Face Tracking and Recognition at a Distance: A Coaxial & Concentric PTZ Camera System

Hyun-Cheol Choi, Unsang Park, *Member, IEEE*, Anil K. Jain, *Fellow, IEEE* and Seong-Whan Lee, *Fellow, IEEE*

***Abstract*—Face recognition has been regarded as an effective method of subject identification at a distance because of its covert and remote sensing capability. However, face images have a low resolution when they are captured at a distance (say, larger than 5 meters) thereby degrading the face matching performance. To address this problem, we propose an imaging system consisting of static and PTZ cameras to acquire high resolution face images up to a distance of 12 meters. We propose a novel Coaxial-Concentric camera configuration between the static and PTZ cameras to achieve the distance invariance property using a simple calibration scheme. We also use a linear prediction model and camera motion control to mitigate delays in image processing and mechanical camera motion. Our imaging system was used to track 50 different subjects and their faces at distances ranging from 6 to 12 meters. The matching scenario consisted of these 50 subjects as probe and additional 10,000 subjects as gallery. Rank-1 identification accuracy of 91.5% was achieved compared to 0% rank-1 accuracy of the conventional camera system using a state-of-the-art matcher. The proposed camera system can operate at a larger distance (up to 50 meters) by replacing the static camera with a PTZ camera to detect a subject at a larger distance and control the second PTZ camera to capture the high resolution face image.**

***Index Terms*—Face recognition at a distance, PTZ camera, coaxial, concentric, tracking.**

I. Introduction

FACE recognition in surveillance environments is crucial to identify potential terrorists and criminals on a watch list. While the performance of face recognition has improved substantially in the last decade [2], [3], the intrinsic (expression, aging, etc.) and extrinsic (illumination, pose, etc.) variations are still the major bottlenecks in face recognition. Face recognition at a distance of over 5 meters introduces another challenge, namely the low image resolution problem. Typical commercial face recognition engines require face images with at least 60 pixels between the eyes (called inter-pupillary distance) for successful recognition, which is difficult to achieve in many surveillance systems. Fig. 1 shows degradations in image resolution as the standoff between the camera and subject increases.

H.-C. Choi is with the Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea (hcchoi@korea.ac.kr).

U. Park is with the Department of Computer Science and Engineering, Michigan State University, E. Lansing, MI 48824, USA (parkunsa@cse.msu.edu).

A. K. Jain is with the Department of Computer Science and Engineering, Michigan State University, E. Lansing, MI 48824, USA and with the Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea (jain@cse.msu.edu).

S.-W. Lee is with the Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea (swlee@image.korea.ac.kr).

An earlier version of this work appeared in [1].

Existing approaches that have studied face recognition at a distance can be essentially categorized into two groups: (i) generating a super resolution face image from the given low resolution image and (ii) acquiring high resolution face image using a special camera system (e.g., a high resolution camera or a PTZ camera). While reconstructing a high resolution face image from its low resolution counterpart can improve image quality and help the face recognition process, the performance of this approach highly depends on the training data. High-resolution cameras can potentially overcome the low resolution problem, but either they expect the subject to be at a fixed location/distance or the camera has to be manually focused on the subject. The above mentioned limitations have lead to the extensive use of Pan-Tilt-Zoom (PTZ) cameras, since PTZ cameras provide an inexpensive way to automatically track and obtain close-up face images of subjects of interest. However, the field of view of PTZ cameras is severely limited when it zooms into an object. Therefore, systems with paired static and PTZ cameras have emerged as a promising method to achieve tracking and zooming capability for wide surveillance areas; the static camera provides the wide field of view and then directs the PTZ camera to obtain high resolution images of target objects. The main challenge faced by such a system arises in registering the image coordinates of static camera and the pan and tilt angles of the PTZ camera. Due to the lack of depth information, the image coordinates of the static camera are not in one to one correspondence with pan and tilt angles of the PTZ camera. A direct estimation of the depth using a 3D sensor or stereography method could be a possible solution, but they are either too expensive or not sufficiently accurate.

Dedeoglu et al. [33] recognized faces in low resolution

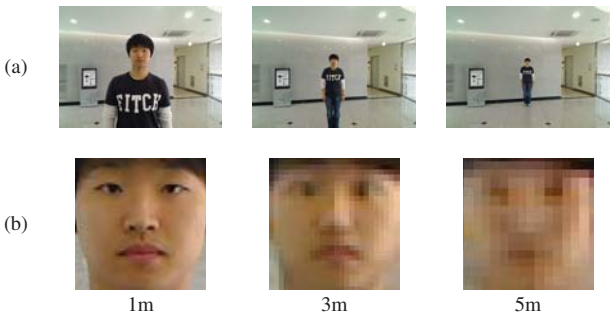


Fig. 1. Images at three different distances (1∼5m): (a) images captured by a webcam (Logitech, Pro9000, image size of 640×480) and (b) face images cropped and resized. The inter-pupillary distances (IPDs) are 35, 12, and 7 pixels from left to right, respectively.

TABLE I
A COMPARISON OF SURVEILLANCE SYSTEMS THAT USE PTZ CAMERA

Surveillance system	#Static cameras	#PTZ cameras	Tracking: single (s) or multiple (m) persons	Prediction	Speed control	Operating area (meters)	Face recognition (accuracy, #images in probe, #images in gallery, #subjects in probe, #subjects in gallery)
Bernardin et al. (2007) [4]	0	1	m	No	Yes	indoor (5 m)	No
Mian et al. (2008) [5]	0	1	s	No	No	indoor (N/A)	No
Yang et al. (2008) [6]	0	1	s	No	No	indoor (7.8 m)	No
Kumar et al. (2009) [7]	0	1	s	No	No	outdoor (N/A)	No
Varcheie et al. (2009) [8]	0	1	s	Yes	Yes	indoor (N/A)	No
Venugopalan et al. (2010) [9]	-	1	s	Yes	Yes	indoor (0.6∼1.5 m)	No (Iris recognition)
Varcheie et al. (2011) [10]	0	1	m	Yes	No	indoor (12 m)	No
Everts et al. (2007) [11]	0	2	s	No	No	indoor (N/A)	No
Liao et al. (2009) [12]	0	2	m	No	No	outdoor (N/A)	No
Del et al. (2010) [13]	0	2	m	Yes	No	outdoor (80 m)	No
Liao et al. (2010) [14]	0	2	m	Yes	No	N/A	No
Wheeler et al. (2010) [15]	0	2	m	Yes	No	outdoor (15∼20 m)	(N/A, 30, 268, 9, 244)
Zhou et al. (2010) [16]	0	2	s	No	No	in/outdoor (30∼100 m)	No
Bodor et al. (2004) [17]	1	1	s	No	No	in/outdoor (N/A)	No
Marchesotti et al. (2005) [18]	1	1	s	No	No	outdoor (N/A)	No
Funahasahi et al. (2004) [19]	1	1	s	No	No	indoor (1.5 m)	No
Yoon et al. (2009) [20]	1	1	s	No	No	indoor (1.5∼2.5 m)	No
Amnuaykanjanasin et al. (2005) [21]	1	1	s	No	No	outdoor (5 m)	No
Prince et al. (2006) [22]	1	1	m	No	No	indoor (N/A)	(100%, 100, 220, 100, 220)
Chen et al. (2008) [23]	1 (Omni)	1	m	No	No	indoor (5 m)	No
Scotti et al. (2005) [24]	1 (Omni)	1	m	No	No	in/outdoor (N/A)	No
Tarhan et al. (2011) [25]	1 (Omni)	1	s	No	No	indoor (6 m)	No
Lu et al. (2008) [26]	1	1	m	No	No	indoor (N/A)	No
Yao et al. (2009) [27]	1	1	s	Yes	Yes	indoor (15 m)	No
Sivaram et al. (2009) [28]	1	2	s	No	No	indoor (6 m)	No
Xu et al. (2010) [29]	1	2	m	Yes	Yes	outdoor (80 m)	No
Stillman et al. (1999) [30]	2	2	m	No	No	indoor (N/A)	No
Hampapur et al. (2003) [31]	2	2	s	No	No	indoor (6.25 m)	No
Krahnstoever et al. (2008) [32]	4	4	m	Yes	No	outdoor (N/A)	No
Proposed method	1 or 2	1	s & m	Yes	Yes	indoor (12 m)	single person: (91.5%, 102978, 10150, 50, 10050) multi-person: (93.4%, 36574, 10009, 3, 10003)

