34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

An Face-based Visual Fixation System for Prosthetic Vision

Xuming He*, Junae Kim* and Nick Barnes*

Abstract— **Recent studies have shown the success of face recognition using low resolution prosthetic vision, but it requires a zoomed-in and stably-fixated view, which will be challenging for a user with the limited resolution of current prosthetic vision devices. We propose a real-time object detection and tracking system capable of fixating human faces. By integrating both static and temporal information, we are able to improve the robustness of face localization so that it can fixate on faces with large pose variations. Our qualitative and quantitative results demonstrate the viability of supplementing visual prosthetic devices with the ability to visually fixate objects automatically, and provide a stable zoomed-in image stream to facilitate face and expression recognition.**

I. INTRODUCTION

The normally sighted eye consists of a high resolution fovea of 0.3-2 degrees, which typically performs high acuity tasks such as face recognition, facial expression recognition and reading, with peripheral vision which reduces exponentially in terms of retinal cell density from the fovea [16]. Eye movements are used to bring target objects to the fovea, maintaining a stabilized fixation and enabling the fovea to view an object long enough that it produces a stable, rather than blurred image [10]. It has been shown that a longer stabilized fixation boosts the performance of high acuity tasks, e.g., face recognition [5], [11].

The state-of-the-art prothetic vision shares a similar, if not more challenging, task as the human vision system in terms of maintaining fixation. To date, for the visual percept resulting from electrical stimulation, called phosphenes, the largest numbers reported are around 100 (eg., [1]), while a large set of functional results are demonstrated on a 60 electrode array [6]. This small total number of visual field elements creates a difficulty for high acuity tasks as it would appear necessary to devote all the resolution to the task at hand. This leaves no surrounding peripheral visual field to maintain fixation in the manner of normal human vision. For example, in face recognition, Thompson [13] demonstrates that it is possible to discriminate a small number of faces to more than 90% accuracy based on just 32x32 phosphenes, with performance dropping significantly with resolution reduction. Here the entire field of view was taken up by the face – without the typical human visual fixation mechanism, this would be difficult to achieve in a normal human environment. Figure 1-A shows a typical example of indoor face identification scenario, in which the input from camera is displayed in normal and zoomed-in ways. Note that, using about 1000 simulated phosphenes, we can barely see the person in the whole image.

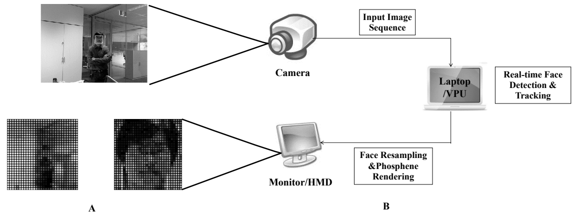
Camera
Input Image Sequence
Laptop /VPU
Real-time Face Detection & Tracking
Monitor/HMD
Face Resampling &Phosphene Rendering
A
B

Fig. 1. (A) Simulated phosphene rendering of an image of a scene and an object inside.Top: input image of video camera; Bottom: simulated phosphene rendering of the whole image and the face with 30x30 resolution. (B) An overview of our face fixation system for visual prothesis.

Many image processing methods have been applied in bionic eye devices to address the problem of the limited view or resolution of retinal implants, such as Gaussian filtering, edge extraction, image transformation [3], [4], [18]. A range of functional vision tasks, including object recognition and navigation, can be facilitated by these image processing steps. However, the existing experimental evaluation is usually constrained to laboratory environments with ideal setup of viewing angle/range [14], [18], which is unrealistic in daily activities. In particular, recognizing objects and faces usually requires zooming on the target object and then applying a transformation to render the full object on the device. This fixation mechanism generally involves higher-level visual concepts such as object. As such, simple image-based processing will not be able to provide such mechanism.

