Research Article Vol. 57, No. 1 / January 1 2018 / Applied Optics 69

applied optics

Hybrid sensing face detection and registration for low-light and unconstrained conditions

Mingyuan Zhou,[1,*] Haiting Lin,[1] S. Susan Young,[2] and Jingyi Yu[1]

[1]University of Delaware, Newark, Delaware 19716, USA
[2]U.S. Army Research Laboratory, 2800 Powder Mill Road, Adelphi, Maryland 20783, USA
*Corresponding author: mzhou@udel.edu

Received 23 August 2017; revised 25 October 2017; accepted 1 December 2017; posted 1 December 2017 (Doc. ID 305383); published 21 December 2017

The capability to track, detect, and identify human targets in highly cluttered scenes under extreme conditions, such as in complete darkness or on the battlefield, has been one of the primary tactical advantages in military operations. In this paper, we propose a new collaborative, multi-spectrum sensing method to achieve face detection and registration under low-light and unconstrained conditions. We design and prototype a novel type of hybrid sensor by combining a pair of near-infrared (NIR) cameras and a thermal camera (a long-wave infrared camera). We strategically surround each NIR sensor with a ring of LED IR flashes to capture the "red-eye," or more precisely, the "bright-eye" effect of the target. The "bright-eyes" are used to localize the 3D position of eyes and face. The recovered 3D information is further used to warp the thermal face imagery to a frontal-parallel pose so that additional tasks, such as face recognition, can be reliably conducted, especially with the assistance of accurate eye locations. Experiments on real face images are provided to demonstrate the merit of our method.

OCIS codes: (100.3005) Image recognition devices; (100.3008) Image recognition, algorithms and filters; (110.2960) Image analysis.

https://doi.org/10.1364/AO.57.000069

1. INTRODUCTION

There is a need for detecting and recognizing humans under low-light conditions in many applications, such as in the military and industrial domains. For example, in covert military operations, it is critical to confirm a person's identity before action. Over the past decades, tremendous advances [1–5] have been made in face detection and recognition in the visible spectrum. Two fundamental challenges remain for low-light conditions. First, in low light and especially in complete darkness, the acquired visible face images are severely corrupted by noise. Reliable face detection/matching is difficult with such noisy data. Although near-infrared (NIR) imagery can be used to mitigate this scenario, the use of continuous NIR illumination is not covert, can easily be detected, and has safety issues for human eyes since prolonged IR exposure can damage the lens, cornea, and retina. In contrast to visible spectrum and NIR imaging, long-wave infrared (LWIR) cameras can capture nearly noise-free images even under complete darkness. One study [6] demonstrated that the unique properties of LWIR facial imagery provide many advantages for face detection and recognition, and that the LWIR facial imagery is robust to illumination, expression change, aging, etc. Second, nearly all existing algorithms and databases assume that the acquired face pose is frontal. In reality, face images captured in uncooperative conditions generally exhibit strong 3D orientations, even for captured thermal faces. These inputs need to be accurately rectified before conducting face detection and recognition. Most existing robust face frontalization approaches [7–11] for normal color images need reliable facial features for pose estimation (such as eye landmarks). The performance is inferior for thermal images where the eyes do not have discriminative features.

In this paper, we propose a collaborative, multi-spectrum hybrid sensing method to achieve face detection and registration under (1) poor lighting conditions, such as in complete darkness, (2) uncooperative conditions in which the face images exhibit strong 3D pose orientations, and (3) covert operations. The method is based on exploiting the "red-eye," or more precisely, the "bright-eye" phenomenon, by which human eyes can be captured by NIR cameras with an NIR flash in the dark. In constructing a new type of hybrid sensor by combining a pair of NIR cameras with a thermal camera, the acquired bright eyes from a pair of NIR cameras can be used to accurately determine the face pose and geometry. Then, the 3D pose orientations captured in the thermal face image can be compensated for in non-frontal face detection and recognition.

The proposed method consists of five components: (1) the newly designed and prototyped hybrid sensing imaging system,

(2) 3D eye localization, (3) additional facial landmark detection, (4) pose estimation, and (5) pose correction and face frontalization.

In the first component, we design a hybrid sensing imaging system that consists of one stereo pair of NIR cameras and a LWIR thermal camera. The thermal camera is used to capture potential targets as reliable sources for face identification. The pair of NIR cameras is used to estimate the target distance and provide auxiliary 3D information. Each NIR sensor in our hybrid sensing system is surrounded by a ring of LED IR lamps, which are controlled by a flash control system we designed to strategically capture the "red-eye," or more precisely, the "bright-eye" effect for extremely efficient and accurate 3D eye localization. In addition, our control system uses IR light sources as flashes, which are more covert and safer than using NIR illuminators.

Using NIR images for eye detection to aid in face detection and face recognition was discussed in Refs. [12–14]. Our proposed system not only captures the eyes in NIR images but also rectifies the eyes in the thermal image for face detection and face registration.

In the second component, we acquire a pair of bright-eye images with a pair of NIR cameras. Then we use them to assist in face pose and geometry recovery. Since the correspondences of the eyes are known in the camera settings, we can efficiently compute the 3D eye locations and project them onto the thermal image to obtain the projected eye locations. This reduces the pose estimation problem from six dimensions (3D location and 3D orientation) into one dimension (1D rotation angle around the axis passing through the eye positions).

In the third component, we need additional facial points on the thermal image to resolve the 1D ambiguity in the pose estimation. There are many successful facial point detection methods in visible spectrum images. Most methods [15–19] use local image features and global context information covering the whole face in their optimization. However, such performance would degrade in thermal images due to a lack of information around the eyes. Thus, we modify Sun *et al.*'s cascaded convolutional neural network (CNN) model [19] to detect three additional face landmarks in the lower face (the nose tip and two mouth corners) on the thermal image. An additional trained simple thermal face detector is used to reduce the searching region generated using projected eye locations into a face bounding box. The modified cascaded CNN model is used in this reduced space for facial landmark detection. Because of the accurate eye locations and tightened face bounding box, facial landmarks are also precisely detected.

In the fourth component, we estimate the head pose using these five thermal face landmarks (eyes, nose tip, and mouth corners). In the fifth component, we perform the pose correction and face frontalization based on Hassner *et al.*'s method [10].

Most existing methods for face frontalization directly obtain the 3D surface of the face from a single or multiple images. For instance, the methods in Ref. [11] obtained the facial geometries by aligning a parameterized 3D face model to the query images. Deep learning was employed in Ref. [9] for canonical view estimation of faces. Hassner *et al.* [10] proposed a single 3D reference surface to do face frontalization for all different query images. While their approaches are effective for normal color images when important facial landmarks such as the eye landmarks are reliably detected, their performance is inferior for thermal images where eyes do not have discriminative features. Since more accurate eye landmarks can be extracted using our proposed hybrid sensing system, we can improve Hassner *et al.*'s method [10] to resolve the pose variation in face recognition. Experiments on real face images show that our technique is effective, accurate, and robust for face detection, pose estimation, and face registration under low-light and unconstrained conditions.

