JOINT ESTIMATION OF HEAD POSE AND VISUAL FOCUS OF ATTENTION

Yingning Huang[*]*, Dingrui Duan*[*]*, Jinshi Cui*[*]*, Franck Davoine*[†]*, Li Wang*[⋆] *and Hongbin Zha*[*]

[*]Key Lab. of Machine Perception (MoE), School of EECS, and
[⋆] Psychology Department, Peking University, Beijing, China
[†] CNRS, LIAMA Sino-European Lab., Beijing

ABSTRACT

Head pose is an important indicator of a person's visual focus of attention (VFoA). A traditional way to recognize VFoA is to consider accurate head pose or gaze estimations. However, these estimations usually degrade drastically in middle or low resolution video data. In this paper, a joint estimation of head pose and VFoA is proposed to address this issue; both head pose and VFoA are iteratively refined until convergence. This approach is evaluated in a specific scenario involving children around a table playing together with toys. Datasets are acquired and annotated by psychologists in Peking university. The experimental results demonstrate the usefulness of the join estimation process to recognize visual focus of attention in middle resolution video sequences.

Index Terms**—** Head pose estimation, visual focus of attention, joint estimation, low resolution

1. INTRODUCTION

The orientation and movements of a human head is an important indicator to infer the intentions of others and facilitate nonverbal communication. However, the essence behind head pose is what people is looking at, i.e., the visual focus of attention (VFoA). Head pose estimation and VFoA recognition systems can be applied in different areas: traffic safety [1, 2], behavior analysis [3, 4], smart meeting rooms[5, 6], intention prediction[7].

There are mainly two different ways of seeing the estimation of head pose. One treats it as a classification problem to get discrete classes of head pose [8, 9], the other treats it as a regression problem to get continuous head pose [10, 11]. When limited to low resolution images, discrete classes are usually considered to estimate head poses.

Some researchers try to reason out VFoA based on head pose, eye gaze or environment context [6, 12]. However, it is not easy to estimate accurate head pose and even harder to estimate gaze from low or middle resolution images, leading to accumulative errors. It can be seen in [13] that coarse head pose estimation is hard to get satisfying performance even using data with clean backgrounds or correct alignment of head regions.

Additional information must be added to break these limitations. In some applications, there are some specific targets considered to be the candidates of VFoA. For example, in children's social psychological adjustment, eye contact and joint attention are very important cues. In this situation, if a child is looking at a participant, and which participant he is looking at, are critical information. More specifically, if we want to detect eye contact, the head of other participants are the candidates of VFoA.

Motivated by the reasons above, our approach considers low resolution video data and uses a joint iterative refining process between coarse head pose estimation and candidates of VFoA. This approach is compared to [14] and [15]. In [14], authors used the initial coarse head pose given by [16], just as we do, and refine the head pose for VFoA candidates using a combined way. We will call it as a combined method in the experiment part. In [15] geometric constraints and head pose are used to recognize the VFoA.

2. JOINT ESTIMATION OF HEAD POSE AND FOCUS OF ATTENTION

2.1. Problem formulation

In our approach, VFoA candidates are used to offer extra information in order to refine the estimation of head pose and provide a new way to do VFoA recognition in low or middle resolution video.

The probabilities of each head pose class are denoted by pc. The initial probability distribution of VFoA target candidates pt is chosen as a uniform distribution. At the same time, locations and sizes of VFoA candidates are recorded as I_{VFoA}. The location and size of heads is recorded as I_{head}.

$$\begin{aligned} pc_n &= f(pc_{n-1}, pt_{n-1}, I_{VFoA}, I_{head}) \\ pt_n &= g(pc_{n-1}, pt_{n-1}, I_{VFoA}, I_{head}) \end{aligned} \tag{1}$$

n notes the number of iteration times. So the goal is to get the most probable head pose and VFoA from $\{pc, pt, I_{VFoA}, I_{head}\}$. An iterative process is adopted to refine pc and pt in Eq.(1).