In this paper, we propose a visual object-based fixation system in which a broader image from the input camera provides the peripheral image, and computer vision techniques are used to provide the ability to zoom and maintain fixation on specific objects. Our system holds the dynamically moving object centered in the viewing area to allow the viewer the possible better recognition performance that a dynamic view offers. In particular, we demonstrate a face-based fixation system for face recognition, as restoring the ability to recognise faces and facial expression are known to be key requirements for a visual prosthetic [7]. Our design is based on recent progress on face detection and tracking in computer vision [8], [12], in which we employ an improved boosted cascade classifier for face detection and incorporate temporal dynamic information to stabilize detection output. A systematical evaluation on a benchmark dataset shows zoomed faces are kept still in continuous videos, and resulting simulated phosphene images are much more stable than baseline methods.

The paper is organized as follows. In Section II, we introduce our system and the details of the proposed fix-

*All authors are with NICTA, Canberra ACT, Australia and CECS, Australian National University, Canberra ACT, Australia.

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

ation algorithm. The experimental evaluation is described in Section III in which we compared with the baselines qualitatively and quantitatively. Finally, we conclude the paper in Section IV.

II. Our Method

Our system captures visual input by a standard camera, and the input frames are fed into a laptop with our video processing and simulated phosphene image display software. The output of the system is a phosphenized face image shown on a head-mounted display or computer screen, and an option button is also provided to user so that a user can choose which face is targeted and how much zooming is needed. An overview of our method is shown in Fig 1-B.

A. *Face detection and tracking*

The face detection component employs a similar approach as in [15] that takes a fixed-size square window at every position on the image plane and classifies the image patch within each window as 'face' or 'non-face'. This process is repeated on several gradually down-scaled images so that faces with different sizes can be detected. We adopt an improved cascade boosted classifier, referred as Lacboost [12], for the detection task. This classifier provides a superior detection performance to Adaboost and requires a smaller number of image features, which is critical for real-time processing in complex real-world scenario. The key idea of Lacboost is to reuse the weak learners from previous stages in the cascade pipeline, and tune the model parameters with an asymmetric cost function that fits better in detection tasks.

Despite the high detection rate in the frame-based image processing, it is still insufficient for a visual prothesis due to lack of temporal stability, background distraction, and pose variation. To overcome these issues, we integrate the face detector with an online visual tracking component in a Bayesian framework. The tracking component incorporates temporal smoothness constraint and a face appearance model for precise localization. The overall module is illustrated in Figure 2 with an example of detection and tracking integration.

More specifically, let $\mathbf{x}_t = (x_t, y_t, s_t)^T$ denote the position and scale of a target face at time t, and I_t is the input image frame. As in [8], we are interested in computing the posterior of the current state $\mathbf{x}_t$ given the image sequence $I_{1:t} = \{I_1, \cdots, I_t\}$. The Bayesian filtering approach computes the posterior in a recursive way,

$$P(\mathbf{x}_t|I_{1:t}) \propto \int_{x_{t-1}} P(\mathbf{x}_t|\mathbf{x}_{t-1})P(I_t|\mathbf{x}_t)P(\mathbf{x}_{t-1}|I_{1:t-1})dx_{t-1} \tag{1}$$

where $P(\mathbf{x}_t|\mathbf{x}_{t-1})$ is the transition probability imposing temporal smoothness, and $P(I_t|\mathbf{x}_t)$ is the data likelihood at time t. We design these two model components as follows.