In this paper, we consider that the target already exists in our scene. We focus on the work of our new hybrid sensing system in precise thermal face detection and registration. The organization of the paper is as follows: our hybrid sensing system is described in Section 2, our prototype is illustrated and the experimental results are shown in Section 3, and the conclusion and discussion of further work are in Section 4.

2. HYBRID SENSING SYSTEM

In this section, we provide the detailed design of our hybrid multi-spectrum and multi-viewpoint sensing system for accurate face detection and pose standardization. Figure 1 shows the overall pipeline of our system, which consists of five subsystems: (1) image acquisition, (2) eye localization, (3) landmark detection, (4) pose estimation, and (5) face frontalization. The first subsystem consists of a LWIR camera and a pair of NIR stereo cameras that are attached to the left and right sides of the LWIR camera. The LWIR camera is used to detect the potential personal target in total darkness based on the person's thermal emissivity and movement, and then captures the person's thermal face image as the reliable source for face identification. After the potential target is ascertained, the two NIR cameras capture two pairs of stereo images, with and without the bright-eye effect, to produce a stronger bright-eye effect. The second subsystem, eye localization, estimates 3D eye positions from the NIR stereo images, efficiently localizes the eyes on the thermal image, and generates valid thermal face bounding boxes that contain the eyes. These face bounding boxes are further tightened by our trained thermal face detector. The landmark detection subsystem uses these face regions as input into our trained thermal facial landmarks detector. Combining thermal eyes locations with the detection results, we obtain a total of five landmarks on the thermal image (two from projecting 3D eye positions and three directly detected from LWIR images). The pose estimation subsystem estimates the head pose based on these five landmarks by projecting a standard 3D head model onto the thermal image and minimizing the total projection error of the landmarks. The last subsystem performs the thermal face pose correction and face frontalization using the matched points between the standard head model and the thermal query images via soft symmetry, as discussed in Ref. [10]. We elaborate each component and subsystem of our system in the following subsections.

A. NIR Bright-Eye Localization

This subsystem explores using the bright-eye effect to efficiently and reliably detect the 2D eye positions on the NIR

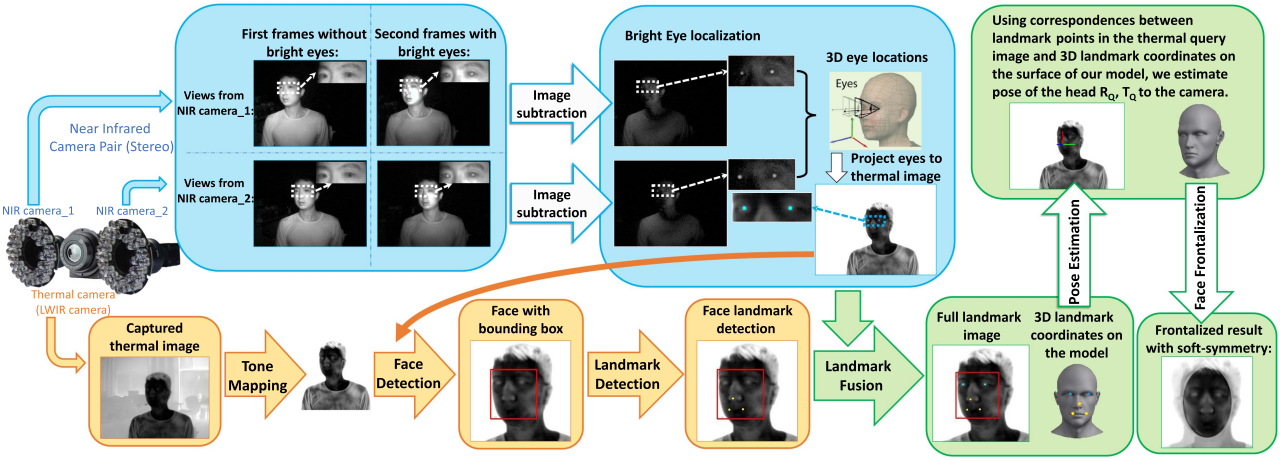


Fig. 1. Pipeline of our hybrid sensing system for face detection and pose standardization.

images, and directly establish correspondences between the eye projections of the image pair. This is different from the traditional depth estimation method, which suffers from correspondence ambiguities. With known correspondences, the 3D eye locations can be accurately recovered through triangulation.

To produce strong bright-eye effects, certain conditions need to be satisfied. The first condition is that (as discussed in Ref. [20]), the bright eyes can be imaged only if the NIR illuminator is beaming along the optical axis of the NIR camera. In such a setting, most of the light from the co-axis NIR illuminator can pass into the eye through the pupil, reflect off the fundus at the back of the eyeball, and come out through the pupil back to the image sensor. This bright-eye effect is similar to a phenomenon in photography known as the red-eye effect, except that now only the NIR part of the spectrum is captured. The bright-eye effect disappears when the NIR illuminator is positioned off the camera's optical axis, because the reflected IR light cannot enter the camera. Figure 2 illustrates the underlying principle of the effects.

The second condition for strong bright-eye effects is that the light should be effectively reflective to the eyes. Behind the retina, there is an ample amount of blood in the choroid, which nourishes the back of the eye. The blood is completely transparent at long wavelength and abruptly starts absorbing at 600 nm [21]. Therefore, we use commodity IR light sources with wavelengths around 800 nm, which are controlled by our flash control system, which we designed to use flashes to effectively acquire the bright-eye effects with NIR cameras.

While low lighting is usually considered harmful for normal imaging situations, it actually strengthens the bright-eye effect in our NIR images. This is because the pupils are fully dilated in the dark and the IR light is minimally absorbed by the ocular pigment. An exceptional advantage of the bright-eye effect is that it is insensitive to pose variations, as shown in Fig. 3. Even if the subject is not facing toward the optical axis of the NIR camera, the amount of IR light reflected by the retina is sufficient to produce the bright-eye effect.

To faithfully extract the positions of the bright eyes from each NIR camera, we instantly capture two sequential frames (within

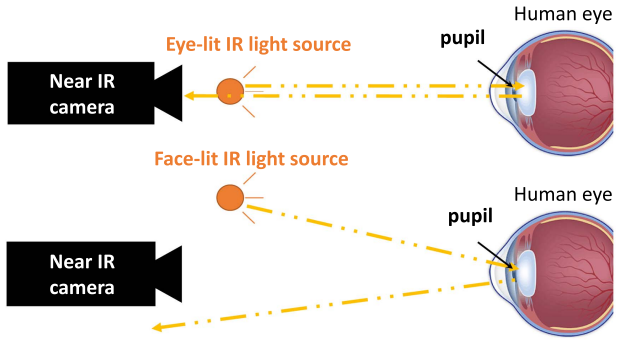


Fig. 2. Principle of the bright-eye effect. When the NIR illuminator is on the optical axis, called the eye-lit IR light source, bright eyes are captured. When the NIR illuminator is off the optical axis, called the face-lit IR light source, no bright eyes are observed in the image.