images using the super-resolution method. Park et al. [34] proposed a stepwise reconstruction of a high-resolution facial image based on the extended morphable face model. The performances of their systems [33] [34] is highly dependent on the training data and the recognition accuracy rapidly drops when the image resolution is less than 16×16. Yao et al. [35] used a high magnification static camera to capture face images at long distances (50∼300 m). However, the camera does not provide pan and tilt motion, resulting in a very small field of view. Bernardin et al. [4] proposed an automatic system for the monitoring of indoor environments using a single PTZ camera. However, their system requires frontal pose in every frame to properly control the PTZ camera and the system has to zoom out when it fails to detect the face. Scotti et al. [24] and Chen et al. [23] used an omnidirectional camera for the monitoring of wide area. Everts et al. [11] and Liao et al. [12] used PTZ cameras to monitor wide areas in zoomed-out mode and used them to zoom-in and capture high resolution images whenever possible. Marchesotti et al. [18] used a pair of static and PTZ cameras to capture high resolution face images. Hampapur et al. [31] used multiple static cameras and a PTZ camera to accurately estimate the 3D world coordinates of a subject's face and then zoom into the face to capture a high resolution image. Stillman [30] used

multiple static cameras to estimate the location of a person in a calibrated scene, where the PTZ camera tracks the detected face. Most of these systems rely on the reconstruction of 3D world coordinates or a crude approximation of the calibration between static and PTZ cameras. The 3D world coordinate estimation is computationally expensive and is not suitable for real time applications. Table I summarizes most of the available approaches to recognize a face at a distance using PTZ camera(s). These methods can be categorized in terms of the number of static and PTZ cameras as below.

- Single PTZ camera: face location is first estimated in the zoomed-out view and the camera is controlled to acquire a high resolution face image. However, the single PTZ camera needs to continuously zoom in and out, so it is very easy to loose track of moving subjects.
- Single static camera and PTZ camera(s): The face location is estimated in the static view and the PTZ camera is controlled to capture a high resolution face image [22], [17], [19], [24], [21], [22], [36], [23], [26], [29]. However, due to the lack of depth information (Z coordinate), it is difficult to accurately estimate the (p, t) values in the static image. So, most of the automatic tracking systems using PTZ cameras provide a limited operating range and do not capitalize on the zooming feature of the PTZ camera. The main challenge faced by such a system is the camera calibration; image coordinates of static camera are calibrated to obtain the pan and tilt angle values of the PTZ camera.
- Dual (multiple) static cameras and PTZ camera(s) [31], [32], [30]: multiple static views allow stereographic reconstruction to estimate the 3D world coordinates. However, the stereographic reconstruction is computationally expensive and has a limited operating range. Multiple static cameras are utilized primarily to increase the surveillance coverage, while multiple PTZ cameras are considered to track multiple subjects concurrently.
- Single static high resolution camera [35]: by using a telescope attached to the camera, face image can be acquired at long distances (indoor: 10$\sim$16 m and outdoor: 50$\sim$300 m), but the field of view is severely limited. By using a high definition video camera, the field of view is increased, but the operating distance becomes smaller compared to the system using PTZ cameras.

Systems using static and PTZ cameras require a camera calibration process to correlate the world coordinates, image coordinates of static cameras, and parameters that control the PTZ cameras. To facilitate this calibration process, we propose a Coaxial-Concentric camera system that uses PTZ and static cameras with a relative camera calibration scheme between the image coordinate of static camera, (x_s^i, y_s^i), and PTZ camera parameters, (p, t, z). Compared to other camera systems proposed in the literature, our approach has the following advantages: (i) calibration process does not involve the world coordinates, (ii) only one relative calibration process is required and the calibrated system can be easily deployed at a different location with no recalibration, (iii) face images can be captured irrespective of the distance between the camera and subject, and (iv) by predicting subject's location and a camera speed control scheme, we obtain a smooth PTZ camera control capability.

The Coaxial-Concentric camera system developed by us was evaluated in a face recognition test with 50 probe subjects and 10,050 gallery subjects. The probe images were captured at distances ranging from 6 to 12 m whereas the gallery subjects are typical mug shots captured at a distance of about 1 m. A rank-1 identification accuracy of 91.5% was obtained in case of single person tracking. For multi-person tracking in four different scenarios with 3 subjects, a rank-1 accuracy of 93.4% was obtained.

II. Camera Calibration

A. Problem Formulation

We first define the variables used to describe the proposed camera system.

- $\mathbf{w_{obj}}$= $(x_{obj}, y_{obj}, z_{obj})$: target (face) location in the real world coordinate system
- $\mathbf{w_{calib}}$= $(x_{calib}, y_{calib}, z_{calib})$: real world coordinate at calibration distance (z_{calib}) corresponding to $\mathbf{w_{obj}}$
- $\mathbf{m_s}$= (x_s^i, y_s^i): image coordinate of the i^{th} static camera
- $\theta_{\mathbf{ptz}}$= (p, t, z): pan, tilt, and zoom parameters to control the PTZ camera; $\theta_{\mathbf{pt}}$ represents (p, t)
- $\mathbf{d}$= (d_x, d_y, d_z): displacement vector from the focal point of the static camera to the center of rotation of the PTZ camera

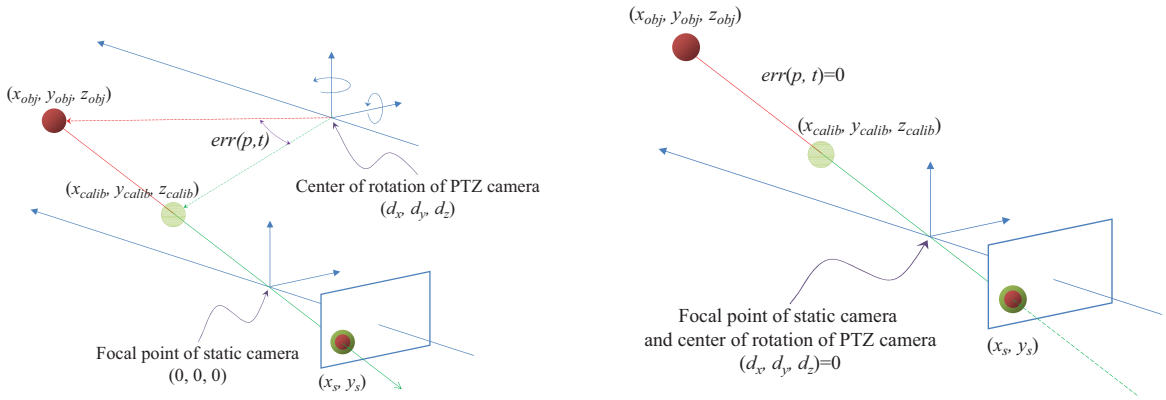
Our objective is to drive the PTZ camera via the $\theta_{\mathbf{ptz}}$ parameters towards the face location $\mathbf{w_{obj}}$ to capture a high resolution face image (inter-pupillary distance greater than 60 pixels). To determine the desired $\theta_{\mathbf{ptz}}$, we can either try to directly estimate $\mathbf{w_{obj}}$ or use the relationship between $\mathbf{m_s}$ and $\theta_{\mathbf{ptz}}$.

B. Coaxial-Concentric Camera Calibration

The conventional camera calibration process typically refers to establishing the relationship between the world coordinate and static image coordinate systems [37], [38]. The calibration process in PTZ camera systems for the high resolution face image acquisition involves calculating the relationship between the world coordinate and $\theta_{\mathbf{ptz}}$ parameters via the image coordinates of static camera; the calibration between the world coordinate and static image coordinate is not needed. Therefore, the calibration process involves calculating the mapping function from $\mathbf{m_s}$ to $\theta_{\mathbf{pt}}$. The zoom (z) parameter is obtained based on the estimated object (face) size (see Sec. II-D). The mapping function F can be calculated by a linear equation using a set of corresponding ground truth values of $\theta_{\mathbf{pt}}$ and $\mathbf{m_s}$ as:

$$\begin{bmatrix} p \\ t \end{bmatrix} = F \begin{bmatrix} x_s \\ y_s \\ 1 \end{bmatrix} = \begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \end{bmatrix} \begin{bmatrix} x_s \\ y_s \\ 1 \end{bmatrix} \quad (1)$$

We find a set of corresponding point pairs between $\mathbf{m_s}$ and $\theta_{\mathbf{pt}}$ by manually driving the PTZ camera to a number



(a) Non-Coaxial-Concentric configuration (b) Coaxial-Concentric configuration

Fig. 2. Schematics of (a) Non-Coaxial-Concentric and (b) Coaxial-Concentric camera systems. Targets (faces) at two different locations being projected to the same spot on the image plane shares the same pan and title angles for the PTZ camera control in the proposed Coaxial-Concentric camera system.