The data likelihood integrates a generic face detection and an appearance-based person-specific face matching procedure. In the generic face detection, we represent its result as a mixture model:

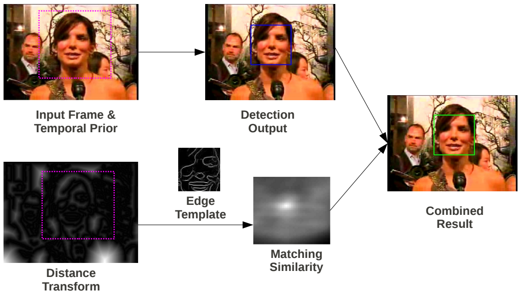

Fig. 2. Detection and tracking algorithm: integration of detection result and edge template based tracking in our method.

$$P_d(\mathbf{x}_t|I_t) = \sum_{o_t} P(x_t, y_t, s_t|o_t)P(o_t|I_t) \tag{2}$$

where o_t is a binary indictor variable specifying whether the face is visible in image frame I_t. For face detection, we transform the detection score to the range of $[0, 1]$ so that it can be viewed as the probability of $P(o_t = 1|I_t)$. Let the detection window be $\hat{\mathbf{x}}_t^d$, we assume the conditional probability of state given visibility has the form

$$P(\mathbf{x}_t|o_t) \propto \begin{cases} K(\mathbf{x}_t, \hat{\mathbf{x}}_t^d; \sigma_d) & \text{if} \quad o_t = 1 \\ \epsilon & \text{if} \quad o_t = 0 \end{cases} \tag{3}$$

where $k(x, \hat{x}; \sigma)$ is a Gaussian kernel function $\exp(-\frac{||x-\hat{x}||^2}{\sigma^2})$ and ϵ is a small positive constant ($\epsilon \ll 1$).

For the appearance-based face matching component, we build a weighted edge template model of faces based on the first detection result. Given the detected face window, we normalize it into 50x50 and apply Canny edge detector to extract the edge template M_e. The edge template is also associated with a weight mask W_e that emphasizes the central area of the detection window. We update the edge template in an online fashion such that it can adapt to the pose changes. Specifically, we define the weight mask as a two-dimensional Gaussian on the normalized face window and its standard deviation equals to the half of window size. An example of matching score can be seen in Figure 2.

During face fixation at time t, we search for the best matches of the template with the edges in input frame I_t in the neighborhood of the face detection result $\hat{\mathbf{x}}_t^d$ as well as of the localization result $\hat{\mathbf{x}}_{t-1}$ at $t-1$ (see the following for its definition). We denote the search region as N_t. The matching distance is defined by a weighted Chamfer distance [2], which can be computed efficiently using a distance transform:

$$\begin{aligned} \hat{\mathbf{x}}_t^a &= \arg\min_{\mathbf{x}\in N_t} D_{Chamf}(Canny(R_t(\mathbf{x})), M_e; W_e) \\ &= \arg\min DT(Canny(R_t(\mathbf{x}))) \cdot (M_e \odot W_e) \end{aligned} \tag{4}$$

where D_{Chamf} is the Chamfer distance and $R_t(\mathbf{x})$ are the image windows in N_t. DT represents the distance transform. We use $\cdot$ for inner product and $\odot$ for element-wise product between vectors. Our system searches three different scale factors $s_t = \{s_{t-1}-\delta, s_{t-1}, s_{t-1}+\delta\}$ where s_{t-1} is the scale of a target face at previous time $t-1$. δ is the discretized scale step.

The minimum matching results is represented by a Gaussian distribution, $P_a(\mathbf{x}_t|I_t) \propto K(\mathbf{x}_t, \hat{\mathbf{x}}_t^a; \sigma_a)$, and overall data likelihood is defined by combining the above two components probabilistically:

$$P(I_t|\mathbf{x}_t) \propto P_d(\mathbf{x}_t|I_t)P_a(\mathbf{x}_t|I_t) \tag{5}$$

We adopt a simple linear Gaussian transition probability $P(\mathbf{x}_{t+1}|\mathbf{x}_t) \propto K(\mathbf{x}_{t+1}, \mathbf{x}_t; \sigma_p)$ to enforce the temporal smoothness, where σ_p controls the temporal smoothness. In order to compute the prediction in real time, we employ the following approximation because the posterior $P(\mathbf{x}_{t-1}|I_{1:t-1})$ is usually peaky around its mode $\hat{\mathbf{x}}_{t-1}$:

$$P(\mathbf{x}_t|I_{1:t}) \approx P(\mathbf{x}_t|\hat{\mathbf{x}}_{t-1})P(I_t|\mathbf{x}_t) \tag{6}$$

$$\hat{\mathbf{x}}_t = \arg\max P(\mathbf{x}_t|I_{1:t}) \tag{7}$$

Note that we combine image cues from multiple sources to maintain the robustness and stability of fixation.

