SoICT 2017

The eighth international Symposium on Information and Communication Technology
Nha Trang, December 7-8, 2017

IMPORTANT DATES

> **SUBMISSION DEADLINE**
~~August 31, 2017~~
→ September 15, 2017

> **AUTHOR NOTIFICATION**
October 22, 2017

> **CAMERA-READY SUBMISSION**
October 30, 2017

> CONFERENCE DATES
December 7-8, 2017

SoICT 2017 E-Updates

> Click here for the latest information

Previous conferences

2010 2011 2012
2013 2014 2015 2016

Organizers

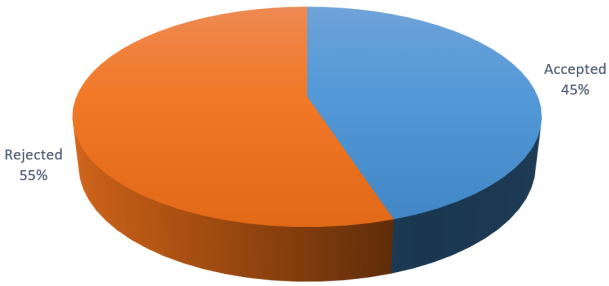
Financial Sponsors

Scuba diving in Nha Trang

Like Share Nguyen Xuan Hoai and 683 others like this.

List of Accepted Papers

Accepted: 45%



Canada China Finland France Germany Japan

North Korea Pakistan Singapore South Korea United States of America Vietnam

1. **Hai Anh Tran, Quang Duc Tran, Linh Giang Nguyen, Abdelhamid Mellouk, Hieu Mac and Van Tong.**
A novel SDN controller based on Ontology and Global Optimization for heterogeneous IoT architecture
2. **Duc Anh Nguyen and Ngoc Hung Pham.**
A Test Data Generation Method for C/C++ Projects
3. **Truong Cao Duy and Duong Tuan Anh.**
Discovering Motifs in a Database of Shapes under Rotation Invariant Dynamic Time Warping
4. **Chi Tran, Tien Dinh and Viviane Gascon.**
Meta-heuristics to solve a districting problem of a public medical clinic
5. **Tam N. Nguyen and Ngoc Q. Ly.**
Abnormal Activity Detection based on Dense Spatial-Temporal Features and Improved One-Class Learning