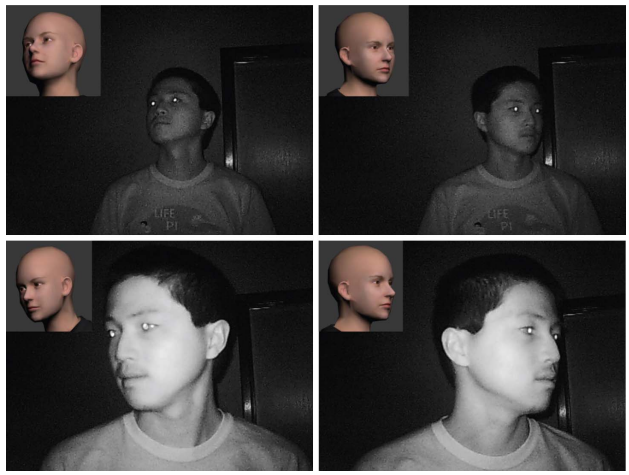


Fig. 3. Bright-eye effects of different poses. The bright-eye effect is insensitive to pose variations under low-lighting conditions.

70 ms) with and without the bright-eye effect. The first frame is the face-lit-only image without bright eyes using off-axis illuminating flashes. The second frame is the eye-lit image with bright eyes using on-axis illuminating flashes. We calculate the difference map between the two frames. With this difference map, simple global thresholding will reveal the eye locations.

There are some special cases that need to be dealt with, such as when the scene has other reflective objects on the faces, such as glasses. To filter out the false positive spots on the difference map, we use the eye features on the eye-lit NIR image to distinguish eye and non-eye objects for those potential eye spots. First, the shapes of the spots are used to filter out some of the false positive spots. As shown in Fig. 4(a), the false positive spots have different shapes and sizes compared with the bright-eye spots, which are usually circular and small. Second, when the spots are of similar shapes, as shown in Fig. 4(b), we compare the features that are extracted from the corresponding positions on the face-lit frame with the standard features of true eyes to further filter out false positive spots.

B. 3D Eye Localization and Projection on the LWIR Image

After the detection of the eyes in the NIR stereo image pair, we first recover the 3D locations of the eyes through triangulation and project the 3D locations of the eyes back to the LWIR image. These are the eye landmarks on the thermal face image.

The triangulation of 3D eye locations is conducted by solving a linear system. As illustrated in Fig. 5, we assume the first NIR camera coordinate system to be the world coordinate system. The projection matrix of each NIR camera is denoted as $P_i = K_i[R_i|T_i]$, where K_i, R_i, and T_i ($i = 1, 2$, which denotes the NIR camera index) are the intrinsic, extrinsic rotation, and translation matrices, respectively, relative to the first NIR camera. Similarly, we define $P_t = K_t[R_t|T_t]$ as the projection matrix of the LWIR camera.

In letting one unknown 3D eye coordinate be (x, y, z) and the corresponding known homogeneous coordinates on the stereo NIR images be (u_1, v_1) and (u_2, v_2), we have the following relations:

$$s_i \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = P_i \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \qquad i = 1, 2, \tag{1}$$

where s_i is unknown scalar parameter.

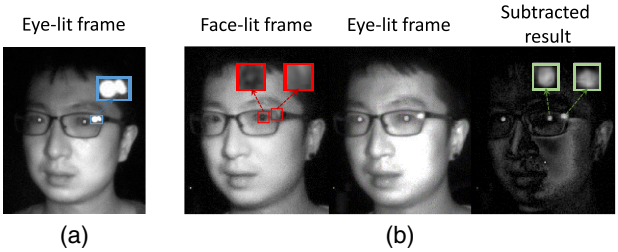


Fig. 4. Filtering out the false positives in bright-eye detection. (a) The false positive spots have different shapes and sizes compared with the bright-eye spots, which are usually circular and small. (b) For the same shapes and sizes, eye feature patterns in the face-lit image are used to filter out the false positive spots.

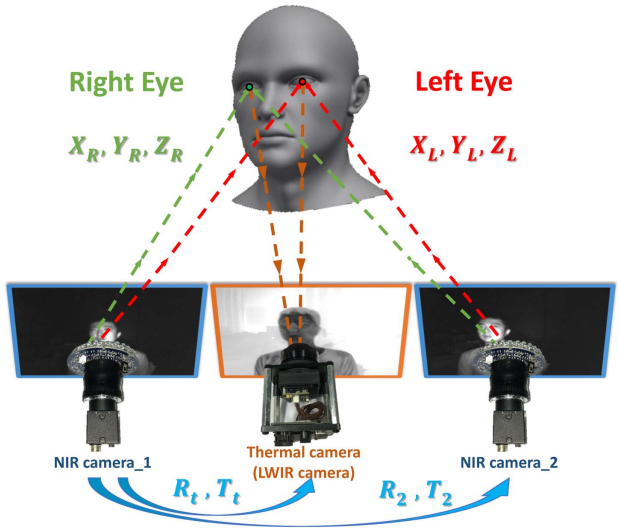


Fig. 5. Triangulation of 3D eye locations.

Two projection matrices, P_1, P_2, one for each NIR camera can be obtained through the camera calibration and can be expressed as

$$P_i = \begin{bmatrix} p_{i,11} & p_{i,12} & p_{i,13} & p_{i,14} \\ p_{i,21} & p_{i,22} & p_{i,23} & p_{i,24} \\ p_{i,31} & p_{i,32} & p_{i,33} & p_{i,34} \end{bmatrix}, \qquad i = 1, 2. \tag{2}$$

By combining Eqs. (1) and (2), and eliminating the unknowns s_1 and s_2, we derive a linear system:

$$J \begin{bmatrix} x \\ y \\ z \end{bmatrix} = b,$$

$$J = \begin{bmatrix} p_{1,11} - u_1 p_{1,31} & p_{1,12} - u_1 p_{1,32} & p_{1,13} - u_1 p_{1,33} \\ p_{1,21} - v_1 p_{1,31} & p_{1,22} - v_1 p_{1,32} & p_{1,23} - v_1 p_{1,33} \\ p_{2,11} - u_2 p_{2,31} & p_{2,12} - u_2 p_{2,32} & p_{2,13} - u_2 p_{2,33} \\ p_{2,21} - v_2 p_{2,31} & p_{2,22} - v_2 p_{2,32} & p_{2,23} - v_2 p_{2,33} \end{bmatrix}$$

$$b = \begin{bmatrix} u_1 p_{1,34} - p_{1,14} \\ v_1 p_{1,34} - p_{1,24} \\ u_2 p_{2,34} - p_{2,14} \\ v_2 p_{2,34} - p_{2,24} \end{bmatrix}. \tag{3}$$

By solving Eq. (3), we obtain the unknown 3D eye coordinates as the following:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = (J^T J) \backslash (J^T b). \tag{4}$$

Then, we obtain the position of the eye on the thermal image (u_t, v_t) by projecting the recovered 3D eye location onto the thermal image based on the calibration parameters. The homogeneous coordinates of the eyes on the thermal image u_t, v_t can be expressed as

$$s_t \begin{bmatrix} u_t \\ v_t \\ 1 \end{bmatrix} = P_t \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \tag{5}$$

where s_t is a scalar and P_t is the projection matrix of the thermal camera, which can be also obtained through the calibration. We denote the eye landmarks of the thermal face as $(u_{t,l}, v_{t,l})$ and $(u_{t,r}, v_{t,r})$ for the left eye (LE) and right eye (RE), respectively.