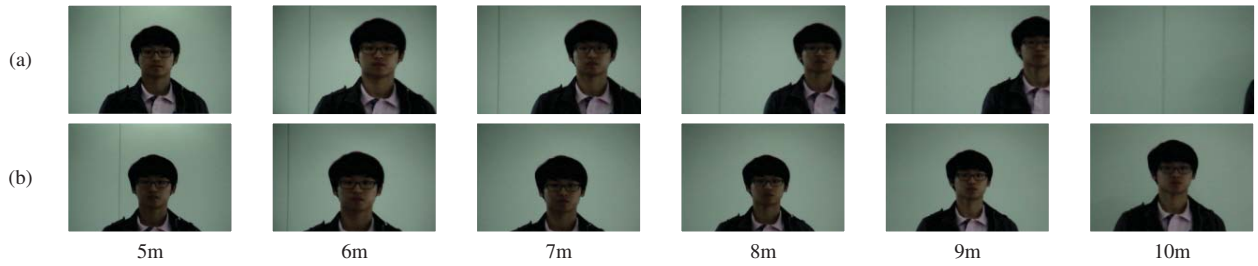


Fig. 3. Facial images at a distance of 5 to 10 m. The PTZ camera was controlled by a static camera (a) in Non-Coaxial-Concentric configuration and (b) in the proposed Coaxial-Concentric configuration ($\|\mathbf{w_{calib}}\| = 5m$). The IPDs are ~60 pixels for images in row (b). Note that the proposed system keeps the target face in the center of the image frame.

of different positions (15 in our case) in the static view. Even though a non-linear mapping function gave smaller residual error in our experiments, we chose to use the linear method for computational efficiency.

Fig. 2(a) shows that the world coordinates of a target $\mathbf{w_{obj}}$ appearing at two different locations correspond to the same image coordinate in the static view $\mathbf{m_s}$. As a result, the desired $\theta_{\mathbf{pt}}$ values obtained from the image coordinates of the static view may not always give the correct pan and tilt values to accurately capture the image of an object in the PTZ view. The error between the desired and calibrated $\theta_{\mathbf{pt}}$ values is defined as follows:

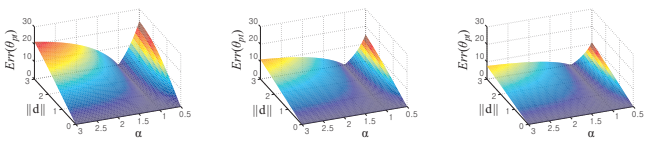
$$\begin{aligned} Err\left(\theta_{\mathbf{pt}}\right) &= \cos^{-1}\left(\frac{(\mathbf{w_{obj}}-\mathbf{d})\cdot(\mathbf{w_{calib}}-\mathbf{d})}{\|\mathbf{w_{obj}}-\mathbf{d}\|\,\|\mathbf{w_{calib}}-\mathbf{d}\|}\right) \\ &= \cos^{-1}\left(\frac{(\alpha\mathbf{w_{calib}}-\mathbf{d})\cdot(\mathbf{w_{calib}}-\mathbf{d})}{\|\alpha\mathbf{w_{calib}}-\mathbf{d}\|\,\|\mathbf{w_{calib}}-\mathbf{d}\|}\right) \\ &\because \mathbf{w_{obj}} = \alpha\mathbf{w_{calib}} \end{aligned} \tag{2}$$

The condition to achieve the minimum error can be derived as:

$$\begin{aligned} & Err\left(\theta_{\mathbf{pt}}\right) = 0 \\ \Leftrightarrow & \left(\frac{(\alpha\mathbf{w_{calib}}-\mathbf{d})\cdot(\mathbf{w_{calib}}-\mathbf{d})}{\|\alpha\mathbf{w_{calib}}-\mathbf{d}\|\,\|\mathbf{w_{calib}}-\mathbf{d}\|}\right) = 1 \\ \Leftrightarrow & (\alpha-1)^2\left\{(\mathbf{w_{calib}}\cdot\mathbf{d})^2 - \|\mathbf{w_{calib}}\|^2\|\mathbf{d}\|^2\right\} = 0 \\ \Leftrightarrow & (\alpha-1)^2\|\mathbf{w_{calib}}\|^2\|\mathbf{d}\|^2\left(\cos^2\theta_{(\mathbf{w_{calib}},\mathbf{d})}-1\right) = 0 \\ \Leftrightarrow & \alpha = 1 \; or \; \|\mathbf{d}\| = 0 \\ & (\because \|\mathbf{w_{calib}}\| > 0 \; and \\ & \quad \cos\theta_{(\mathbf{w_{calib}},\mathbf{d})} \neq 1 \; unless \; \mathbf{w_{calib}}\times\mathbf{d} = 0) \end{aligned} \tag{3}$$

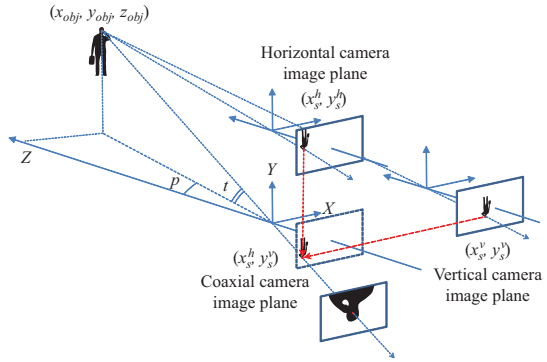
where $\theta_{(\mathbf{w_{calib}},\mathbf{d})}$ is the angle between $\mathbf{w_{calib}}$ and $\mathbf{d}$. The expanded derivation of Eq. (3) is provided in the Appendix.

In order to minimize the error in $\theta_{\mathbf{pt}}$, at least one of the following conditions must be satisfied: (1) the object must be observed at the calibrated distance ($\alpha = 1$) or (2) the focal point of static camera and center of rotation of PTZ camera coincide ($\|\mathbf{d}\| = 0$). In case an object is located at farther

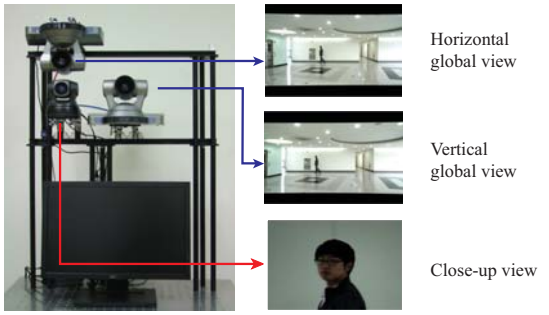


(a) $\|\mathbf{w_{calib}}\| = 5m$ (b) $\|\mathbf{w_{calib}}\| = 10m$ (c) $\|\mathbf{w_{calib}}\| = 15m$

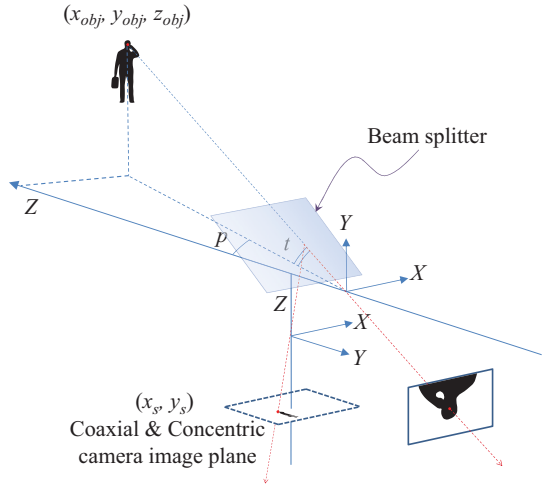
Fig. 4. Localization error ($Err(\theta_{\mathbf{pt}})$) in degrees between the desired and estimated directions of the PTZ camera with respect to the ratio (α) between ($\mathbf{w_{obj}}$) and ($\mathbf{w_{calib}}$) and the distance ($\mathbf{d}$) between the static and PTZ cameras. The error is minimized when $\alpha = 1$ or $\|\mathbf{d}\| = 0$.