6. **Giao Bui Cong and Anh Duong Tuan.**
An Application of Similarity Search in Streaming Time Series under DTW: Online Forecasting
7. **Hai Vu, Anh Tuan Pham, Van Giap Nguyen and Thanh-Hai Tran.**
Pedestrian Localization and Trajectory Reconstruction in a Surveillance Camera Network
8. **Khue Doan, Minh Nguyen Quang and Bac Le.**
Applied cuckoo algorithm for association rule hiding problem
9. **Quang Tran Minh, Trong Nhan Phan, Akihiko Takahashi, Tam Thai Thanh, Son Nguyen Duy, Mong Nguyen Thanh and Chau Nguyen Hong.**
A Cost-effective Smart Farming System with Knowledge Base
10. **Ryosuke Miura, Yuuki Takano, Shinsuke Miwa and Tomoya Inoue.**
GINTATE: Scalable and Extensible Deep Packet Inspection System for Encrypted Network Traffic
11. **Quang Tran Minh, Thanh Nguyen Chi and Michel Toulouse.**
Toward a Crowdsourcing-Based Urban Flood Mitigation Platform
12. **Thuc Nguyen Trac and Phuc Do.**
Managing and visualizing citation nework using graph database and LDA model
13. **Phol Le Van, Binh Nguyen Thanh and Ioannis Parissis.**
Mutants Generation For Testing Lustre Programs
14. **Minh Hoa Nguyen, Tung Long Vuong, Thanh Ha Le, Thi Thuy Nguyen, Do Van Nguyen and Dinh Nam Nguyen.**
Moving object detection in compressed domain for high resolution videos
15. **Canh Pham Van, Manh Hoang Dinh, Dinh Hoa Nguyen, Thi Huyen Dang and Xuan Huan Hoang.**
Limiting the Spread of Epidemics within Time Constraint on Online Social Networks
16. **Vinh Van, Thien Thai and Quoc-Minh Nghiem.**
Combining Convolution and Recursive Neural Networks for Sentiment Analysis
17. **Tran Phuong and Tho Le.**
Efficient Secure Text Retrieval on Multi-Keyword Search
18. **Van-Quyet Nguyen and Kyungbaek Kim.**
Estimating the Evaluation Cost of Regular Path Queries on Large Graphs
19. **Van Tung Pham, Haihua Xu, Xiong Xiao, Nancy Chen and Eng-Siong Chng.**
Pruning strategies for partial search in spoken term detection
20. **Van Ho and Ioanis Nikolaidis.**
A Network Coding Based Improvement To The Schedule Template Construction Technique
21. **Trí Nguyễn Hồ Duy, Thức Nguyễn Trác and Phúc Đỗ.**
Creating prior-knowledge of Source-LDA for topic discovery in citation network
22. **Hieu Mac, Duc Tran, Van Tong, Linh Giang Nguyen and Hai Anh Tran.**
DGA Botnet Detection Using Supervised Learning Methods
23. **Van-Hau Nguyen.**
SAT Encodings of Finite-CSP Domains: A Survey
24. **Thuy Le Thi, Tho Quan Thanh and Tuoi Phan Thi.**
Ontology-Based Entity Coreference Resolution For Sentiment Analysis
25. **Sanggon Lee.**
Non-connection Wireless IoT network control system using WiFi and Bluetooth Beacon
26. **Viet-Vu Vu and Hong-Quan Do.**
Graph based Clustering with background knowledge
27. **Diem Truong, Duc-Thuan Vo and Trang Uyen Nguyen.**
Vietnamese Open Information Extraction
28. **Ning Pang, Fengjing Yin, Xiaoyu Zhang and Xiang Zhao.**
A Robust Approach for Multivariate Time Series Forecasting
29. **Tran Manh Thang and Nguyen Khanh Van.**
FDDA: A Framework For Fast Detecting Source Attack In Web Application DDoS Attack
30. **Phung Van Duy, Hoang Minh Tu and Nguyen Tuan Anh**.
DTA Hunter System: A new statistic-based framework of predicting future demand for taxi drivers
31. **Ly Vu, Cong Thanh Bui and Quang Uy Nguyen.**
A Deep Learning Based Method for Handling Imbalanced Problem in Network Traffic Classification
32. **Quang-Hung Nguyen, Nam Thoai and Anhtu Ngoc Tran.**
Implementing Genetic Algorithm Accelerated By Intel Xeon Phi
33. **Dang Tuan Nguyen and Trung Tran.**
Phrasal Graph-based Method for Abstractive Vietnamese Paragraph Compression
34. **Najlah Gali, Radu Mariescu-Istodor and Pasi Franti.**
Functional Classification of Websites
35. **Muhammad Mustafa Hassan.**
Participatory Consensual Assessment Technique End-Users Assessing Outcomes of Participatory Design
36. **H. Chau Le, Thanh-Long Mai and The Ngoc Dang.**
Spectrum Utilization of Coarse Granular Routing Elastic Optical Network
37. **Takahide Matsunaga, Ryouta Yagi, Kenichi Takahashi, Takao Kawamura and Kazunori Sugahara.**
Entrusting a Responsibility of Personal Information Management to Users from a Service Provider
38. **Ngo Xuan Bach, Le Cong Linh and Tu Minh Phuong.**
Cross-Domain Intention Detection in Discussion Forums
39. **Nguyen Ngoc Khai and Hoang Truong.**
A Compositional Type Systems for Finding Log Memory Bounds of Transactional Programs

Pedestrian Localization and Trajectory Reconstruction in a Surveillance Camera Network

Hai Vu
International Research Institute MICA
HUST - CNRS/UMI - 2954 - INP Grenoble ,
Hanoi University of Science and Technology
hai.vu@mica.edu.vn

Van-Giap Nguyen
University of Information and Communication
Technology, Thai Nguyen University
giapnv.ictu@gmail.com

Anh-Tuan Pham
†University of Technology and Logistics,
‡International Research Institute MICA
HUST - CNRS/UMI - 2954 - INP Grenoble,
Hanoi University of Science and Technology
tuan-anh.pham@mica.edu.vn

Thanh-Hai Tran
International Research Institute MICA
HUST - CNRS/UMI - 2954 - INP Grenoble,
Hanoi University of Science and Technology
thanh-hai.tran@mica.edu.vn