C. Additional Face Landmark Detection

In this section, we locate three more face landmarks to resolve the 1D ambiguity of the head pose. Besides the two eye landmarks (LE and RE) on the thermal image, we detect the additional three landmarks. These landmarks are the nose tip (N), and the left and right corners of the mouth (LM and RM). We use a deep cascaded CNN to detect these landmarks on the thermal face images.

From the eye landmarks identified, our system locates a potential sufficiently large region that covers the face, where the size of the region is proportional to the target distance and the distance between the two eyes. We apply a cascade thermal face detector that is trained on our thermal face data set onto the potential region to tighten the face bounding box.

We designed our own additional thermal face landmark detector based on Sun *et al.*'s deep-cascaded CNN [19]. Figure 6 shows an overview of our modified model. Because of the weak discriminative features of the eyes in the LWIR images, we eliminated the part of the eye detection in Sun's model. We adjusted three input regions in the first level of the model, which cover the nose (N), the lower face (LF), and the mouth (M). Each deep structure in the first level is adjusted correspondingly since the sizes of the three input regions change. We eliminated the parts of the eye landmark refinement and correspondingly adjusted the size of the local search regions in the remaining levels to improve the efficiency and accuracy of our required landmarks detection, which is shown in Section 3.D. Since the precise face region can be predicted with the aid of the accurate eye locations, the cascade thermal facial landmark detector is used only in the face region. Therefore, we achieve accurate detection of the three additional facial landmarks.

D. Pose Estimation

In this section, we perform the pose estimation using five facial landmarks: LE, RE, N, LM, and RM. We adopt a standard 3D human head model and extract these corresponding five standard 3D facial points. Since we can obtain the calibrated LWIR camera's intrinsic and distortion parameters, we can estimate the relative pose of the head model to the thermal camera by solving a perspective-n-point (PnP) problem. We use the method in Ref. [22] to solve the PnP problem where the projection error is minimized:

$$\text{res} = \sum_i \text{dist}^2 \left(K_t [R_Q | T_Q] \begin{bmatrix} \boldsymbol{X}_i \\ 1 \end{bmatrix}, m_{t,i} \right), \tag{6}$$

where $\boldsymbol{X}_i, m_{t,i} = (u_{t,i}, v_{t,i}), i = 1, \ldots, 5$ are the 3D points in the thermal camera coordinate system and their corresponding distortion-corrected 2D image projections on the thermal image; $\text{dist}(\boldsymbol{a}, \boldsymbol{d})$ computes the 2D distance between points $\boldsymbol{a}$ and $\boldsymbol{d}$; and K_t is the pre-calibrated thermal camera intrinsic matrix. The rotation and translation matrices of the head pose R_Q, T_Q relative to the LWIR camera are estimated by minimizing Eq. (6).

The accuracy of pose estimation with a single 3D shape is based on localizing the facial features, which is demonstrated in Ref. [10]. Failures in facial landmark detection will lead to failures in most pose estimation and correction algorithms. It is important to note here that (1) our hybrid solution can provide high pose estimation and correction accuracies, since we have highly accurate facial feature localizations, especially eye localizations, and (2) our solution can still estimate a rewarding approximated pose using only the two derived 3D eye positions in the case of failures in the detection of other facial features, although this estimated pose may have error (ambiguity) in one dimension (the pitch of the head pose). However, the human head does not usually have a large pitch angle, looking either up or down.

When we have the two recovered 3D eye positions, we use only the two pairs of 3D-to-2D points to perform pose estimation as described in the following. The 3D RE and LE positions are noted as $X_r = (x_r, y_r, z_r)$ and $X_l = (x_l, y_l, z_l)$, respectively, in the world coordinate system, which is the first NIR camera coordinate system. First, we calculate a right vector $\vec{r} = R_t(X_l - X_r)$ and its normalized vector $\hat{r}$. Second, we calculate an upward vector $\vec{u} = \hat{r} \times (0, 1, 0)$ and a forward vector

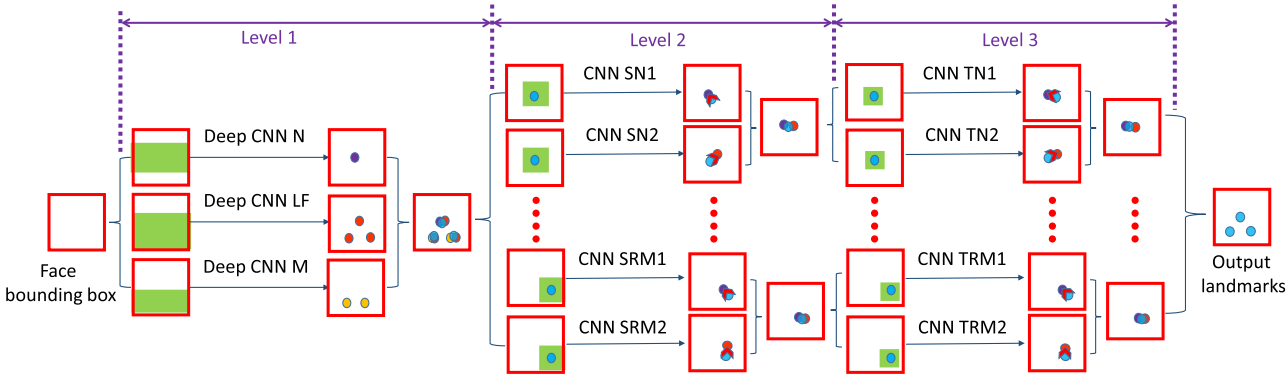


Fig. 6. Modified cascaded CNN for thermal facial landmarks (N, LM, and RM) detection based on [19].

$\vec{f} = \hat{r} \times \hat{u}$. Finally, we initialize R_Q as $[\hat{r}^T, \hat{u}^T, \hat{f}^T]$. After this, T_Q can be solved through Eq. (6).

E. Pose Correction and Face Frontalization

In this section, we perform our face frontalization after pose estimation by adopting Hassner *et al.*'s method in Ref. [10]. This is a warping process as described in the following. First, we project the head model into the 2D reference image I_r with n valued pixels using a projection matrix $C = K_t[R_q\, T_q]$, where R_q is the identity matrix so that the model is facing forward, and T_q is a translation vector adjustable according to the desired output size. For each pixel m_j, $j = 1, 2, \ldots, n$, of the reference view (frontalized face) shown in Fig. 7(c), we store the 3D point $X_j = (x_j, y_j, z_j)^T$ on the surface of the head model, which is projected to m_j, i.e.,

$$m_j \sim CX_j. \tag{7}$$

Second, we assign the thermal intensity value for each prestored 3D point X_j from its projection onto the original thermal image I_o [query image, such as the example shown in Fig. 7(a)]. Denoting the projection as m_j', we have

$$m_j' \sim K_t[R_Q|T_Q]X_j, \tag{8}$$

$$I_r(m_j) = I_o(m_j'). \tag{9}$$

The sampled thermal intensities $I_o(m_j'), j = 1, 2, \ldots, n$, from bi-linear interpolation assigned to $I_r(m_j)$ produce the initial frontalized result. Figure 7(d) shows the initial frontalized view.