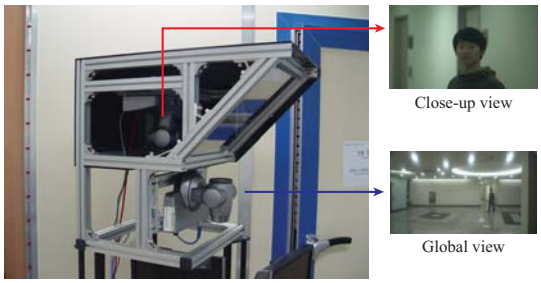
(a) Schematic of the proposed camera system with two cameras



(b) Actual implementation of the proposed system described in (a)



(c) Schematic of the proposed camera system with a beam splitter



(d) Actual implementation of the proposed system described in (c)

Fig. 5. Schematic of the proposed Coaxial-Concentric camera systems and the corresponding face images obtained at global and close-up views: (a) two static cameras are placed above and beside the PTZ camera to generate the virtual camera in a coaxial position w.r.t. the PTZ camera and (c) beam splitter divides a beam of light into the static and PTZ cameras. (b) and (d) are the images of the actual camera system corresponding to (a) and (c), respectively, and their static and PTZ views.

or closer distance than the calibrated distance, the object will not be in the center of the PTZ camera's field of view, as shown in Fig. 3(a). The first condition is difficult to satisfy in practice because the object can appear at any distance from the camera independent of the calibrated distance. However, the second condition can be satisfied using the proposed camera configuration, which we call the Coaxial-Concentric camera configuration. Fig. 4 shows the simulation results of the amount of error with different values of α and $\|\mathbf{d}\|$. It can be seen that the error is always zero when $\alpha = 1$, regardless of $\|\mathbf{d}\|$ or $\|\mathbf{w_{calib}}\|$. The error also increases with $\|\mathbf{d}\|$ at fixed α and $\|\mathbf{w_{calib}}\|$. The overall error decreases as $\|\mathbf{w_{calib}}\|$ increases.

The proposed Coaxial-Concentric configuration of static and PTZ cameras has the following properties: (i) coaxial; the axes of both the cameras are parallel so that the views of static and PTZ cameras overlap and (ii) concentric; focal point of static camera and center of rotation of PTZ camera coincide ($\|\mathbf{d}\| = 0$). Due to the infeasibility of designing such a hardware system[1], we propose two types of camera systems that effectively satisfy the requirements of coaxial and concentric camera configurations as follows.

1) Camera system with dual static cameras: We configure two static cameras, one above (horizontal camera), and one beside (vertical camera) the PTZ camera, so that the X coordinate (Y coordinate) of the horizontal (vertical) camera's focal point coincides with the X coordinate (Y coordinate) of the PTZ camera's center of rotation as shown in Fig. 5(a) [1]. All cameras are also configured to have parallel camera axes. The mapping function F from the static image coordinate to the pan-tilt parameters can thus be estimated as $(p, t) = F(x_s^h, y_s^v, 1)$ from the coordinates of the horizontal and vertical static cameras. However, this configuration is computationally demanding since it has to estimate corresponding points in the two static camera images.

2) Camera system with a beam splitter: A beam splitter is an optical device that splits a beam of light into two. We configure a hexahedral dark box with one of its side tilted by 45 degrees and attached to a beam splitter as shown in Figs. 5(d) and 6. The inside of the hexahedral box also needs to be sufficiently dark to get sharp images. PTZ camera is configured inside the dark box and the static camera is placed outside the box. The incident beam is split at the beam splitter

[1]The concentric configuration requires two different cameras physically overlapped.

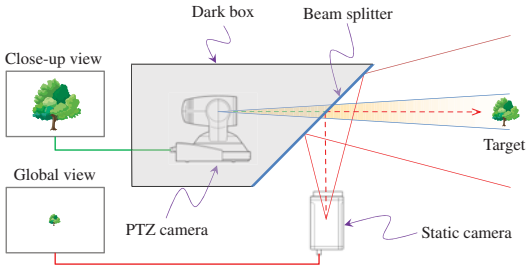


Fig. 6. Cross-sectional diagram of the proposed camera system using a beam splitter.

and captured by both PTZ and static cameras to provide almost the same image[2] to both the cameras. All the camera axes are effectively parallel in this configuration. This configuration enables the use of a single static camera to estimate the pan and tilt parameters of the PTZ camera.

Fig. 3 shows the effectiveness of the proposed Coaxial and Concentric system over the Non-Coaxial-Concentric system. In Fig. 3, the mapping function F is calculated at $\|\mathbf{w_{calib}}\| = 5m$, the resolution of the PTZ camera images is 640×480, the zoom is controlled to capture the target face with ∼70 pixels of IPD, and the face images are captured from five to ten[3] meter range. The Coaxial-Concentric system captures the face in the center of the image at all distances, while the Non-Coaxial-Concentric system ($\|\mathbf{d}\| = 35cm$) is not even able to capture the face as the distance increases. The proposed Coaxial-Concentric camera system can also be operated at a distance of less than 5 m or larger than 12 m. However, for distances larger than 12 m the static camera used in our system (with a resolution of 1280×720) cannot reliably detect the subject and his face location[4].

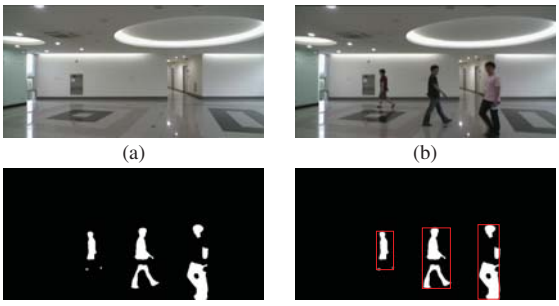


Fig. 7. Object detection: (a) background image, (b) input image, (c) background subtraction to obtain blobs, and (d) detected objects.

C. Subject Tracking

We use a conventional background subtraction method [39], followed by morphological operations to obtain the "blob" associated with the subjects in the field of view (Fig. 7(c)). Background subtraction method is a commonly used technique to segment an object from the background. However, the segmentation often fails when the color of the object is similar to the background as shown in Fig. 7(c). Thus, we utilize a heuristic clustering method to combine nearby blobs as shown in Fig. 7(d) to improve the segmentation.

Typical blob tracking processes utilize the size of overlapping area [40], or other blob features such as color or distance between the blobs. A combination of various blob features (e.g., color, distance, etc.) are also used to create the so-called matching matrices [41]. In many cases, Kalman filter is used to predict the position of the blob in a frame and match it with the closest blob [42]. The use of blob trajectory [42] or blob color [40] helps to solve occlusion problems.

After the blob detection, we compare the detected blobs in each frame to associate an ID with smooth spatio-temporal continuity. Given a captured image, I_i, $i = 1, \ldots, N$, detect blobs $B^i_{o,v}$, $v = 1, \ldots, V$ in each image. Let V_i represent the number of blobs, V, in the i^{th} image. Then, the addition or removal of a blob (person) can be decided by comparing V_{i-1} and V_i. The blobs in the i^{th} image can be associated with those in $(i-1)^{th}$ image by comparing the similarities between $B^{i-1}_{o,v}$ and $B^i_{o,v}$. This person tracking is essentially associating the membership of blobs detected in each image, $I_i, i = 1, \ldots, N$, or in successive images, I_{i-1} and I_i. Rather than comparing $B^{i-1}_{o,v}$ and $B^i_{o,v}$, we introduce the predicted blob, $B^i_{p,r}, r = 1, \ldots, R$ and compare $B^i_{p,r}$ and $B^i_{o,v}$. The predicted blob, $B^i_{p,r}$, is computed using $B^1_{o,v}$,...,$B^{i-1}_{o,v}$. The prediction starts after a minimum number ($= C_w$) of frames are captured. Algorithm 1 summarizes the person tracking algorithm.

Algorithm 1 Blob tracking algorithm

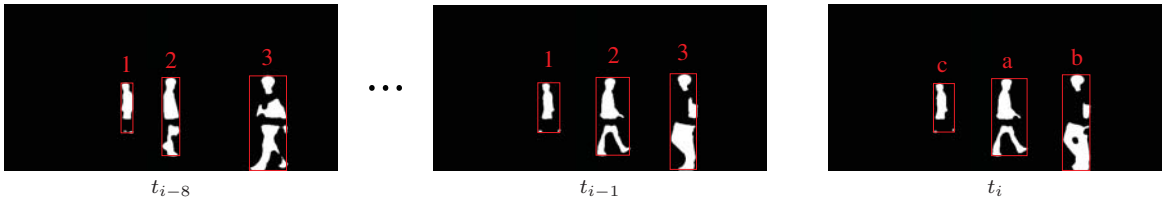
```
for i = 1 → N do
    if i ≤ C_w then
        B^i_{p,r}, ← B^i_{o,v}
    else
        B^i_{p,r}, ← predict(B^1_{o,v},...,B^{i-1}_{o,v})
        Associate membership of B^i_{o,v} with B^{i-1}_{o,v}
      based on s(B^i_{p,r}, B^i_{o,v}).
    end if
end for
```

In order to calculate of the similarity between blobs in successive frames, we use three different attributes of each blob, head coordinates, color, and size. These attributes are represented by $B^{i,head}_{o,v}$, $B^{i,color}_{o,v}$, and $B^{i,size}_{o,v}$, for the v^{th} blob in the i^{th} image. The location of the head is estimated using the height of the blob; the height of head is empirically estimated as one seventh the height of a blob . The prediction of a blob property is calculated by using a linear prediction model. Let x_{i-1} and y_{i-1} denote the location of a blob in the $(i-1)^{th}$ frame and t_{i-1} be the time (ms) of observation of x_{i-1} and y_{i-1}. Then, the predicted head position (x_i, y_i), in the i^{th} image can be computed from a number (=C_w) of previously estimated values using the following two-step

[2]Images are slightly different due to the differences in camera optics (e.g., lens, charge-coupled device, etc.).