ABSTRACT

In this paper, we propose a high accuracy solution for locating pedestrians from video streams in a surveillance camera network. For each camera, we formulate the vision-based localization service as detecting foot-points of pedestrians in the ground plane. We address two critical issues that strongly affect the foot-point's detection results: casting shadows and pruning detection results due to occlusion. For the first issue, we adopt a removing shadow technique based on a learning-based approach. For the second issue, a regression model is proposed to prune the wrong foot-point detection results. The regression model plays a role in estimating the position by using the human factors such as height, width and its ratio. A correlation of the detected foot-points and the results estimated from the regression model is examined. Once a foot-point is missed due to uncorrelation problem, a Kalman filter is deployed to predict the current location. To link the trajectory of the human in the camera network, we base on an observation about the same ground-plane/floor in view of cameras then the transformation between a pair of cameras could be computed offline. In the experiments, a high accuracy performance for locating the pedestrians and a real-time computation are achieved. The proposed method therefore is particularly feasible to deploy the vision-based localization service in scalable indoor environments such as hall-way, squares in public buildings, offices, where surveillance cameras are common used.

CCS CONCEPTS

• **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

KEYWORDS

Human Localization, Surveillance camera, Vision-based tracking, Shadow Removal

SoICT2017, Dec., 2017, Nha Trang, VietNam

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

1 INTRODUCTION

Nowadays, the use of vision sensors is becoming more popular for security and monitoring in public areas. It opens opportunities for developing the vision-based localization technology in different scenarios such as homeland security, crime prevention [9], accident prediction and detection [20], monitoring patients, elderly and children at home [18]. These applications always require positioning information from video streams collected by a surveillance camera network. However, a vision-based localization technique requires many sub-tasks from single/multiple surveillance cameras such as: achieving human silhouette (extraction issues), motion trajectories (tracking issues), and human identification (re-identification issues). The entire system's performance is strongly impacted by the relevant techniques used in each task. For instance, two first tasks of human extraction and tracking often suffer from complicated background, noises, object occlusion, lighting conditions, casting shadows or quality of image/video. As a consequence, deploying a perfect solution is intractable. In this study, we propose to use a cue from the surveillance's contexts, that is cameras share partially the same view of ground-plane. Moreover, in most of situations, people can be assumed as standing on the ground-plane (floor). Their common activities are walking with their foot-point (or cross-leg) touching the ground-plane. These observations are important cues to (1) form the proposed vision-based localization techniques; (2) to more easily merge trajectories in a camera network. It is more useful, particularly, when surveillance environment is partially viewed by cameras. Constructing human trajectories needs to overcome the non-overlapping fields-of-view issue.

To this end, the foot-points are detected from a 2-D image sequence. The foot-point definition is illustrated in Fig. 1. Given an image sequence, the system extracts human silhouette from the video stream. An extracted foot-point at a pixel $p(x, y)$ on a 2-D image is transformed to 3-D coordination $P(X, Y, Z)$ in the world coordination. While the transformation from 2-D point to 3-D is

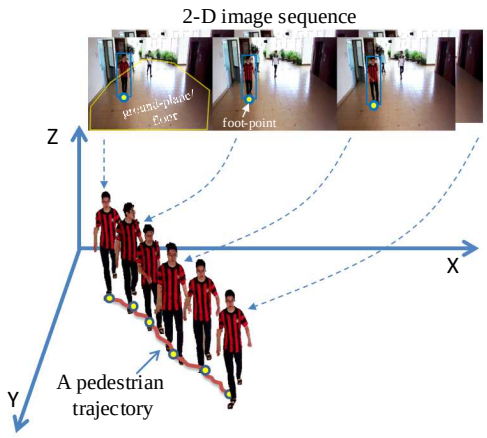


Figure 1: The foot-point definition. Positioning information and a pedestrian trajectory in a real world coordinate is build from the extracted foot-points.

implemented by a standard calibration procedure, extracting the foot-points is a more complicated procedure. We meet some problems such as casting shadow and object occlusion in practice. To resolve these, we firstly perform the background subtraction to separate foreground from background. The foreground may consist of artifacts. They are pruned in a post-processing procedure consisting of filtering and morphological operators. We then put the foot-point candidates to a regression model. This regression model is built through a learning procedure in order to estimate corresponding human height, width (or a ratio height/width) from the detected positions. A correlation between the estimated results and the detected ones is evaluated to eliminate outliers of the detected foot-points. The inlier results are put to a tracking module using a Kalman filter [6]. In this study, we limit the number of people to one in the surveillance camera network. However, the proposed method could be extended to multiple people by applying further techniques such as multiple object tracking and/or re-identification algorithms [13].