Third, we apply the soft-symmetry processing as in Ref. [10] to refine the initial result. Those pixels corresponding to the 3D points that are poorly visible from the original thermal view are reassigned with the intensities of their symmetric correspondence on the other side of the face. We average the initial and refined output to obtain the final result. Figure 7(e) shows our final frontalized face image.

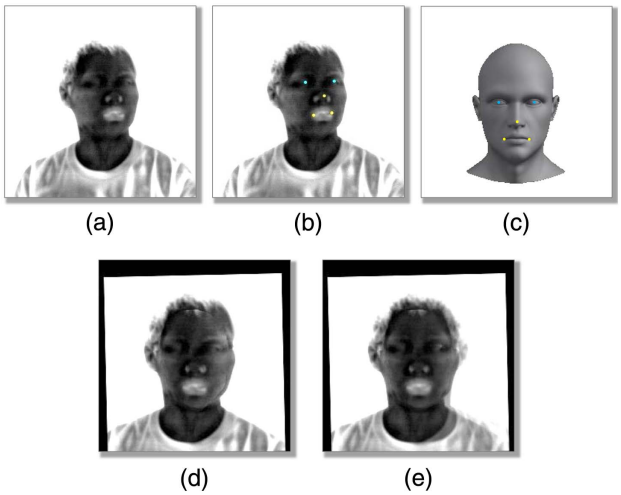


Fig. 7. Frontalization overview. (a) Query image. (b) Fused landmarks image. (c) Reference view with detected landmarks rendered from the 3D head model; each of its colored pixels has a corresponding 3D point coordinate on surface of the 3D model. (d) Use pose estimation and a specified reference projection matrix to back-project query intensities to the reference coordinate system. (e) Frontalized result with soft symmetry.

3. EXPERIMENTS

A. Hybrid Sensing Prototype System

Figure 8 shows the prototype of our hybrid sensing system. In the center of the system is the Tamarisk 640 thermal camera. We use two Flea2 monochrome cameras with NIR pass filters as our NIR cameras. Each NIR camera is surrounded with a Logisaf L002-48-94 IR board (940 nm), which is a component of the Logisaf CCTV camera, to simulate the on-axis IR light source, which is eye-lit IR lights. We also place two extra Logisaf L002-60-94 IR boards (940 nm) above the two NIR cameras as the off-axis light source, which is face-lit IR lights.

We have resolved two challenges in our hybrid system design. The first challenge is the synchronization of the IR illuminators, the pair of NIR cameras, and the LWIR camera for image acquisition. Synchronization between the pair of NIR cameras is achieved through the camera operation APIs provided by Point Grey. It is more challenging to synchronize the on-axis and off-axis IR light sources with the camera capturing. For this purpose, we designed an IR flash control system (shown on the left of Fig. 8) comprising one Phidget Interfacekit board and two relay boards; it not only synchronizes our IR light sources with the camera system but also makes these sources into flashes. Using NIR flashes makes our system more covert than using continue NIR illumination. The control system is programmed to first turn on the off-axis IR lights, simultaneously trigger the NIR cameras to capture the first frame (without bright eyes) and turn off these lights, and then turn on the on-axis IR lights, trigger the cameras for the second frame (with bright eyes), and turn off them. The working range of this prototype to acquire bright eyes can be found in Fig. 9.

The second challenge is the LWIR camera's calibration and the cross-modality cameras' calibration, i.e., the calibration between LWIR and NIR cameras. In the LWIR camera's calibration, the printed chessboard or other pattern is not visible due to its uniform temperature in the LWIR image. To efficiently calibrate the LWIR camera, we designed a white pattern mold [shown in Fig. 10(a)] with circular holes on the board. We kept the mold at normal temperature and used a black exothermic board behind the mold to produce a higher temperature than the temperature of the board. This produces a clear contrast between the circular holes on the mold and the pattern mold

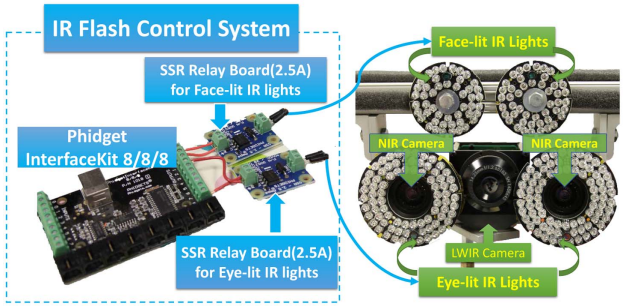


Fig. 8. Our hybrid sensing system prototype.

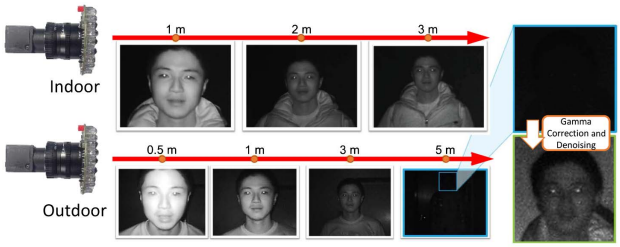


Fig. 9. Working range of our prototype.

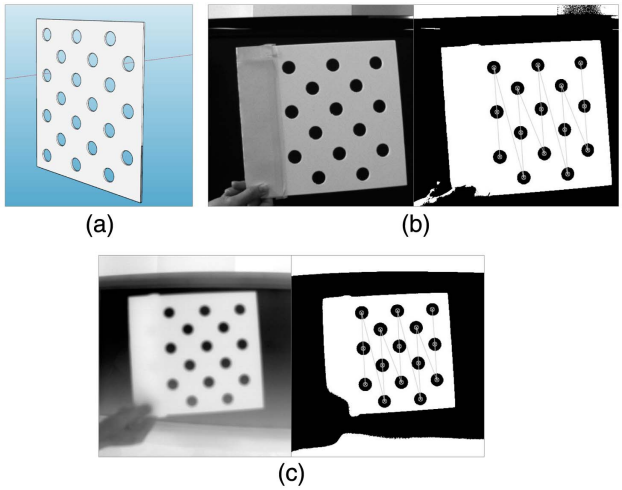


Fig. 10. Cross-modality camera calibration. (a) Calibration mold. (b) NIR image and its thresholded image. (c) LWIR image and its thresholded image.