[3]Our system can handle objects up to 12 meters.

[4]This is the highest image resolution static and PTZ type of cameras available in the market with regular video frame rate (=30).



t_{i-8} t_{i-1} t_i

(a) (b)

Features	1	2	3
Head position	(241, 182)	(384, 180)	(537, 178)
Color	(43, 32, 33)	(69, 66, 63)	(119, 103, 107)
Size	(37, 107)	(56, 169)	(69, 206)

(c)

Features	a	b	c
Head position	(384, 184)	(535, 181)	(240, 185)
Color	(90, 88, 85)	(125, 111, 114)	(50, 37, 39)
Size	(77, 167)	(58, 207)	(47, 106)

(d)

	a	b	c
1	0.82	0.63	0.99
2	0.99	0.81	0.82
3	0.80	0.99	0.63

(e)

	a	b	c
1	0.79	0.68	0.97
2	0.91	0.80	0.90
3	0.91	0.97	0.73

(f)

	a	b	c
1	0.78	0.70	0.98
2	0.96	0.89	0.81
3	0.88	0.98	0.70

(g)

	a	b	c
1	0.51	0.30	0.95
2	0.87	0.58	0.60
3	0.65	0.95	0.32

(h)

Fig. 8. Example of person tracking with three features (head position, color, and size) in static view: blobs in (a) are from previous image frames (t_{i-8} *and* t_{i-1}) and (b) current image frame (t_i); (c) features of predicted blobs at t_i; (d) observed blobs at t_i; (e) scores based on head positions; (f) scores based on the color of torsos; (g) scores based on the size; (h) final scores obtained by combining scores in (e), (f), and (g). The symbols 1, 2, and 3 are used for identified blobs (subjects) and a, b, and c are used for unidentified blobs. The tracking process finds the correct association of a, b, and c with 1, 2, and 3.

recursive update

$$\begin{aligned} M_{i-1} &= \begin{bmatrix} b_1 & b_2 \\ b_3 & b_4 \end{bmatrix} \\ &= D_{i-1}{K_{i-1}}^T \left(K_{i-1}{K_{i-1}}^T\right)^{-1} \\ where\ D_{i-1} &= \begin{bmatrix} x_{i-1} & \dots & x_{i-C_w} \\ y_{i-1} & \dots & y_{i-C_w} \end{bmatrix} \\ and\ K_{i-1} &= \begin{bmatrix} t_{i-1} & \dots & t_{i-C_w} \\ 1 & \dots & 1 \end{bmatrix}. \end{aligned} \tag{4}$$

The predicted position of the head is

$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} = M_{i-1} \begin{bmatrix} t_i \\ 1 \end{bmatrix}. \tag{5}$$

The color and size can be similarly predicted by Eqs. (4) and (5). The predicted blob properties, $B_{p,r}^{i,head,color,size}$, are averaged with the observed properties in the previous image, $B_{o,r}^{i-1,head,color,size}$, to smooth noisy estimates. The predicted blob properties are finally used to calculate the similarity between each blob in successive images.

We define the similarity score between the r^{th} predicted and v^{th} observed head positions in the i^{th} frame as

$$s^{head}(B_{p,r}^{i,head}, B_{o,v}^{i,head}) = 1 - \frac{\left\| B_{p,r}^{i,head} - B_{o,v}^{i,head} \right\|}{\|(width, height)\|} \tag{6}$$

where $width$ and $height$ are horizontal and vertical length of static camera images.

When a subject is moving, his limbs are often fragmented, whereas the torso part is rather stable. Furthermore, the color of the torso is more stable than that of, e.g., his shirt sleeve's. Therefore, we estimate the torso region with respect to the ratio of height and width of the blob and compute the similarity between blobs based on the average RGB colors of the torso regions as

$$s^{color}(B_{p,r}^{i,color}, B_{o,v}^{i,color}) = 1 - \frac{\left\| B_{p,r}^{i,color} - B_{o,v}^{i,color} \right\|}{\|(255, 255, 255)\|} \tag{7}$$

where the component values of RGB color are stored as integer numbers in the range [0, 255].

We also consider the size of blobs in terms of their width and height as

$$\begin{aligned} s^{size}(B_{p,r}^{i,size}, B_{o,v}^{i,size}) &= \prod_{d \in (w,h)} \left(1 - \frac{\left| B_{p,r}^{i,size} - B_{o,v}^{i,size} \right|}{L(d)}\right) \\ where\ L(w) &= width\ and\ L(h) = height. \end{aligned} \tag{8}$$

This method shows better performance than using the diagonal length of a blob because the diagonal length can have the same value for two different blobs with different shapes. The final similarity score is calculate by taking the summation of three scores as

$$\begin{aligned} s(B_{p,r}^{i}, B_{o,v}^{i}) = {} & \omega_1 \cdot s^{head}(B_{p,r}^{i,head}, B_{o,v}^{i,head}) + \\ & \omega_2 \cdot s^{color}(B_{p,r}^{i,color}, B_{o,v}^{i,color}) + \\ & \omega_3 \cdot s^{size}(B_{p,r}^{i,size}, B_{o,v}^{i,size}), \end{aligned} \tag{9}$$

with equal weights. Fig. 8 shows an example of the blob similarity comparison process. For each blob being tracked, a random ID is assigned to differentiate it from the other blobs. When a blob is identified in the high resolution face images captured by the proposed system, the blob is assigned with a permanent ID.

D. Zoom Control

The height of detected objects in static camera images is used for zoom control. We manually measure ten magnification factors of the PTZ camera to ensure that the distance between the two eyes is at least 100 (60) pixels in the PTZ view with a resolution of 1280×720 (640×360) pixels and their corresponding blob heights from a set of training data. A quadratic mapping function between the height (h) of the blob and zoom values (z) of the PTZ camera is obtained by

$$z = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \begin{bmatrix} h^2 & h & 1 \end{bmatrix}^T. \quad (10)$$

E. System Configuration

There are two different implementations of the proposed system: (1) System without a beam splitter: two Sony EVI-HD1 cameras are used as static cameras to obtain the vertical and horizontal global views and one Sony EVI-D100 camera is used as a PTZ camera to track and acquire high resolution face images at a distance. The image resolutions are 720×360 and 720×486 pixels for the static and PTZ views, respectively. (2) System with beam splitter: two Sony EVI-HD1 cameras set to 1280×720 pixel resolution are used as static and PTZ cameras. All image acquisition and processing modules are implemented in C++ and utilize the OpenCV Library [43]. The PTZ camera is controlled using the standard RS-232 serial port. The tracking and camera control components run in real time (8 fps) on a quad core Intel 2.8 GHz machine.

The system is decomposed into static camera processing and dynamic camera control modules (Fig. 9). The former includes image capture, background subtraction, and object and head tracking. The latter performs face location prediction and camera control (i.e., pan-tilt and speed control). The static processing module sends target locations of faces in each frame to the dynamic camera control module. The PTZ camera control module adjusts pan-tilt angles to observe the target(s) in the field of view.

III. Camera Control for Smooth Tracking

There are two components in the PTZ camera control module: the pan and tilt parameter controller (PTC) and the motion velocity controller (MVC). The PTC predicts the next head location given the previous head trajectory. The estimated head location is converted to pan and tilt values. Given a set of pan and tilt values, the MVC controls the velocity of pan and tilt motion. While there have been a few previous studies on the static image processing part [44], no systematic study has been reported on the dynamic camera control part.

A. Pan and Tilt Controller

The objective of the camera control is to keep the subject being tracked in the center of the PTZ camera view. By setting the head location to the center of the PTZ camera view, the possibility of losing track of the face in the next frame is minimized. Controlling the camera with the current location of the head and its corresponding pan and tilt values does not provide robust tracking capability due to delays in image processing and mechanical camera motion. To solve this problem, we use a linear prediction model similar to the prediction model used in the blob tracking process (see Sec. II-C).