2 RELATED WORK

Nowadays, many localization techniques are available in the indoor environments such as: GPS, RFID, Bluetooth, Infrared, ultrasound, Wireless Local Area Network (WLAN) and vision-based techniques [16]. While RFID is more mature technique to use in indoor positioning system [11], its anti-interference ability is poor. This technique therefore is more suitable for the positioning of goods. The Bluetooth techniques in complex environments are less stable and easily disturbed by noise signal [7]. Ultrasonic technology use ultrasonic waves to measure distance between fixed-point station and the mobile target to localize. These methods need to set up dedicated equipments at multiple nodes in the monitoring environments. Comparing with these techniques, the vision-based localization shows significant advantages. It is more natural for human beings, and provides much extracted informative features at a given time. Although this technique usually requires a huge computational costs due to high-dimensional visual data. This issue is able to overcome thanks to recent powerful computation systems.

A vision-based localization can be deployed in two different manner: Fixed and Mobile camera systems. While the second category is often performed by mobile robot for deploying a navigation service (e.g., for supporting visually impaired people), its related works are out-of-the-scope of this study. We focus on the relevant works belonging to the first one which utilizes the fixed surveillance camera in an indoor/outdoor environment. Regarding to this topic, readers can refer a recent review on intelligent multi-camera video surveillance [19]. According to Wang's survey [19], the vision-based localization techniques spread in a board of topics of computer vision such as multi-camera calibration, topology of camera networks, object tracking and activity analysis, and so on. For the first topic, camera calibration is a fundamental problem in computer vision [3]. In the context of surveillance camera networks, this is a crucial procedure to merge field-of-views (for analyzing human activities, trajectories, so on) in different scenes and/or with various camera's configurations. The authors in [17] solve the problem of tracking objects across camera views and computing the homographies between overlapping camera views. In [12], the authors propose an interesting approach of simultaneously estimating the translations between two synchronized but disjoint cameras and the track of a moving object in the 3D space. Both of these approaches require correspondences of tracks observed in different camera views.

Briefly, an initial step a vision-based localization system often deals with extracting moving subjects from scene/backgrounds. To do this, one can utilize the background subtractions [1], motion extraction based on difference frames or optical flow [15] and so on. However, it is noticed that because the inherent ambiguities of lighting conditions, it is difficult to correctly estimate from vision data; and also the scene structure (e.g., depth variations) causes degrading the quality of the target detection. The second step aims to prune the pedestrian detection results. Because the extracted object suffers from many artifacts such as the gradual and sudden illumination changes (such as clouds), high frequency background oscillations such as tree branches or sea waves (outdoor environments). Therefore, the target results must be further processing. The third step of vision-based localization is tracking moving object, particularly, tracking objects across camera views. It is a challenging task because of the following circumstances: complex object shapes, cluttered background, loss of information caused by representing the real world 3-D into a 2-D image sequence, illumination variations, occlusions, shadows, etc [19]. To resolve these issues, prediction tools such as the Kalman or particle filters are commonly used to estimate the object's location where object blob may be occluded and hence the features cannot be extracted [8].

Obviously, selecting the appropriate approaches in each component task will achieve various aspects. In this study, we propose a compact solution to improve accuracy of the vision-based localization in a surveillance camera situation. Key points in the proposed method aim to increase the accuracy of the foot-points detected from the 2-D image sequences. The proposed method therefore resolves both topology of cameras as well as increases accuracy of the tracking object tasks.

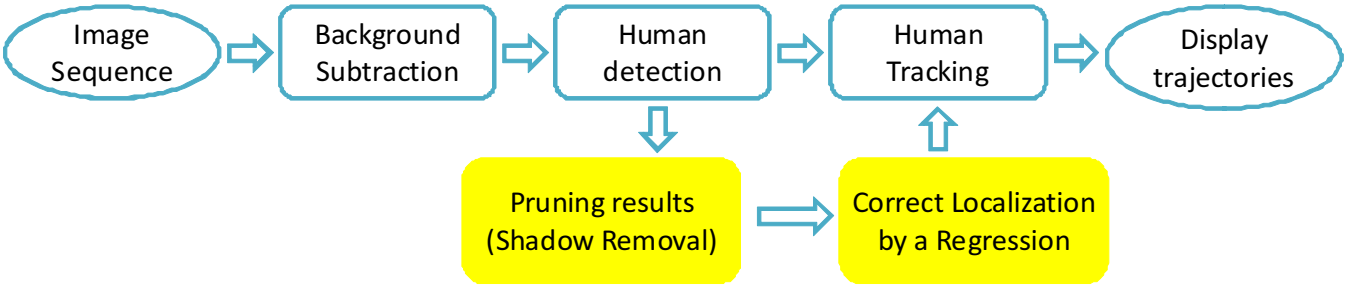


Figure 2: The general proposed framework.