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

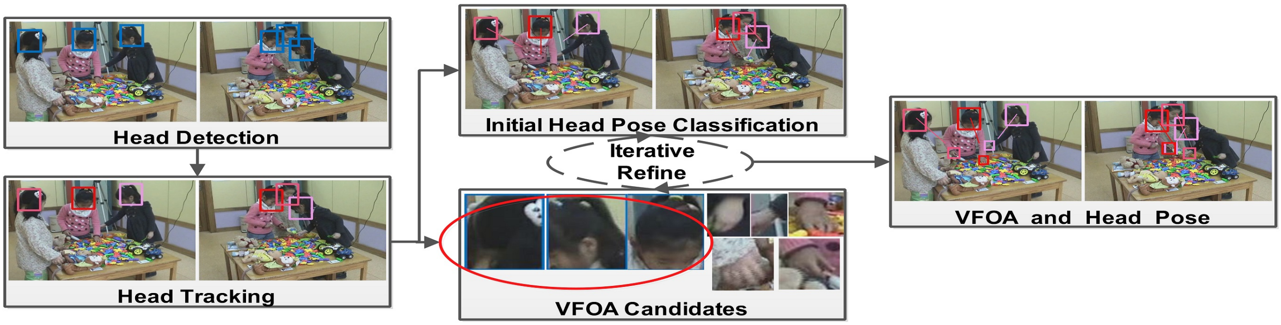

Fig. 1. Framework of this paper. Original data was collected by the psychology researchers from Peking University

2.2. Joint estimation model

Head are considerd as VFoA candidates. Head regions $R_{head} = \{x_{head}, y_{head}, w_{head}, h_{head}\}$ can be obtained using available head detection and tracking algorithms. Furthermore, hand information can be added as another kind of VFoA candidates, which is recorded as $R_{hand} = \{x_{hand}, y_{hand}, w_{hand}, h_{hand}\}$.

The initial probability distribution of head pose is given by initial head pose classification, which can be written as $\{pc(1), pc(2), ..., pc(K)\}$. K is the total number of head pose classes. The initial probability distribution pt of all VFoA candidates is set as an uniform distribution.

2.2.1. Refining the probability distribution of VFoA

The probability that a target j is the VFoA of a person i is calculated as in Eq. (2):

$$pt_n^i(j) = \sum_{k=1}^{K} pc_n^i(k) \cdot p_{sd} \cdot p_{ac} \cdot p_{hc} + pt_{n-1}^i(j) \tag{2}$$

$pc_n^i(k)$ is the probability that head i belongs to the kth head pose class at the nth iteration. p_{sd}, p_{ac} and p_{hc} are defined as:

- Size and distance factor: We suppose that a target object j is more probably a VFoA if it is large and close to the head i. s_j denotes the size of target j. d_{ij} is the distance between head i and target j.

$$p_{sd} = s_j / d_{ij}^2 \tag{3}$$

- Angle consistency factor: if the line connecting the object and the head is consistent with head pose, this factor would be bigger. a_k is the head pose degree represented by class k. $a_{i\to j} = atan((x_i - x_j)/(y_i - y_j))$, $a_{i\to j}$ is the angle between head i and target j.

$$p_{ac} = e^{-(a_k - a_{i\to j})^2/2\sigma_1^2} \tag{4}$$

- Height consistency factor: if the head pose belongs to the head-up class, candidate in the higher place more likely to be noticed and vice versa. h represents the vertical location. In our case, classes 1 to 7 are for head-down faces and classes 8 to 14 are for head-up faces.

$$p_{hc} = \begin{cases} e^{-(h_i/2 - h_j)^2/2\sigma_2^2} & pose \in headDown \\ e^{-(h_i - h_j)^2/2\sigma_2^2} & pose \in headUp \end{cases} \tag{5}$$

σ_1 and σ_2 are constants respectively set to 10 and 20 consider to the 3 sigma rule and the visual angle of human beings. At the beginning of iteration, the probabilities of all VFoA candidates are set to zero, i.e., $pt_0^i(j) = 0$.

pt is not normalized because it is possible that people look at non-candidate (i.e. non-head, non-hand) targets.