Table 1. Intrinsic Calibration Result

Parameters	NIR Camera_1	NIR Camera_2	Thermal Camera
c_x	311.2695	302.0133	322.0687
c_y	258.8132	276.4582	246.3135
f_x	1046.5966	1055.7334	841.6957
f_y	1047.2157	1056.2581	840.2238
k_1	−0.1241	−0.2270	−0.3813
k_2	−3.4701	1.4192	−0.5758
p_1	0.0018	0.0024	−0.0026
p_2	−0.0012	0.0004	−0.0009
k_3	−998.1297	572.7115	−70.6259
k_6	−1046.1720	595.1688	−79.4948
RMS	0.2017	0.1810	0.3015

in the LWIR image [shown in Fig. 10(c)]. After extracting centers of these circles in the pattern, we used a calibration code [23] for the LWIR camera calibration. Because of the color contrast between the black board and the white mold, it is also easy to detect the centers of the circles on the captured images from the Flea2 cameras. Figure 10(b) shows an example. Therefore, this specially designed mold solves the cross-modality calibration problem.

B. Calibration Results

Table 1 shows the intrinsic calibration results for the pair of NIR cameras and the thermal camera, K_1, K_2, and K_t, respectively. Parameters (f_x, f_y) represent the camera focal lengths, (c_x, c_y) represent the optical centers expressed in pixels coordinates, $k_1, k_2, \ldots, k_6$ represent the radial distortion coefficients (k_4, k_5 are fixed to 1 in calibration), and p_1 and p_2 are the tangential distortion coefficients. RMS in the last row of Table 1 illustrates the root-mean-squared distances in pixels between detected image points and projected ones. Table 2 shows the extrinsic calibration results, R_2, T_2, R_t, and T_t, and the reprojection error for different stereo pairs of cameras. This error is also calculated by RMS for all points in all the available views from each stereo pair. The small calibration errors, both in the intrinsic and extrinsic calibration result tables, indicate that our calibration method with the designed tool is accurate and reliable.

C. Data Set Collection

We acquired our data set using our proposed hybrid sensing system. The data set contains 10 subjects. For each person, we captured two sets, one with glasses and another without glasses. For each set, we acquired thermal face images having 11 horizontal pose variation angles (−75°, −60°, −45°, −30°, −15°, 0°, 15°, 30°, 45°, 60°, 75°), and seven vertical pose variation angles (−60°, −40°, −20°, 0°, 20°, 40°, 60°). We also collected additional 23 random pose images with random roll rotations. Therefore, we have 200 pose images for each person. In total, there are 2000 images in the whole data set. We labeled each image with the positions of three facial landmarks (N, LM, and RM). We randomly selected 65% of the data set for training of our facial landmarks detector and 10% for validation. The remaining images are used for testing. Example training data are shown in Fig. 11. Our collected data are also used for the training of our thermal face detector.

D. Face Detection and Registration Results

1. Results of Landmark Detection

In this section, we compared the performance of landmark detection using the method in Ref. [18] with our method. The training input to the learning-based method in Ref. [18] requires labeled eye landmarks. However, it is very difficult to manually label the eye landmarks in the thermal images due to lack of information around the eyes. This can be seen in Fig. 14(a), where the eye region of the thermal image is compared with the eye region of the NIR image.

We used our projected eye locations in the thermal images as the eye landmarks, together with three labeled facial landmarks (N, LM, and RM) to train Sun's structure in

Table 2. Extrinsic Calibration Result

Pairs	Rotation Vector	Translation Vector	Reprojection Error
NIR_Cam_1 to NIR_Cam_2	(−0.01237; 0.00816; −0.01713)	(113.34144; 0.42439; 2.83881)	0.19021
NIR_Cam_1 to LWIR_Cam	(−0.00629; −0.00042; −0.00852)	(55.06529; −0.79071; −5.88902)	0.28276

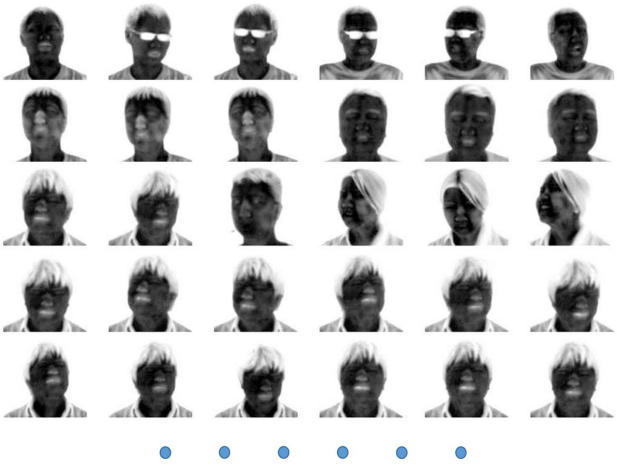


Fig. 11. Sample training data for the facial landmark detector.

Ref. [19]. This is the best scenario of training Sun's structure for all five thermal facial landmarks detection. For each testing data, these five "labeled" facial landmarks are used as the ground truth to compare with the detected landmark results.

We used the same performance measurements as those in Ref. [19] to calculate the average detection error and the failure rate for each facial point. These performance measurements indicate the accuracy and reliability of the algorithm. The detection error is measured as

$$\text{err} = \sqrt{(u_t - u_t')^2 + (v_t - v_t')^2}/l, \tag{10}$$

where (u_t, v_t) and (u_t', v_t') are the ground-truth position and the detected position, respectively, and l is the width of the face bounding box returned by the thermal face detector. A failure is counted if an error is larger than 5%.

Figure 12 summarizes the results of the comparison of landmark detection of the learning-based method in Ref. [19] and our method on the testing data. Our method achieved higher accuracy not only on thermal eye detection, but also on other thermal facial landmark (N, LM, and RM) detection. More than 9.5%, 18.9%, and 11.5% relative accuracy improvements on average errors are achieved for N, LM, and RM, respectively, by our method. Our thermal eye detection has no failure, or training and testing error. For the learning-based method, the lack of discriminative information around the eyes (low contrast) in the thermal image causes inaccurate eye detection and less accurate detection of the other three facial landmarks.

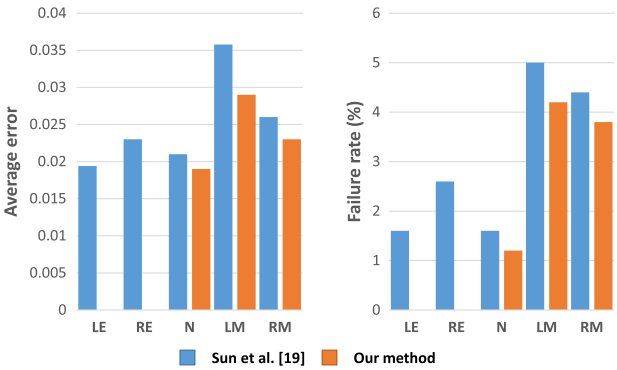


Fig. 12. Average detection errors and failure rates of the Sun structure [19] and ours on our testing data.

Compared with the learning-based method [19] that uses only thermal images, there are two advantages to our method for achieving more accurate facial landmark detection. First, our method locates the eye landmarks in the thermal images more accurately from the 3D eye positions obtained by two NIR images in our hybrid sensing system. Second, our method predicts the face bounding box more precisely based on the information of the thermal eye positions and actual distance between the eyes in 3D. This improves the detection of the other three facial landmarks (N, LM, and RM) in the thermal images.