B. Display in simulated prosthetic vision

Detected face regions are cropped and normalized to target size for display on a boinic eye device, including retinal implants, head-mounted displays or computer screens for simulation. In this paper, we report our results based on the simulated phosphene display that commonly used in retinal implants simulation [14]. Our phosphene image representation is based on Gausssian kernel profiles placed on a 35x30 rectangular grid. The center value and standard deviation of each Gaussian kernel are proportional to the pixel intensity at its center position. For each phosphene, we limit the intensity value to 6 bit. We refer the readers to [9] for more details.

III. Experimental Evaluation

A. Dataset and Setup

Our quantitative experimental evaluation is based on a publicly available YouTube face video dataset[1] [17]. The video clips involve camera and head movements which induce changes in face position and orientation, such as side-to-side, nodding, and tilting movements. We categorized those video clips into 5 classes according to the maximum pose variation in each clip. Each class spans 15-degree intervals and we consider 5 classes as ($[0°, 15°]$, $[15°, 30°]$, $[30°, 45°]$, $[45°, 60°]$ and more than 60°). We choose 20 clips per class in this experiments. To make the comparison consistent, we adjust the starting time of each sequence such that in the first frame, the faces show their frontal view, and limit the length of each video clip to 50 frames.

We implemented our software system on a Laptop with Intel Core i7-2620M CPU running at 2.7GHz. Overall, our system runs in realtime and generates output at 18 frames per second on average. We also implement two baseline methods of face detection/tracking for comparison purposes. The first is the classical Viola-Jones face detector [15] applied to each frame. The second is a color-histogram based tracking, which builds a face appearance model from an initial face detection.

To measure the fixation accuracy, we compute the distance between the center of face fixation outcome and the center of the ground truth (manually labeled) on the image plane, normalized by the size of ground truth face windows. The average distance on the dataset will indicate how close the predicted face locations are to the ground truth. Specifically, we normalize the image plane such that the ground-truth face window has a size of 50×50 across all the instances. Thus the average accuracy score for a clip with T frames is computed as follows.

$$F_A = \frac{1}{T}\sum_{t=1}^{T}\sqrt{(\hat{x}_t - x_t^g)^2 + (\hat{y}_t - y_t^g)^2}/s_t^g \times 50 \tag{8}$$

where (x_t^g, y_t^g, s_t^g) is the ground truth at frame t. We also average over clips to obtain the mean fixation accuracy $\bar{F}_A$.

B. Results and Discussion

Our system can reliably detect and robustly track faces within distance of 0.5 to 5 meters in a normal indoor environment using a consumer camera. In Figure 3, we show three examples of face fixation results of our system. The first example is taken in an office environment and shows four individual frames across a live sequence with marked face areas in high resolution are demonstrated. The second example is from the Youtube dataset. For each we also show the resulting simulated phosphene images. In these examples, we can see that the zoomed face window provides rich and informative cues for identity and expression recognition in the restricted resolution phosphene images, which would be missed by viewers if we could not fixate on the face area.