2.2.2. Refining the probability distribution of head pose

Now we assign to each target a probability of being the VFoA of each person. The probability can be used to refine the initial head pose estimation.

$$\begin{aligned} & pc_{n+1}^i(k) = \lambda\left(\sum_{j=1}^{N} pt_n^i(j) \cdot p_{ac} + pc_n^i(k)\right) \\ & s.t. \quad \sum_{k=1}^{N} pc_{n+1}^i(k) = 1 \end{aligned} \tag{6}$$

$pt_n^i(j)$ is the result calculated with Eq. (2). $pc_n^i(k)$ is the probability of head i belongs to the kth class in the tth iteration. N is the total number of candidate targets and λ is the normalization parameter. The probability of head i belongs to head pose class k is updated through Eq. (6).

2.2.3. Iterative refining process

An iterative process is based on Eq. (2) and Eq. (6). At each iteration n, a candidate with maximal probability $pt_n^i(j)$ is considered as the VFoA. The head pose class with maximal probability is seen as the refined head pose class.

$$target_n(i) = \begin{cases} \arg\max_j(pt_n^i(j)) & \max_j(pt_n^i(j)) > threshold \\ Null & \max_j(pt_n^i(j)) \leq threshold \end{cases} \tag{7}$$

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

$$pose_n(i) = \arg\max_k (pc_n^i(k)) \tag{8}$$

A threshold is set because participants are not always looks at a VFoA candidate. If the probability is below the threshold, it is supposed that the person is looking at other objects. The iterations end when both VFoA and head pose estimations converge to stable states.

3. PRE-PROCESSING

In this section, we describe three methods that are required prior to jointly estimate the head poses and VFoA.

3.1. Head detection

Reliable initial head regions are very important for head pose estimation and can provide trust-worth candidates for VFoA. Head detection is divided into two main parts: creating multivariate information and voting for each pixel based on this information.

[17–19] are implemented into our algorithm to detect head, upper body and skin color. In addition, hair color detection runs on each frame. For each pixel i, we evaluate the probability that its 5*5 neighborhood patch $patch_i$belongs to a head region. This probability is given by Eq. (9):

$$p^i_{head} = \frac{h_i}{Z} \sum_{patch_i} \mathbf{V}^j_{fore}(\alpha \mathbf{V}^j_{up} + \beta \mathbf{V}^j_{head} + \gamma \mathbf{V}^j_{skin} + \lambda \mathbf{V}^j_{hair}) \tag{9}$$

$j \in patch_i$. h_i is the height of pixel i. It is supposed that heads are more likely to appear at higher locations in the image. $\alpha, \beta, \gamma, \lambda$ are constants. Z is the normalization parameter. $\mathbf{V_{fore}}$ is the binary result given by background substraction which marks all foreground pixels with ones and background pixels with zeros which is similar with the rest $\mathbf{V}$.

$$pixel_i \in \left\{ \begin{array}{cc} head & p^i_{head} \geq threshold \\ non-head & p^i_{head} < threshold \end{array} \right. \tag{10}$$

Using Eq. (10), we can get the pixels which are regarded as belonging to head regions. Holes are filled into connected component and morphological algorithm is used to refine the results. When there are merged components, skin or hair color region is used to segment the components. Then the bounding boxes are given according to the gravity centers and boundaries of each component and finish the head detection step.

3.2. Head tracking

Here we describe the tracker used to connect and refine the detections in each frame produced in the previous stage. In this step, CamShift[20] is used.