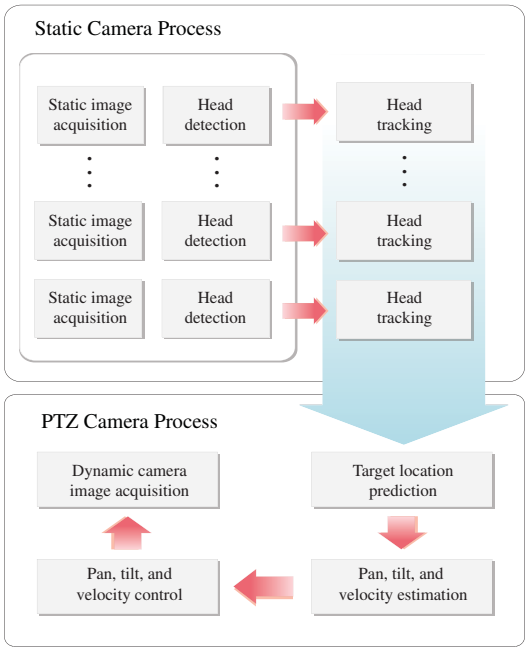


Fig. 9. Schematic of the process flow in the proposed camera system.

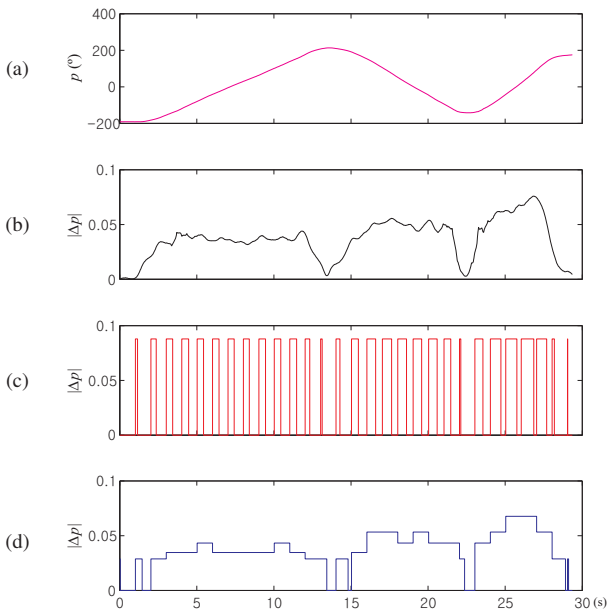


Fig. 10. Motion velocity-time graphs: (a) ground truth pan value-time graph of a PTZ camera during the tracking process, (b) ground truth velocity-time graph and simulated velocity-time graph by using (c) fixed velocity and (d) average velocity.

B. Motion Velocity Controller

The PTZ camera in our system provides 24 levels of pan speeds from 2 to 300 degrees/sec and 20 levels of tilt speeds

from 2 to 125 degrees/sec. In typical PTZ camera systems, a fixed speed is used at each camera control command. However, the fixed speed strategy can cause non-smooth control of the camera, resulting in a higher probability of losing the subject or resulting in a blurred image. In our system, the PTZ camera speed is calculated based on the current and the next predicted head location (average speed). Fig. 10 shows a comparison of the two different camera velocity control methods: (i) fixed velocity and (ii) average velocity. Fig. 10(a) indicates ground truth pan values of a PTZ camera in tracking a moving object, extracted from a 30-second static video (60 fps) and Fig. 10(b) is the ground truth pan velocity-time graph. Assuming that the speed of PTZ camera is controlled once every second, Figs. 10(c) and 10(d) show simulated results of pan velocity-time graph for fixed velocity and average velocity methods. While the fixed speed method shows discontinuous speed profile, the average speed method shows a smoother profile that is more similar to the ground truth pan velocity-time graph.

IV. Application to Face Recognition

In order to verify the face recognition capability of the proposed system in surveillance applications, we conducted face recognition tests at a distance of up to 12 meters. We compared the face identification accuracies using both the conventional static camera and the proposed camera systems to show the effectiveness of the proposed system. All the data were collected using the two-camera system with a beam splitter because of its lower computational complexity compared to the three-camera system. Our earlier results with the three-camera system can be found in [1].

A. Experimental Data

We captured probe images of subjects by using the proposed system with a beam splitter in two different surveillance scenarios as follows.

1) Single person tracking: We captured videos of 50 subjects at a distance ranging from 6 to 12 m using both static (Fig. 11(c)) and PTZ cameras (Fig. 11(b)). All the video data were collected indoors at Korea University campus; the subjects were Korea university students. Each subject was asked to walk starting at about 12 m from the camera up to about 6 m distance by making an S-shaped path to evaluate the tracking capability of the proposed system. The average duration of each video is about 25 seconds at 30 fps.

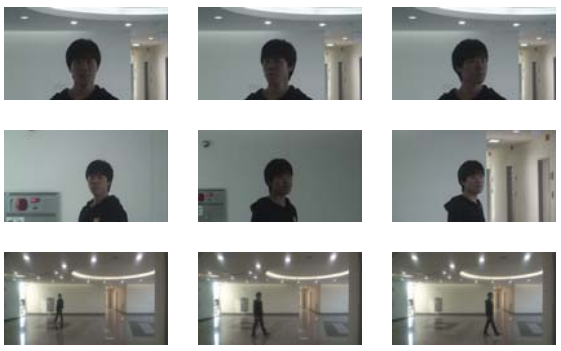


Fig. 11. Gallery and probe images captured by the proposed system: (a) frontal, left and right facial images for gallery and probe images captured by (b) PTZ camera and (c) static camera.

2) Multi-person tracking: We captured 40 videos of 3 subjects at a distance ranging from 5 to 10 m in 4 different scenarios as shown in Fig. 12: (1) people are not moving, (2) people are moving without overlap, (3) people are crossing each other, and (4) people are passing each other in the same direction. After 200 frames have been captured for a subject, the camera system automatically moves towards other subjects not yet identified. Each video is manually segmented according to the subjects in the field of view to establish the ground truth to evaluate the face recognition performance.

TABLE II
Face recognition accuracy of conventional static and proposed PTZ camera systems

Approach	Rank-1 accuracy(%)	
	Single person	Multi-person
Static view (conventional surveillance system)	0	0
PTZ view, 1 frame, (coaxial camera system)	55.1	42.6
PTZ view, 1 frame, t_r=0.31	63.1	50.0
PTZ view, 1 frame, t_r=0.45	68.2	64.7
PTZ view, fusion of 2 frames, t_r=0.45	79.9	78.0
PTZ view, fusion of 5 frames, t_r=0.45	88.3	89.0
PTZ view, fusion of 10 frames, t_r=0.45	91.5	93.4

The gallery data consists of three images per subject captured at about 1 m distance from the camera at three different poses (Fig. 11(a)). Additional 10,000 images of 10,000 subjects from the MORPH database [45] were added to the gallery to increase the complexity of face recognition in the identification mode.[5]

B. Results and Analysis

We performed face recognition experiments using all the frames in the collected video data set as probe; 102,978 (36,574) images of 50 subjects and 10,150 (10,009) images of 10,050 (10,003) subjects as gallery for single person (multi-person) tracking. A commercial face recognition engine, FaceVACS [46], was used for face detection and recognition. We rejected probe images with matching scores less than 0.31 and 0.45[6] in the PTZ view to compare the results of previous experiments with dual static cameras. The range of matching scores provided by FaceVACS is [0,1]. The probe images from static views show almost complete failure of face recognition and the rejection scheme did not help in improving the identification accuracy. Table II shows the Rank-1 face identification accuracies obtained from the static and PTZ

[5]Even though the face images in MORPH are different from the faces in probe videos in terms of pose, overall face size, and ethnicity, it is the only large scale public domain face image database available.

[6]The matching scores 0.31 and 0.45 correspond to the smallest non-zero score and a score with 40% rejection rate, respectively.

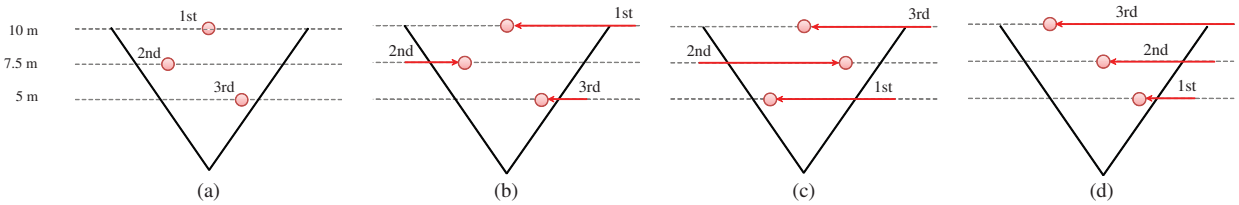


Fig. 12. Four different scenarios to evaluate the tracking capability with multi-person tracking: (a) subjects are stationary, (b) subjects are moving without overlap, (c) subjects are crossing each other while moving into different directions, and (d) subjects are passing each other while moving in the same direction.

views. The single person recognition results in Table II are slightly lower than those in [1] because of the increase in the gallery size (from 10,020 to 10,050) and different populations in probe data set (from 20 to 50 subjects). The threshold score used for rejection is indicated by t_r. While the identification accuracy of the PTZ view is 55.1% (42.6%) in single (multi) person tracking, that of the static view is no better than random guess. Frame level fusions using the score-sum method [47] with contiguous 2, 5, and 10 frames after rejection scheme (t_r = 0.45) shows further improvement of 23.3% (28.7%) in the identification accuracy. For example, in the fusion with 5 frames, the matching scores of the probe image at time t to all the gallery images are summed with those of probe images at time $t-1$, ..., $t-4$. The identity is decided based on the summed scores. Figs. 14 and 13 show example probe images that were successfully matched and not successfully matched at rank-1. Major reasons of the failures are (i) inability to track a face, (ii) off-frontal facial pose, (iii) motion blur, and (iv) non-neutral facial expression.