3 PROPOSED METHOD

3.1 The general framework

A general proposed framework is illustrated in Fig. 2. The proposed method starts from extracting a moving subject from an image sequence in a surveillance camera. As a consequence, the foot-point is extracted in each frame. While top-panel consists of common procedures, the proposed method mainly extends current techniques with two new steps marked the yellow boxes. These procedures aim to increase precision of the foot-points detected from 2-D image sequences. To do this, the outliers of moving subjects are eliminated through a correlation evaluation. This evaluation measures a confidence between the detection results and the estimated/predicted ones which yields from a regression model. The inliers are put into a tracking module using a Kalman filter. Because the observations of the Kalman filter consist of mostly inliers, the tracking module is a straightforward implementation. At the final step, fully trajectories across the multi-camera is built and displayed.

We observe that the better Region-Of-Interests (ROIs) and foot-points are extracted, the more accurate of tracking and localization phases are. For each camera, a standard calibration [3] with respect to a 3-D world coordinate system is done beforehand. This calibration procedure estimates both the intrinsic parameters (such as focal length, principal point, skew coefficients and distortion coefficients) which allow to correct lens distortion issues; and extrinsic parameters (such as the camera's center and orientation in the world coordinate system). Therefore, the homography matrix $\boldsymbol{H} \in \boldsymbol{R}^{3\times3}$ is identified. Given a foot point $\boldsymbol{p} = [x, y, 1]^T$ detected from a 2-D image, its corresponding position $\boldsymbol{P} = [X, Y, Z]$ 3-D world coordination is calculated by a transformation $T : \boldsymbol{P} = \boldsymbol{H}\times\boldsymbol{p}$. Because the positioning information is derived from a point where cross-leg touching on the floor, $Z = 0$ is assigned. Moreover, in the proposed method, we assume that the cameras share a ground-plane/floor. It is able to reconstruct the whole trajectory of the human across cameras at non-overlapping regions. This assumption is reasonable because in common surveillance environments (e.g., in a hall-way or large lobby of a building) the ground-plane/floor is often available.

3.2 Background subtraction (BGS) technique

A surveillance camera commonly operates in fixed and/or with minimal change of background. Some accidental situations could change background/scene like door opening; changing position of fixed objects (pots, fire extinguishers). Therefore, extracting a pedestrian could simply base on a background subtraction (BGS) technique. To obtain a trade-off between computational time and detection rate, we utilize the improved Adaptive Gaussian Mixture Model (GMM) algorithm [21]. Each pixel in the image sequence is presented as a Gaussian mixture model. We update the background using a recursive filter. We assume that $\eta(t)$ is a learning rate set for the recursive filter. For each pixel, this parameter is calculated as follows:

$$\mu(t) = (1 - \eta(t)) \times \mu(t-1) + (I(t) - \mu(t-1)) \times \eta(t) \qquad (1)$$

where $I(t)$ is the pixel value in the input image at time t, and $\mu(t-1)$, $\mu(t)$ are mean values calculated at $t-1$ and t. In this work, we set $\eta(t)$ to 0.03. The improved adaptive GMM technique obtains stable result along common trajectories in an indoor surveillance environment. Some illustrations of the BGS results using Adaptive GMM are given in Fig. 3. However, there are critical issues that the BGS results could be very noisy or contaminated by shadow. Fig. 3(c)-(d) illustrates some casting shadows and artifact/noises pixels. Specially, shadow of moving objects normally spreads over the surface that the light is obscured (corner, assigned the two surfaces) and often larger than objects of interest. These issues degrade the quality of further tasks such as detection and tracking. We then conduct following post-processing procedure to remove these noises and shadows.

3.3 Pruning BGS results

To resolve the issues of background subtraction, we deploy a series of morphological operators and filtering techniques. First, a down-scale operator is applied to remove tiny/small noise regions. We then apply a median filter to fill-up black holes or unconnected regions. An up-scale operator is utilized to recover the original size of the image. Effectiveness of this process can be observed in Fig.3(e). Based on these post-processing, the largest blob is identified as the target. However, as shown in Fig.3(e), it still exists casting shadows of the subject. We continue pruning the detection results by applying a shadow removal technique.