Figure 13 shows an example of the comparison between our face landmark detection and the one in Ref. [19]. Figure 13(a) shows the result of our fused eye landmarks and three more facial landmark detections. Figure 13(b) shows all five facial landmark detection results using [19]. Figure 13(c) shows the eye landmark location comparison.

2. Detection Results with Eyeglasses

We also tested our eye landmark detection method on images with eyeglasses; examples are shown in the first row of Fig. 15(c). This is another advantage of our method over the learning-based method for eye detection on a thermal image. Our system uses extra NIR images, which are more informative in cases such as targets wearing eyeglasses. From Fig. 14(b), we can see eyeglasses totally block the eyes on the thermal image. In contrast, they have little effect on NIR images on which our eye localization is performed.

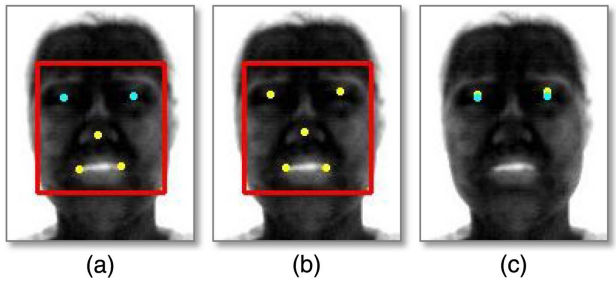


Fig. 13. Comparison between CNN eye detection [19] (yellow points) and our method (cyan points). (a) Our fused result. (b) Full landmark detection result from [19]. (c) The comparison.

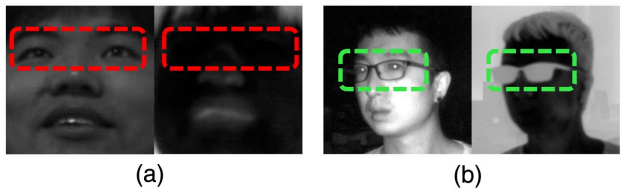
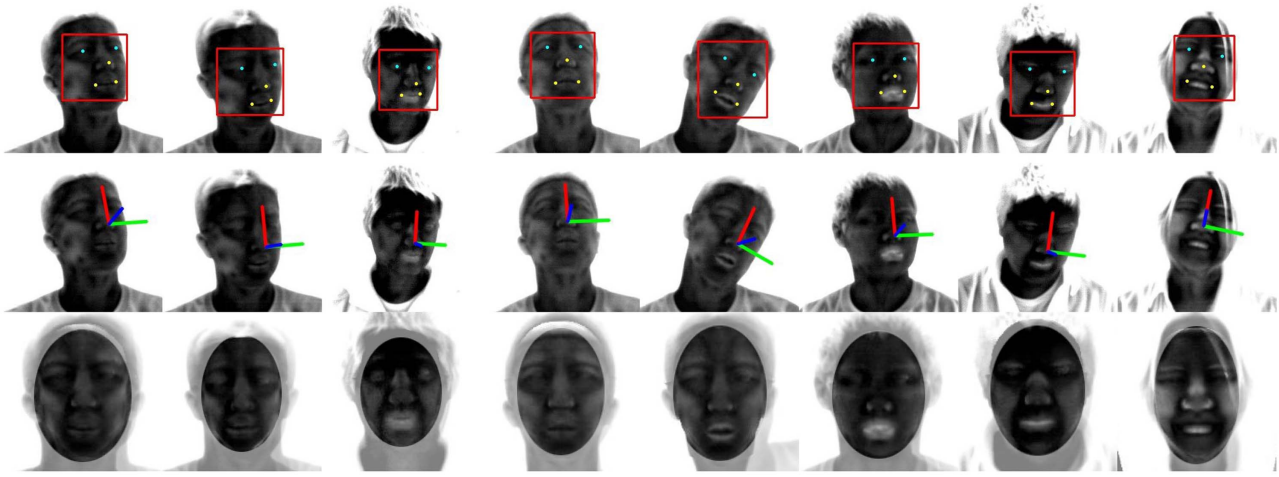
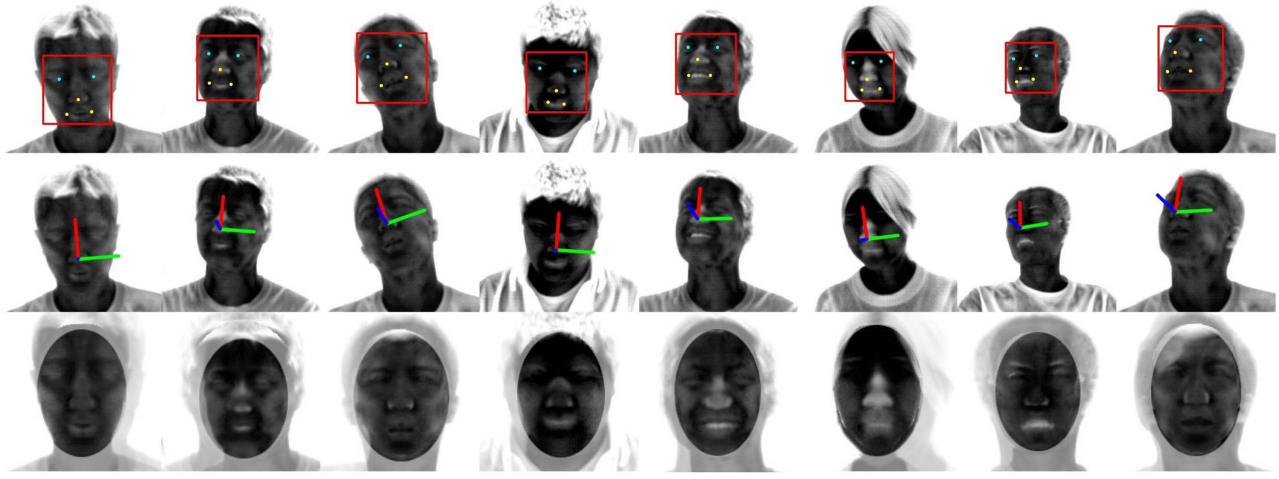