We also show quantitative results of detection and tracking stability and the comparison with other baseline face localization methods. The performance is measured by the average face localization accuracy $\bar{F}_A$ as in Equation 8, which accounts for different sizes of face instances. For the detection-only method, if it fails, we use a default left-upper corner position of the image plane as its output. The overall average scores across the whole evaluation dataset are shown in Table I. Our method yields improved fixation performance compared to the baselines: the average fixation precision of our system is 4.07 pixels deviation from the manual labeling, showing 47.16% and 12.74% improvement over two baselines respectively. Figure 4 shows the resulting simulated phosphene images from another example in the Youtube dataset. We notice that the detector alone often failed to find the face in the sequence due to facial orientation. The tracker can follow the facial regions most of time but often drifts away from the center of face (in the second example, it is shifted to the person's neck) due to background distraction or clutter. On the other hand, our system generates stable phosphene face images in these challenging conditions.

IV. Conclusion

We have presented a prosthetic vision eye fixation system in which a full-view image from a high-resolution camera

[1] Available from http://www.cs.tau.ac.il/~wolf/ytfaces/

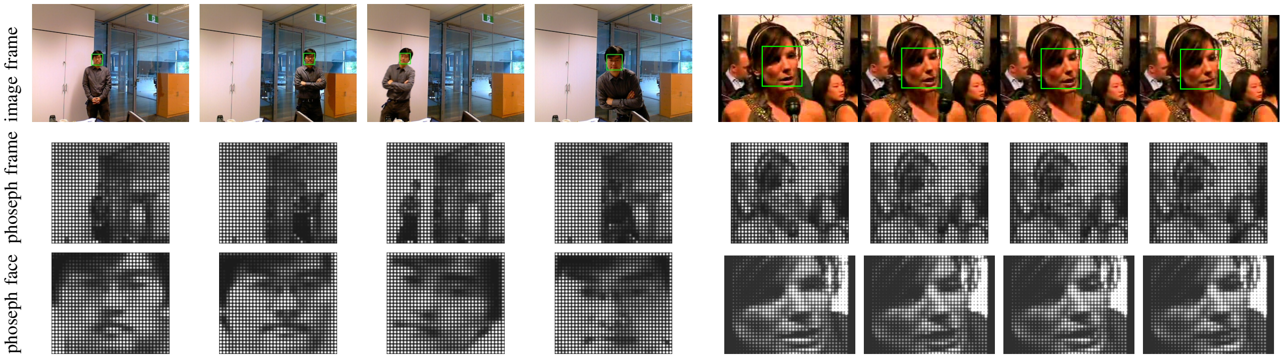

Fig. 3. Examples of face fixation in video sequences. Left: face fixation at varying distances. Right: fixation with background distraction and clutter.

TABLE I

THE COMPARISON OF OUR SYSTEM VS. BASELINES W.R.T FIXATION ACCURACY IN PIXELS.

Accuracy	Detector	Tracker	Our System
F_A	6.00 ± 11.49	4.59 ± 2.79	4.07 ± 2.32

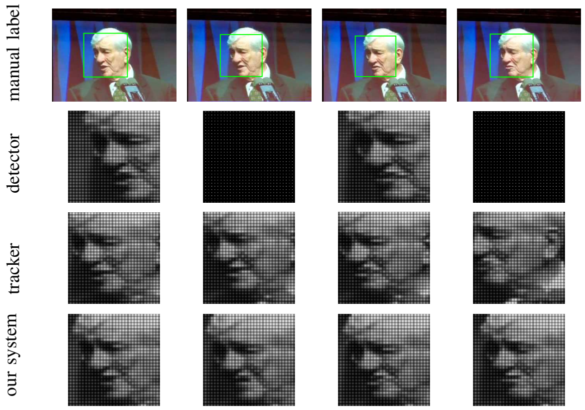

Fig. 4. Example of comparison with baselines. The detection output in the second row (black indicates detection failure), tracker result at the third row, and our system output in the last row.

provides the peripheral view, and state-of-the-art computer vision techniques is used to provide the ability to zoom and maintain fixation. Based on a series of simulated experiments on a natural face video dataset, we demonstrate that our visual object-based fixation system is capable of detecting and tracking faces under various challenging environments, and generating a consistent phosphenized face image sequence to compensate for camera and face motions. This result, we believe, will be particularly useful for facilitating face recognition, and restoring the ability to recognise facial expression.