For each appearance, a normalized LAB color histogram is used to represent it. So the appearance likelihood between target $W(i)$ and target $W(j)$ can be calculated as the Euclidean distance $E(i, j)$ between their LAB histograms, and calculated as $D_{app}(i, j) = 1 - E(i, j)/2$. The location distance $D_{dst}(i, j)$ between $W(i)$ and $W(j)$ is computed as $D_{dst}(i, j) = (W(i) \cap W(j))/(W(i) \cup W(j))$.

The similarity $D(i, j)$ between detections i and j respectively in frames at times t and $t + 1$ is given by Eq. (11).

$$\begin{aligned} D(i,j) = D_{app}(&W^{\text{detect}}_{t+1}(i), W^{\text{detect}}_{t}(j)) \\ \cdot D_{dst}(&W^{\text{detect}}_{t+1}(i), W^{predict}_{t+1}(j)) \end{aligned} \tag{11}$$

When somebody is missed in head detection, history information W^{store}_t containing all people is used to get corresponding $W^{predct1}_{t+1}(i)$. The confidence in successfully recovering a failed detection is computed as Eq (12).

$$\begin{aligned} C_k = D_{app}(&W^{predct1}_{t+1}(k), W^{store}_{t}(k)) \\ \cdot D_{dst}(&W^{predct}_{t+1}(k), W^{store}_{t}(k)) \end{aligned} \tag{12}$$

If $C(i)$ is bigger than a threshold learned by experience, then this recover is acceptable and we create a new detection based on the $W^{predict1}_{t+1}$, otherwise we just ignore it.

3.3. Initial head pose classification

Considering the resolution of video data and the precision of further computation, getting face poses of children is considered as a discrete classification problem in [16]. Children's head poses in per-frame are estimated by four steps: Gabor wavelet transformation, principal component analysis, classification of multiple classifiers and majority voting. There are 14 classes of head pose designed for our data. Seven classes of head-down category:$\{-90, -60, -30, 0, 30, 60, 90\}$ and also the same seven classes of head-up category.

4. EXPERIMENTAL RESULTS

4.1. Dataset and Annotation

The data is collected by psychology researchers from Peking university studying children's social withdrawal in peer-play scenarios. It can be seen in Fig. 1, a table covered with toys is in the center of the camera view, three children stand around the table and play together with toys. The video is segmented into 22 clips and each clip lasts for 1 minute. One of clip was annotated for the experiments. The clip we used has 1799 frames and the resolution is 720*480.

In social withdraw study, head and hand are important targets. But hand detection is challenging, so hands are manually labeled to provide the VFoA candidates.

The annotated information includes:

- Head and Hand region: R_{head}, R_{hand},
- Head pose: $pose(i) \in \{0, 1, ..., 13\}$, i notes the ith person;
- VFoA: $target(i)$,the target that ith person is looking at.

4.2. Head Tracking

When $(A_{track} \cap A_{grd})/(A_{track} \cup A_{grd}) > 0.6$ (A_{track} is the area giving by tracking process and A_{grd} is the area of ground truth), it is considered as a successful tracking. Here a universal threshold 0.6 is used in the evaluation. There is no Id-switch in this video clip.

	Child1	Child2	Child3	Average
accuracy	0.9983	0.9944	0.8821	0.9573

Table 1. The accuracy of head tracking.

4.3. Head Pose Estimation

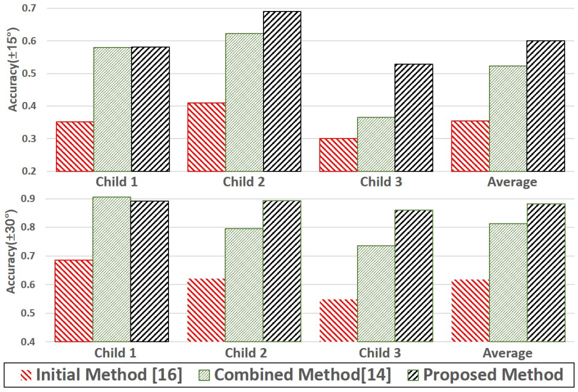

Fig. 2. The comparative experiment of head pose estimation. The combined method is proposed by [14].