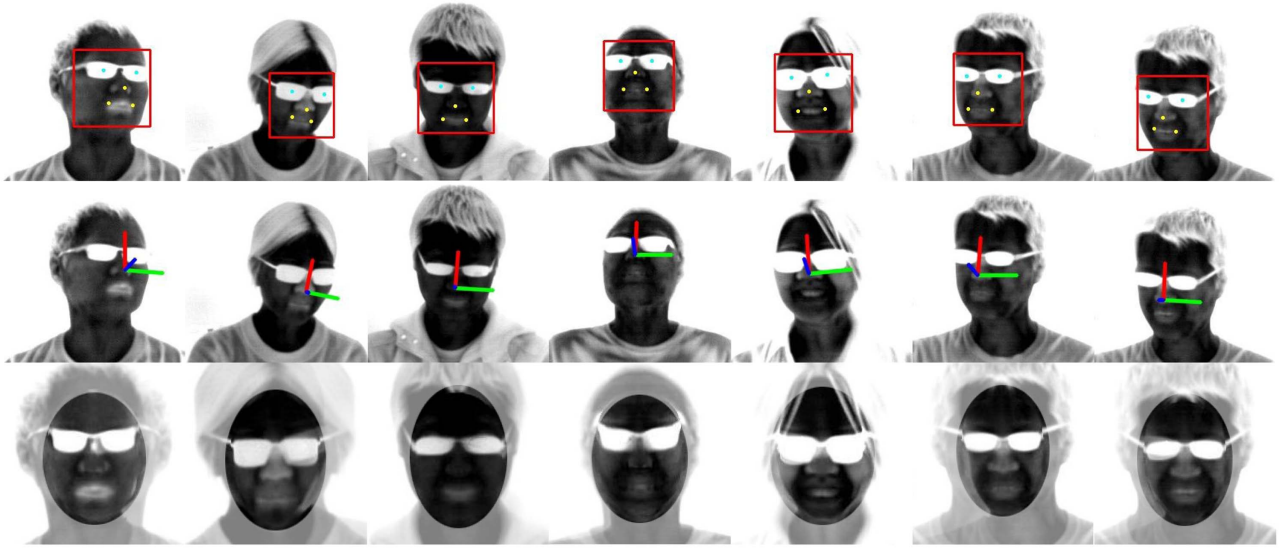
Fig. 14. Comparison between the NIR image and thermal image. (a) Eye region comparison. (b) Comparison with the effect of eyeglasses.



(a) Experimental results with horizontal rotation angles between 0° and 60° and vertical rotation angles between −45° and 45°.



(b) Experimental results with horizontal rotation angles between −60° and 0° and vertical rotation angles between −45° and 45°.



(c) Experimental results with eyeglasses appearances.

Fig. 15. Experimental results. For each subfigure, the first row shows the landmark detection results, the second row shows the pose estimation results, and the third row shows the pose correction results.

3. Thermal Face Image Pose Correction and Frontalization Results

Figure 15 presents examples of several scenarios. Figure 15(a) shows the results of face images with horizontal rotation angles between 0° and 60° and vertical rotation angles between −45° and 45°. The first row shows the landmark detection results. The second row shows the pose estimation results. The third row shows the pose correction results. Figure 15(b) shows the results of face images with horizontal rotation angles between −60° and 0° and vertical rotation angles between −45° and 45°. Figure 15(c) shows the results of face images with eyeglasses. These results demonstrate that our method is effective and robust in head pose estimation.

4. DISCUSSION AND CONCLUSION

In this paper, we present a new collaborative, multi-spectrum sensing system by combining computational imaging and illumination to achieve face detection and registration for low-light personnel recognition under uncooperative conditions. Our method uses the special bright-eye effect of human eyes to assist in 3D eye localization. Using the 3D eye position, we can efficiently propose tight and compact face region candidates for additional face landmark detection in the thermal image. This will lead to the development of a new class of face registration and face recognition algorithms with 2D + 3D concepts.

Our active hardware thermal eye localization subsystem is more accurate, robust, and fast. It also allows for fast face/head movement, which is common for unconstrained conditions. The proposed hybrid sensor, which runs only on-demand flash control in which the flash is turned on for eye locating only when potential targets are detected from a thermal camera, also makes covert operations closer to attainable and reduces power consumption, which is critical in outdoor operations. Experiments illustrate that our method is robust, accurate, and efficient in face detection and registration under low-light conditions.

In the future, we can further optimize our hybrid sensing system for face detection and recognition. First, we will extend the NIR flashing working range by increasing the intensity of our NIR illuminators. Second, we will increase the amount of training data to train a better facial landmark detector. Third, we will further improve pose estimation accuracy by using more non-coplanar landmarks, for instance, to include the boundary of the chin. Fourth, we will conduct thermal face recognition. With global context detection, local facial feature detection, and pose correction from our system, we believe that the recognition will be accurate and robust. Fifth, we will compensate for the lack of facial details on the thermal image with NIR facial details based on the calibration of 3D triangulation.

Funding. Army Research Office (ARO) (W911NF14-1-0338).

REFERENCES

1. L. Wolf and A. Shashua, "Learning over sets using kernel principal angles," J. Mach. Learn. Res. **4**, 913–931 (2003).
2. M. Nishiyama and O. Yamaguchi, "Face recognition using the classified appearance-based quotient image," in *7th International Conference on Automatic Face and Gesture Recognition* (IEEE, 2006), p. 6.
3. B.-G. Park, K.-M. Lee, and S.-U. Lee, "Face recognition using face-ARG matching," IEEE Trans. Pattern Anal. Mach. Intell. **27**, 1982–1988 (2005).
4. A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Expression-invariant 3D face recognition," in *Audio-and Video-Based Biometric Person Authentication* (Springer, 2003), pp. 62–70.
5. L. Wiskott and C. Von Der Malsburg, "Recognizing faces by dynamic link matching," Neuroimage **4**, S14–S18 (1996).
6. R. S. Ghiass, O. Arandjelović, A. Bendada, and X. Maldague, "Infrared face recognition: a comprehensive review of methodologies and databases," Pattern Recogn. **47**, 2807–2824 (2014).
7. Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 1701–1708.
8. T. Hassner, "Viewing real-world faces in 3D," in *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 3607–3614.
9. Z. Zhu, P. Luo, X. Wang, and X. Tang, "Recover canonical-view faces in the wild with deep neural networks," arXiv: 1404.3543 (2014).
10. T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," arXiv: 1411.7964 (2014).
11. V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel, "Exchanging faces in images," in *Computer Graphics Forum* (Wiley, 2004), Vol. **23**, pp. 669–676.
12. T. Bourlai, J. Von Dollen, N. Mavridis, and C. Kolanko, "Evaluating the efficiency of a night-time, middle-range infrared sensor for applications in human detection and recognition," Proc. SPIE **8355**, 83551B (2012).
13. J. Dowdall, I. Pavlidis, and G. Bebis, "Face detection in the near-IR spectrum," Image Vis. Comput. **21**, 565–578 (2003).
14. B. Klare and A. K. Jain, "Heterogeneous face recognition: matching NIR to visible light images," in *20th International Conference on Pattern Recognition (ICPR)* (IEEE, 2010), pp. 1513–1516.
15. L. Liang, R. Xiao, F. Wen, and J. Sun, "Face alignment via component-based discriminative search," in *European Conference on Computer Vision* (Springer, 2008), pp. 72–85.
16. M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2010), pp. 2729–2736.
17. X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2012), pp. 2879–2886.
18. P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," IEEE Trans. Pattern Anal. Mach. Intell. **35**, 2930–2940 (2013).
19. Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2013), pp. 3476–3483.
20. T. E. Hutchinson, "Eye movement detector with improved calibration and speed," U.S. patent 4,950,069 (21 August 1990).
21. J. Van de Kraats and D. van Norren, "Directional and nondirectional spectral reflection from the human fovea," J. Biomed. Opt. **13**, 024010 (2008).
22. V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: an accurate O(n) solution to the PnP problem," Int. J. Comput. Vis. **81**, 155–166 (2009).
23. Z. Zhang, "A flexible new technique for camera calibration," IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1330–1334 (2000).