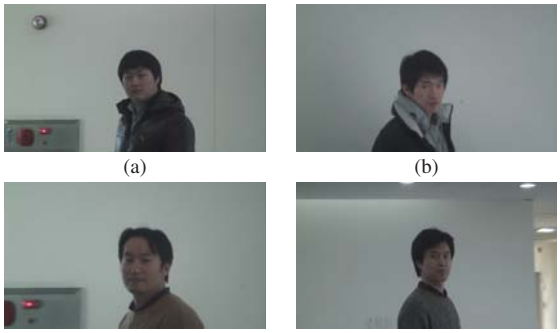


Fig. 13. Example probe images successfully matched at rank-1.

Fig. 14. Example probe images that could not be matched at rank-1 due to (a) tracking failure, (b) off-frontal pose, (c) motion blur and (d) non-neutral expression.

V. Conclusions and Future Work

We have proposed a novel Coaxial-Concentric camera system that can capture and track high resolution face images (with inter-pupillary distance of about 100 pixels) at any distance in the range of 6 to 12 meters for face recognition. The Coaxial-Concentric camera configuration provides a large operating distance to track moving persons and recognize them with high accuracy. We have introduced a linear prediction model and a pan and tilt motion velocity control method for robust tracking. The face recognition results show the effectiveness of the proposed system for fully automatic subject tracking and identification at a distance of up to 12 meters.

The limitations of the current system are as follows: (i) the static and PTZ cameras have to be manually adjusted[7] to satisfy the Coaxial-Concentric conditions because the focal point and center of rotation cannot be directly handled from outside the cameras; (ii) the operating distance is limited to $\sim$12 m due to the limitation of object detection in static camera; and (iii) the system can recognize a face only when it is close to the frontal pose, which is an inherent limitation of the state of the art face matchers. We plan to seek a more efficient method of calibration between static and PTZ cameras in the Coaxial-Concentric configuration. We also plan to extend the operating distance beyond 12 meters by using either a high definition static camera or multiple PTZ cameras to employ multi-stage zooming process. In limited scenarios, manual control of the PTZ camera can also be considered to increase the operating distance.

[7]The manual adjustment is required only once at the initial system setup. The complete system can be deployed to other places with no further manual adjustment.

Appendix

The condition to minimize the error between the desired and calibrated $\theta_{\mathbf{pt}}$ values in Eq.((3)) is derived as:

$$\begin{aligned} & Err\left(\theta_{\mathbf{pt}}\right) = 0 \\ \Leftrightarrow & \cos^{-1}\left(\frac{(\alpha\mathbf{w_{calib}} - \mathbf{d}) \cdot (\mathbf{w_{calib}} - \mathbf{d})}{\|\alpha\mathbf{w_{calib}} - \mathbf{d}\| \, |\mathbf{w_{calib}} - \mathbf{d}|}\right) = 0 \\ \Leftrightarrow & \left(\frac{(\alpha\mathbf{w_{calib}} - \mathbf{d}) \cdot (\mathbf{w_{calib}} - \mathbf{d})}{\|\alpha\mathbf{w_{calib}} - \mathbf{d}\| \, \|\mathbf{w_{calib}} - \mathbf{d}\|}\right) = 1 \\ \Leftrightarrow & (\alpha\mathbf{w_{calib}} - \mathbf{d}) \cdot (\mathbf{w_{calib}} - \mathbf{d}) = \\ & \qquad \|\alpha\mathbf{w_{calib}} - \mathbf{d}\| \, \|\mathbf{w_{calib}} - \mathbf{d}\| . \end{aligned} \tag{11}$$

Expand the right hand side by the definition of vector norm,

$$(\alpha\mathbf{w_{calib}} - \mathbf{d}) \cdot (\mathbf{w_{calib}} - \mathbf{d}) = \sqrt{(\alpha\mathbf{w_{calib}} - \mathbf{d}) \cdot (\alpha\mathbf{w_{calib}} - \mathbf{d})} \cdot \sqrt{(\mathbf{w_{calib}} - \mathbf{d}) \cdot (\mathbf{w_{calib}} - \mathbf{d})}. \quad (12)$$

By squaring both sides,

$$\{(\alpha\mathbf{w_{calib}} - \mathbf{d}) \cdot (\mathbf{w_{calib}} - \mathbf{d})\}^2 = \{(\alpha\mathbf{w_{calib}} - \mathbf{d}) \cdot (\alpha\mathbf{w_{calib}} - \mathbf{d})\} \cdot \{(\mathbf{w_{calib}} - \mathbf{d}) \cdot (\mathbf{w_{calib}} - \mathbf{d})\}. \quad (13)$$

By expanding brackets and simplifying,

$$\begin{aligned}
&\left\{\alpha \left\|\mathbf{w_{calib}}\right\|^2 - (\alpha+1)\left(\mathbf{w_{calib}} \cdot \mathbf{d}\right) + \left\|\mathbf{d}\right\|^2\right\}^2 = \\
&\qquad \left(\alpha^2 \left\|\mathbf{w_{calib}}\right\|^2 - 2\alpha\left(\mathbf{w_{calib}} \cdot \mathbf{d}\right) + \left\|\mathbf{d}\right\|^2\right) \cdot \\
&\qquad \left(\left\|\mathbf{w_{calib}}\right\|^2 - 2\left(\mathbf{w_{calib}} \cdot \mathbf{d}\right) + \left\|\mathbf{d}\right\|^2\right) \\
&\Leftrightarrow 2\alpha \left\|\mathbf{w_{calib}}\right\|^2 \left\|\mathbf{d}\right\|^2 + (\alpha+1)^2 \left(\mathbf{w_{calib}} \cdot \mathbf{d}\right)^2 = \\
&\qquad \left(\alpha^2+1\right) \left\|\mathbf{w_{calib}}\right\|^2 \left\|\mathbf{d}\right\|^2 + 4\alpha\left(\mathbf{w_{calib}} \cdot \mathbf{d}\right)^2 \quad (14) \\
&\Leftrightarrow (\alpha-1)^2 \left\{\left(\mathbf{w_{calib}} \cdot \mathbf{d}\right)^2 - \left\|\mathbf{w_{calib}}\right\|^2 \left\|\mathbf{d}\right\|^2\right\} = 0 \\
&\Leftrightarrow (\alpha-1)^2 \left\|\mathbf{w_{calib}}\right\|^2 \left\|\mathbf{d}\right\|^2 \left(\cos^2\theta - 1\right) = 0 \\
&\Leftrightarrow \alpha = 1 \; or \; \left\|\mathbf{d}\right\| = 0 \\
&\quad (\because \left\|\mathbf{w_{calib}}\right\| > 0 \; and \\
&\qquad \cos\theta_{(\mathbf{w_{calib}},\mathbf{d})} \neq 1 \; unless \; \mathbf{w_{calib}} \times \mathbf{d} = 0)
\end{aligned}$$

where $\theta_{(\mathbf{w_{calib}},\mathbf{d})}$ is the angle between $\mathbf{w_{calib}}$ and $\mathbf{d}$. Therefore, the overall error is minimized when α is equal to one or $\mathbf{d}$ is a zero vector; the first condition cannot be satisfied in practice because the object can appear at any distance, but the second condition can be satisfied by using the proposed Coaxial-Concentric configuration.

ACKNOWLEDGMENT

This research was supported by WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-10008).

REFERENCES

[1] H.-C. Choi, U. Park, and A. K. Jain, "Ptz camera assisted face acquisition, tracking & recognition," in *Proc. IEEE Int. Conf. Biometrics: Theory, Applications and Systems*, 2010, pp. 1–6.

[2] S. Z. Li and A. K. Jain (Eds.), *Handbook of Face Recognition*. Springer, 2005.

[3] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 large-scale results," *Technical Report NISTIR 7408, Nat'l Inst. of Standards and Technology*, Mar. 2007.