To do this, we utilize a density-based score fusion scheme where a feature-based approach is taken into account for removing shadow regions. This technique is proposed in a previous work [14]. To consolidate the paper, we briefly explain the shadow removal technique as follows. First, two types of feature are extracted in the

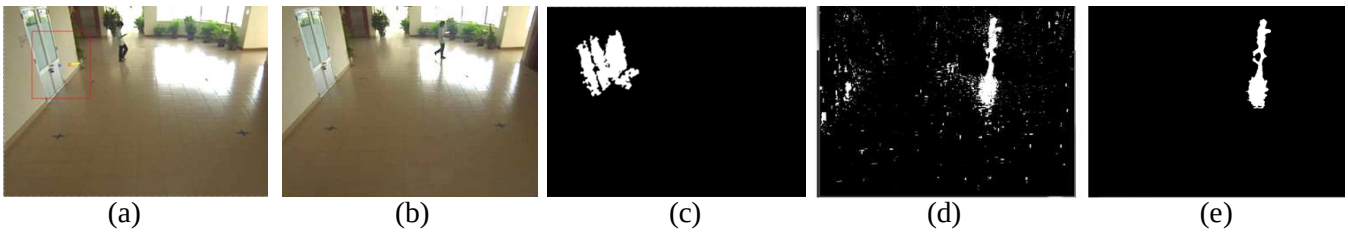


Figure 3: The background subtraction (BGS) results. (a)-(b) Original frames at two different times. (c)-(d) The BGS results of (a) and (b), respectively. It is noticed that a shadow region is observed in (c) whose corresponding region is marked in a red rectangle in original frame (a). (e) Result of (d) by applying a median filter.

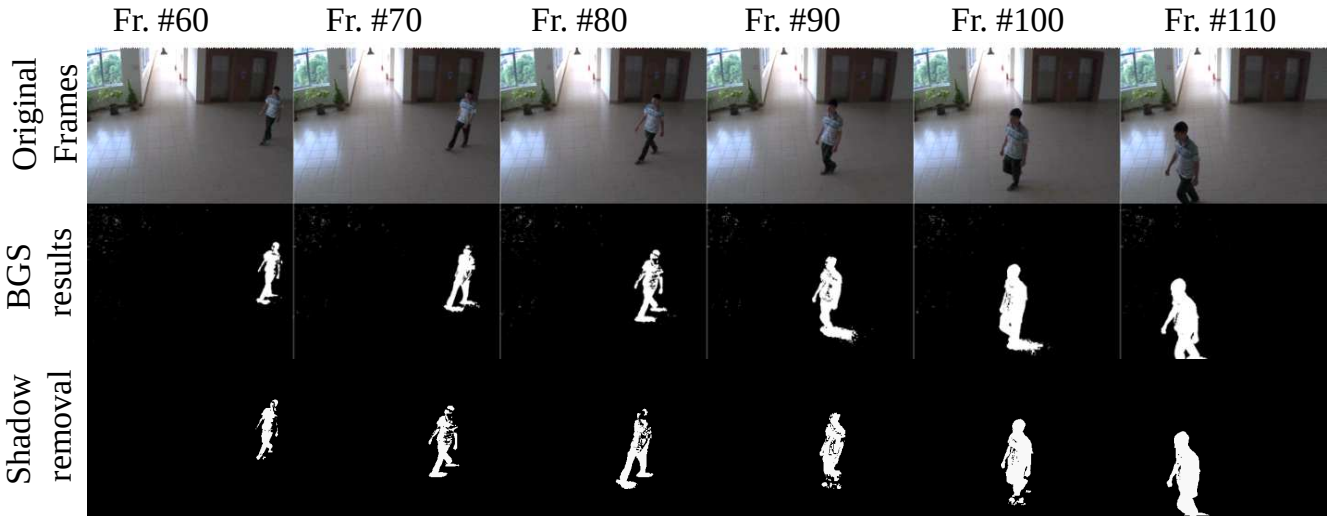


Figure 4: Top row: original frames extracted from an image sequence. Second row: The BGS results using the improved Adaptive GMM technique [21]; Third row: the shadow removal results using a density-based score fusion scheme proposed in [14]