The accuracy(±15°) has improved 69.6% compare to initial method propose by [16], and improved 14.8% compared to [14]. The accuracy(±30°) has improved 42.4% compare to initial method propose by [16], and improved 8.6% compared to [14].

4.4. VFoA recognition

	Child1	Child2	Child3	Average
Multi-VFoA	0.8401	0.8280	0.7655	0.8112
Single VFoA	0.9251	0.8892	0.8427	0.8857
No VFoA	0.8524	0.8427	0.8192	0.8381
Average of All	0.8725	0.8533	0.8091	0.8450

Table 2. The accuracy of VFoA recognition of propose method.

Table 2 shows the result of proposed VFoA recognition method. Multi-VFoA means multiple targets are labeled as VFoA for one person due to the ambiguity in visual angle. Single VFoA means the child is looking at one specific target and there is no ambiguity in telling that. No VFoA means the visual focus of attention is not among the VFoA candidates. When the VFoA candidates with top several highest probability are multi-VFoA in the annotation, it is considered as successful recognition of multi-VFoA.

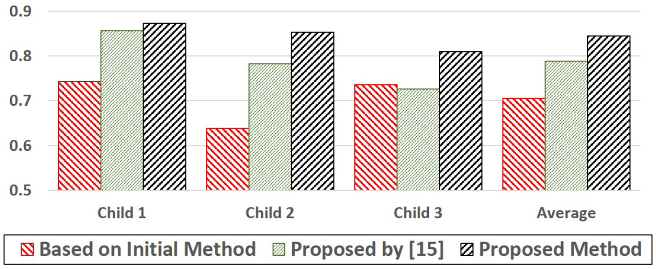

Fig. 3. The comparative experiment of VFoA recognition. The VFoA recognition method is proposed by [15].

The first sequential method is using the VFoA recognition framework of [15] giving the head pose result of [16], the accuracy has been improved 19.8% compared to which. The accuracy has been improved 7.1% compared to [15].

5. CONCLUSION AND FUTURE WORK

Head pose and VFoA are important cues to understand human's behavior. The proposed method uses the result of coarse head pose estimation and VFoA candidates to get refined head pose and VFoA.

Considering the uncertainty in tracking and the clutter background in a natural environment, it becomes a challenging problem to get both accurate head pose and VFoA. But in this situation, iterative refining process still gives better results. The accuracy of head pose estimation is improved 69.6% and 42.4% compared to initial method propose by [16]. The accuracy of VFoA recognition is improved 8.6% compared to sequential method propose by [15]. The proposed method successfully decreases the accumulative errors of traditional sequential methods in VFoA recognition and is more robust to get head pose in low/middle resolution data. The proposed method is expected to be applied into low or middle resolution videos to get more accurate head pose and VFoA.

Acknowledgments: This work was supported by National Basic Research Program (973 Program) of China (No. 2011CB302202), National High Technology Research and Development Program of China (No.2012AA011602), NSFC Grant (No.61273253) and the Sino-French NSFC-ANR Grant (No.61161130528).

References

[1] Erik Murphy-Chutorian, Anup Doshi, and Mohan Manubhai Trivedi, “Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation,” in *Intelligent Transportation Systems Conference, 2007.* IEEE, 2007, pp. 709–714.

[2] Christoph G Keller, Christoph Hermes, and Dariu M Gavrila, “Will the pedestrian cross? probabilistic path prediction based on learned motion features,” in *Pattern Recognition*, pp. 386–395. Springer, 2011.

[3] Bingshu Yang, Jinshi Cui, Hongbin Zha, and Hamid Aghajan, “Visual context based infant activity analysis,” in *Distributed Smart Cameras (ICDSC), 2012 Sixth International Conference on*. IEEE, 2012, pp. 1–6.