[4] K. Bernardin, F. V. D. Camp, and R. Stiefelhagen, "Automatic person detection and tracking using fuzzy controlled active cameras," in *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[5] A. Mian, "Realtime face detection and tracking using a single Pan, Tilt, Zoom camera," in *Proc. Int'l Conf. Image and Vision Computing*, 2008, pp. 1–6.

[6] C. Yang, R. Chen, C. Lee, and S. Lin, "PTZ camera based position tracking in IP-surveillance system," in *Proc. Int'l Conf. Sensing Technology*, 2008, pp. 142–146.

[7] P. Kumar, A. Dick, and T. Sheng, "Real time target tracking with pan tilt zoom camera," in *Conf. Digital Image Computing: Techniques and Applications*, 2010, pp. 492–497.

[8] P. Varcheie and G. Bilodeau, "Active people tracking by a PTZ camera in IP surveillance system," in *Proc. IEEE Int'l Workshop on Robotic and Sensors Environments*, 2009, pp. 98–103.

[9] S. Venugopalan and A. Savvides, "Unconstrained iris acquisition and recognition using COTS PTZ Camera," *EURASIP Journal on Advances in Signal Processing*, 2010.

[10] P. D. Z. Varcheie and G.-A. Bilodeau, "Adaptive fuzzy particle filter tracker for a PTZ Camera in an IP surveillance system," *IEEE Trans. Instrumentation and Measurement*, vol. 60, pp. 354–371, 2011.

[11] I. Everts, N. Sebe, and G. Jones, "Cooperative object tracking with multiple PTZ cameras," in *Proc. Int'l Conf. Image Analysis and Processing*, 2007, pp. 323–330.

[12] H. Liao and W. Chen, "A dual-PTZ-camera system for visual tracking of a moving target in an open area," in *Proc. Int'l Conf. Advanced Communication Technology*, vol. 1, 2009, pp. 440–443.

[13] A. del Bimbo, F. Dini, G. Lisanti, and F. Pernici, "Exploiting distinctive visual landmark maps in pan-tilt-zoom camera networks," *Computer Vision and Image Understanding*, vol. 114, pp. 611–623, 2010.

[14] H.-C. Liao and W.-Y. Chen, "Eagle-Eye: a dual-PTZ-Camera system for target tracking in a large open area," *Information Technology and Control*, vol. 39, pp. 227–235, 2010.

[15] F. Wheeler, R. Weiss, and P. Tu, "Face recognition at a distance system for surveillance applications," in *Proc. IEEE Int'l Conf. Biometrics: Theory Applications and Systems*, 2010, pp. 1–8.

[16] J. Zhou, D. Wan, and Y. Wu, "The chameleon-like vision system," *IEEE Signal Processing Magazine*, vol. 27, pp. 91–101, 2010.

[17] R. Bodor, R. Morlok, and N. Papanikolopoulos, "Dual-camera system for multi-level activity recognition," in *Proc. IEEE/RSJ Conf. Intelligent Robots and Systems*, vol. 1, 2005, pp. 643–648.

[18] L. Marchesotti, S. Piva, A. Turolla, D. Minetti, and C. Regazzoni, "Cooperative multisensor system for real-time face detection and tracking in uncontrolled conditions," in *SPIE Int'l Conf. Image and Video Communications and Processing*, 2005.

[19] T. Funahasahi, M. Tominaga, T. Fujiwara, and H. Koshimizu, "Hierarchical face tracking by using PTZ camera," in *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, 2004, pp. 427–432.

[20] S. Yoon, H. G. Jung, K. R. Park, and J. Kim, "Nonintrusive iris image acquisition system based on a pan-tilt-zoom camera and light stripe projection," *Optical Engineering*, vol. 48, 2009.

[21] P. Amnuaykanjanasin, S. Aramvith, and T. H. Chalidabhongse, "Face tracking using two cooperative static and moving cameras," in *Proc. IEEE Int'l Conf. Multimedia and Expo*, 2005, pp. 1158–1161.

[22] S. Prince, J. Elder, Y. Hou, M. Sizinstev, and E. Olevsky, "Towards face recognition at a distance," in *Proc. Crime and Security*, 2006, pp. 570–575.

[23] C. Chen, Y. Yao, D. Page, B. Abidi, A. Koschan, and M. Abidi, "Heterogeneous fusion of omnidirectional and PTZ cameras for multiple object tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1052–1063, 2008.

[24] G. Scotti, L. Marcenaro, C. Coelho, F. Selvaggi, and C. Regazzoni, "Dual camera intelligent sensor for high definition 360 degrees surveillance," *IEE Proceedings Vision, Image & Signal Processing*, vol. 152, no. 2, pp. 250–257, 2005.

[25] M. Tarhan and E. Altug, "A Catadioptric and Pan-Tilt-Zoom Camera Pair Object Tracking System for UAVs," *Journal of Intelligent & Robotic Systems*, vol. 61, pp. 119–134, 2011.

[26] Y. Lu and S. Payandeh, "Cooperative hybrid multi-camera tracking for people surveillance," *Canadian Journal of Electrical and Computer Engineering*, vol. 33, no. 3, pp. 145–152, 2008.

[27] Y. Yao, B. Abidi, and M. Abidi, "3D target scale estimation and target feature separation for size preserving tracking in PTZ video," *International Journal of Computer Vision*, vol. 82, pp. 244–263, 2009.

[28] G. S. V. S. Sivaram, M. S. Kankanhalli, and K. R. Ramakrishnan, "Design of multimedia surveillance systems," *ACM Trans. Multimedia Computing, Communications and Applications*, vol. 5, pp. 23:1–23:25, 2009.

[29] Y. Xu and D. Song, "Systems and algorithms for autonomous and scalable crowd surveillance using robotic PTZ cameras assisted by a wide-angle camera," *Autonomous Robots*, vol. 29, pp. 53–66, 2010.

[30] S. Stillman, R. Tanawongsuwan, and I. Essa, "Tracking multiple people with multiple cameras," in *Proc. Int'l Conf. Audio-and Video-based Biometric Person Authentication*, 1999.

[31] A. Hampapur, S. Pankanti, A. Senior, Y.-L. Tian, L. Brown, and R. Bolle, "Face cataloger: multi-scale imaging for relating identity to location," in *Proc. IEEE Conf. Advanced Video and Signal Based Surveillance*, 2003, pp. 13–20.
[32] N. Krahnstoever, T. Yu, S. Lim, K. Patwardhan, and P. Tu, "Collaborative real-time control of active cameras in large scale surveillance systems," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, 2008.
[33] G. Dedeoglu, T. Kanade, and J. August, "High-zoom video hallucination by exploiting spatio-temporal regularities," vol. 2, 2004, pp. 151–158.
[34] J. Park and S. Lee, "Stepwise reconstruction of high-resolution facial image based on interpolated morphable face model," in *Proc. Int'l Conf. Audio-and Video-based Biometric Person Authentication*, 2005, pp. 102–111.
[35] Y. Yao, B. Abidi, N. Kalka, N. Schmid, and M. Abidi, "Improving long range and high magnification face recognition: database acquisition, evaluation, and enhancement," *Computer Vision and Image Understanding*, vol. 111, no. 2, pp. 111–125, 2008.
[36] M. Lalonde, S. Foucher, L. Gagnon, E. Pronovost, M. Derenne, and A. Janelle, "A system to automatically track humans and vehicles with a PTZ camera," in *SPIE Int'l Conf. Visual Information Processing*, vol. 6575, 2007.
[37] R. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE Trans. Robotics and Automation*, vol. 3, no. 4, pp. 323–344, 1987.
[38] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *Proc. IEEE CS Int'l Conf. Computer Vision*, 1999, pp. 666–673.
[39] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detecting moving objects, ghosts, and shadows in video streams," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1337–1342, 2003.
[40] S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *Computer Vision and Image Understanding*, vol. 80, no. 1, pp. 42–56, 2000.
[41] S. Intille, J. Davis, and A. Bobick, "Real-time closed-world tracking," in *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, 2002, pp. 697–703.
[42] R. Rosales and S. Sclaroff, "Improved tracking of multiple humans with trajectory prediction and occlusion modeling," in *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, 1998.
[43] Open Computer Vision Library, http://www.intel.com/research/mrl/research/opencv/.
[44] R. Liu, X. Gao, R. C. X. Zhu, and S. Z. Li, "Tracking and recognition of multiple faces at distances," *Advances in Biometrics, LNCS*, vol. 4642, pp. 513–522, 2007.
[45] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, 2006, pp. 341–345.
[46] FaceVACS Software Developer Kit, http://www.cognitec-systems.de/.
[47] A. Ross, K. Nandakumar, and A. Jain, *Handbook of Multibiometrics*. Springer, 2006.