ACKNOWLEDGMENT

The authors acknowledge the contributions of Paulette Lieby and Adele Scott, authors of the simulated prosthetic vision software used in this paper.

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications, and the Digital Economy, and the Australian Research Council (ARC) through the ICT Centre of Excellence Program. This research was also supported in part by ARC through its Special Research Initiative (SRI) in Bionic Vision Science and Technology grant to Bionic Vision Australia (BVA).

REFERENCES

[1] G S Brindley and W S Lewin. The sensations produced by electrical stimulation of the visual cortex. *Journal of Physiology*, 196(2):479–493, May 1968.

[2] G. Brogefors. Hierarchical chamfer matching: A parametric edge matching algorithm. *IEEE TPAMI*, 10:849–865, 1988.

[3] Spencer C. Chen, Gregg J. Suaning, John W. Morley, and Nigel H. Lovell. Simulating prosthetic vision: I. Visual models of phosphenes. *Vision Research*, 49(12):1493–1506, 2009.

[4] LE Hallum, SL Cloherty, and NH Lovell. Image analysis for microelectronic retinal prosthesis. *Biomedical Engineering, IEEE Transactions on*, 55(1):344–346, 2008.

[5] J H Hsiao and G Cottrell. Two fixations suffice in face recognition. *Psychological Science*, 19(10):998–1006, 2008.

[6] MS Humayun, L da Cruz, G Dagnelie, S Mohand-Said, P Stanga, RN Agrawal, RJ Greenberg, and Argus II Study Group. Interim performance results from the second sight(r) argustm ii retinal prosthesis study. In *ARVO*, 2010.

[7] J E Keeffe, K L Francis, C D Luu, N Barnes E L Lamoureaux, and R H Guymer. Expectations of a visual prosthesis: perspectives from people with impaired vision. In *ARVO*, May 2010.

[8] Sanjiv Kumar, Vladimir Pavlovic, and Henry Rowley. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*, 2008.

[9] Paulette Lieby, Nick Barnes, Chris McCarthy, Nianjun Liu, Liu Dennett, Janine Walker, Viorica Botea, and Adele Scott. Substituting depth for intensity and real-time phosphene rendering: Visual navigation under low vision conditions. In *EMBC*, September 2011.

[10] K Rayner. Eye movements in reading and information processing: 20 years of research. *Pscyhological Bulletin*, 124(3):372–422, 1998.

[11] D A Roark, S Barrett, M Spence, H Abdi, and A J O'Toole. Memory for moving faces: Psychological and neural perspectives on the role of motion in face recognition. *Behavioural and Cognitive Neuroscience Reviews*, 2(1):15–46, Mar 2003.

[12] Chunhua Shen, Peng Wang, and Hanxi Li. Lacboost and fisherboost: Optimally building cascade classifiers. In *ECCV*, 2010.

[13] R. W. Thompson. Facial Recognition Using Simulated Prosthetic Pixelized Vision. *Investigative Ophthalmology & Visual Science*, 44(11):5035–5042, 2003.

[14] J.J. van Rheede, Christopher Kennard, and S.L. Hicks. Simulating prosthetic vision: Optimizing the information content of a limited visual display. *Journal of vision*, 10(14):1–14, 2010.

[15] Paul Viola and Michael J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57:137–154, 2004.

[16] V Virsu, R Nasanen, and K Osmoviita. Cortical magnification and peripheral vision. *JOSA A*, 4(8):1568–1578, 1987.

[17] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, 2011.

[18] Ying Zhao, Yanyu Lu, Yukun Tian, Liming Li, Qiushi Ren, and Xinyu Chai. Image processing based recognition of images with a limited number of pixels using simulated prosthetic vision. *Information Sciences*, 180(16):2915–2924, 2010.