development Program of China (No.2012AA011602).

7. REFERENCES

[1] Simon Baron-Cohen and Pippa Cross, “Reading the eyes: Evidence for the role of perception in the development of a theory of mind,” *Mind & Language*, vol. 7, no. 1-2, pp. 172–186, 1992.

[2] Simon Baron-Cohen, Ruth Campbell, Annette Karmiloff-Smith, Julia Grant, and Jane Walker, “Are children with autism blind to the mentalistic significance of the eyes?,” *British Journal of Developmental Psychology*, vol. 13, no. 4, pp. 379398, 1995.

[3] Frank Broz, Hagen Lehmann, Yukiko Nakano, and Bilge Mutlu, “Gaze in hri: from modeling to communication,” pp. 491–492, 2012.

[4] F. Lu, Y Sugano, T Okabe, and Y Sato, “Gaze estimation from eye appearance: A head pose-free method via eye image synthesis.,” *Image Processing IEEE Transactions on*, vol. 24, no. 11, pp. 3680, 2015.

[5] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling, “Appearance-based gaze estimation in the wild,” in *Computer Vision and Pattern Recognition*, 2015, pp. 4511–4520.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in *International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[7] Fei Sha, Yuan Shi, Boqing Gong, and K. Grauman, “Geodesic flow kernel for unsupervised domain adaptation,” in *Computer Vision and Pattern Recognition*, 2012, pp. 2066–2073.

[8] Stylianos Asteriadis, Paraskevi Tzouveli, Kostas Karpouzis, and Stefanos Kollias, “Estimation of behavioral user state based on eye gaze and head poseapplication in an e-learning environment,” *Multimedia Tools and Applications*, vol. 41, no. 3, pp. 469–493, 2009.

[9] Basilio Noris, Jean Baptiste Keller, and Aude Billard, “A wearable gaze tracking system for children in unconstrained environments,” *Computer Vision & Image Understanding*, vol. 115, no. 4, pp. 476–486, 2011.

[10] Kenneth Alberto Funes Mora, Florent Monay, and Jean Marc Odobez, “Eyediap: a database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras,” in *Symposium on Eye Tracking Research & Applications*, 2014, pp. 255–258.

[11] Y Sugano, Y Matsushita, and Y Sato, “Learning-by-synthesis for appearance-based 3d gaze estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1821–1828.

[12] François Chollet, “Keras,” `https://github.com/fchollet/keras`, 2015.

[13] Theano Development Team, “Theano: A Python framework for fast computation of mathematical expressions,” *arXiv e-prints*, vol. abs/1605.02688, May 2016.