examined shadow region. They are chromaticity-based and physical features [5]. Two likelihoods of the shadow-matching scores are calculated from corresponding features. A likelihood ratio as shadow per non-shadow score is calculated. Probabilities of a shadow or non-shadow pixel are estimated on the basis of approximating distributions of the shadow-matching scores using GMM. Results of the shadow removal is illustrated in Fig. 4. In practice, we observe that current shadow removal technique strongly depends on an heuristic parameter which is a probabilistic threshold deciding a shadow/non-shadow pixel. Fig. 5 clearly illustrates this issue. While Fig. 5(a) consists of the false positive pixels (e.g., the detection result is bigger than the ground-truth), the results in Fig. 5(b) miss true positive pixels. Therefore, a technique examining the correctness of the target/moving subject (or the results after applying shadow removal techniques) is proposed in the next section.

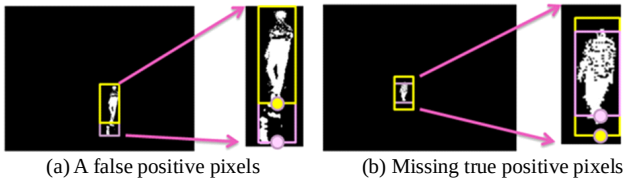
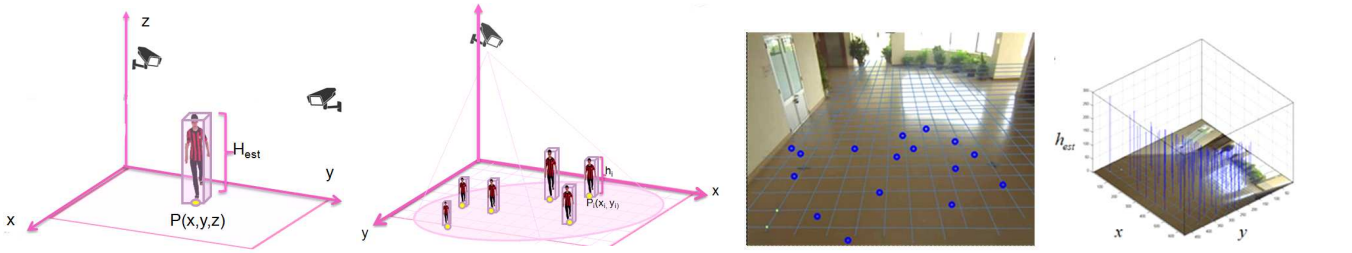


Figure 5: Problems of the shadow removal results. (a) Consisting false positive pixels. (b) Missing true positive pixels. Yellow/Pink box: ground-truth/the shadow removal result

.

3.4 Localization refinement via a regression model

In this section, a learning-based prediction technique is deployed. As the position of a moving object on the ground-plane is measured through their foot-point position. A foot-point itself can infer characteristics such as height, width, or a ratio between width/height of the subject. Theoretically, the further object from a camera is, the smaller object's height and/or its size is and inversely. Estimating such features often utilizes a learned-based model of the regression techniques (e.g., linear, Gaussian Process - GP). A prior knowledge



(a) - Notations using in a GP training

(b) – Setup for a GP training

(c) – Training data for a GP (16 reference points) in practice

(d) The estimated height at every points using a GP trained in (c)

Figure 6: A scheme using a GP to estimate the height of subject depending its position. (a) Notations using in a GP training; (b) A schematic view of the GP training; (c) Training data for a GP with 16 reference points in the practice. (d) The estimated height of subjects in every positions using the GP model trained in (c).

of environment therefore is a good cue to examine the foot-point detection results in Section 3.3.

In signal processing, a Gaussian Process (GP)[10] is defined as a probability distribution over function f such that the set of values of $f(t)$ evaluated at an arbitrary set of points $t_1, ..t_n$ which have joint Gaussian distribution. In view of machine learning, GP can be used to predict the value for unseen points using a model learnt from a training data. In this study, we utilize the GP to predict the features such as height, width of an moving object. The training concept for a GP model consisting of 2-D position $p_i(x_i, y_i)$ and the corresponding height h_i, width w_i of a subject is illustrated in Fig. 6(a)-(b). As shown in Fig. 6(c), the GP is initiated with a number of reference positions. For instance, to train a GP model estimating the object's height, the reference points are assigned by:

$$\begin{aligned} p_1 = (x_1, y_1) &\rightarrow h_1 \\ p_2 = (x_2, y_2) &\rightarrow h_2 \\ &\dots \\ p_n = (x_n, y_n) &\rightarrow h_n \end{aligned}$$

In the estimation phase, given a $p = (x, y)$, object's height H_{est} is estimated by $H_{est} = \varphi(x, y)$. Fig. 6(d) illustrates the estimated height (z-axis) at every positions on a grid of the ground-plane.