[4] Tamas Vajda, “Behavior recognition using pictorial structures and dtw,” in *Automation Quality and Testing Robotics (AQTR), 2010 IEEE International Conference on*. IEEE, 2010, vol. 3, pp. 1–4.

[5] Sileye O Ba and J-M Odobez, “Recognizing visual focus of attention from head pose in natural meetings,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 39, no. 1, pp. 16–33, 2009.

[6] Sileye O Ba and J Odobez, “Multiperson visual focus of attention from head pose and meeting contextual cues,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 1, pp. 101–116, 2011.

[7] Yingning Huang, Jinshi Cui, Franck Davoine, Huijing Zhao, and Hongbin Zha, “Head pose based intention prediction using discrete dynamic bayesian network,” in *Distributed Smart Cameras (ICDSC), 2013 Seventh International Conference on*, 2013, pp. 1–6.

[8] Bingpeng Ma, Wenchao Zhang, Shiguang Shan, Xilin Chen, and Wen Gao, “Robust head pose estimation using lgbp,” in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. IEEE, 2006, vol. 2, pp. 512–515.

[9] Nicolas Gourier, Jérôme Maisonnasse, Daniela Hall, and James L Crowley, “Head pose estimation on low resolution images,” in *Multimodal Technologies for Perception of Humans*, pp. 270–280. Springer, 2007.

[10] Yongmin Li, Shaogang Gong, and Heather Liddell, “Support vector regression and classification based multi-view face detection and recognition,” in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 2000, pp. 300–305.

[11] Yong Ma, Yoshinori Konishi, Koichi Kinoshita, Shihong Lao, and Masato Kawade, “Sparse bayesian regression for head pose estimation,” in *Pattern Recognition, 2006. 18th International Conference on*. IEEE, 2006, vol. 3, pp. 507–510.

[12] Michael Voit and Rainer Stiefelhagen, “Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios,” in *Proceedings of the 10th international conference on Multimodal interfaces*. ACM, 2008, pp. 173–180.

[13] Erik Murphy-Chutorian and Mohan M Trivedi, “Head pose estimation in computer vision: A survey,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 607–626, 2009.

[14] Dingrui Duan, Lu Tian, Li Cui, Jinshiand Wang, Hongbin Zha, and Aghajanz Hamid, “Gaze estimation in childrens peer-play scenarios,” in *Advanced Sensing / Visual Attention and Interaction, International Joint Workshop on*, 2013, pp. 1–6.

[15] Lu Tian, Dingrui Duan, Jinshi Cui, Li Wang, Hongbin Zha, and Aghajanz Hamid, “Video based children’s social behavior classification in peer-play scenarios,” in *2013 2nd Pattern Recognitio, International Asian Conference on*, 2013, pp. 1–6.

[16] Junwen Wu and Mohan M Trivedi, “A two-stage head pose estimation framework and evaluation,” *Pattern Recognition*, vol. 41, no. 3, pp. 1138–1158, 2008.

[17] Michael Jones and Paul Viola, “Fast multi-view face detection,” *Mitsubishi Electric Research Lab TR-20003-96*, vol. 3, pp. 14, 2003.

[18] M. Castrillón Santana, O. Déniz Suárez, M. Hernández Tejera, and C. Guerra Artal, “Encara2: Real-time detection of multiple faces at different resolutions in video streams,” *Journal of Visual Communication and Image Representation*, pp. 130–140, April 2007.

[19] Praveen Kakumanu, Sokratis Makrogiannis, and Nikolaos Bourbakis, “A survey of skin-color modeling and detection methods,” *Pattern recognition*, vol. 40, no. 3, pp. 1106–1122, 2007.

[20] John G Allen, Richard YD Xu, and Jesse S Jin, “Object tracking using camshift algorithm and multiple quantized feature spaces,” in *Proceedings of the Pan-Sydney area workshop on Visual information processing*. Australian Computer Society, Inc., 2004, pp. 3–7.