To examine the confidence of the detected foot-point in Section 3.3, we evaluate a correlation between H_{est} and H_{det} as follows:

$$\frac{|H_{est} - H_{det}|}{H_{est}} < \tau \tag{2}$$

where:

- H_{est} is the subject's height estimated using the GP model.
- H_{det} is the subject's height detected in Section 3.3 (the results after applying the shadow removal techniques).

The above evaluation means H_{det} is highly correlated or is corrected detection if τ is small enough. Otherwise, the foot-point extracted from the shadow removal result is not confident therefore it can be considered as an outlier. Consequently, we keep only inliers to serve to the tracking algorithm that is explained in the next section.

3.5 The object tracking procedure

Object tracking is an important task of vision-based localization. In this study, we use a conventional Kalman filter method [6] to track a moving subject. A Kalman filter composes of two steps: prediction and correction. The prediction step uses the state models to predict the new state of the variables. The correction step uses the current observation to update the object's state.

By deploying the Kalman Filter, the foot-points is tracked in 2-D image coordination. The state vector of the foot-point at a given time $k-1$ is simply presented by $\boldsymbol{x_{k-1}} = (x, y, u_x, u_y)$ consisting of its coordinates (x, y) and velocity (u_x, u_y) in two directions x and y. We assume that surveillance camera captures at the fixed frame rate and the pedestrian moves at a common speed. Observation vector is defined at each time when the result of correlation evaluation (Eq. (2)) returns the corrected detection. A state transition model $\boldsymbol{F_k}$ allows to predict the state vector $\boldsymbol{z_k} = (x, y)$ at time k:

$$\boldsymbol{x_k} = \boldsymbol{F_k} \times \boldsymbol{x_{k-1}} + w_k \tag{3}$$

Where w_k is process noise, which is assumed to follow a normal distribution with covariance Q_k : $w_k \sim N(0, Q_k)$. Observation model $\boldsymbol{H_k}$ maps the true state space into the observed space:

$$\boldsymbol{z_k} = \boldsymbol{H_k} \times \boldsymbol{x_k} + v_k \tag{4}$$

where $\boldsymbol{H} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, v_k is observation noise which is assumed to be zero mean Gaussian white noise with covariance R_k : $v_k \sim N(0, R_k)$.

3.6 Trajectory reconstruction across camera views

After extracting the trajectories on each camera, we aim to reconstruct them across camera views even without overlapping field of views. To do this, the most important point is to find an appropriate transformation between a view of a reference camera and others. Commonly, if cameras are already calibrated with a single 3D world coordinate system, the transformation can be computed in a straightforward way. Otherwise, it has to be resolved by correspondence-based approaches [2].

(a) (b) (c)

Figure 7: Calibration using the ground-plane information. (a)-(b) Field-of-view of two different cameras. 4 references points are marked in red, blue, yellow, green circles. The grid (red-points) of ground-plane is constructed using a bird-eye view projection. (c)- The point correspondences from the views of two cameras.

Table 1: Characteristics of the evaluation sequences

SeqID	Camera	Number of frames	Length of trajectory (meter)	Lighting condition
Seq#1	Cam1	1540	41.5	natural
Seq#2	Cam2	828	24.5	neon-light
Seq#3	Cam3	946	34.8	natural

In this study, we observe that the surveillance cameras share the same a ground-plane. This is an important cue to find a transformation between two cameras. This transformation helps to link trajectories even their appearing/disappearing across independently among the cameras. For each field-of-view's camera, as shown in Fig.7(a)-(b), we mark four reference points (in red, blue, yellow, green circles). Because size of the white-tile is prior known, we generate the grid with red points using a bird-eye's view projection [4]. Similarly, we generate a grid-points with second camera (with very small overlapping field-of-views). Because the fixed coordination is marked in both cameras (yellow coordinations), we find a list of corresponding points between two field-of-views cameras. The links between these correspondences are shown in Fig.7(c)-(d). Based on these matching points, we extract parameters of a perspective transformation of two cameras [4]. The transformation matrix will be used to transform a trajectory from a current view to a reference one.

4 EXPERIMENTAL RESULTS

4.1 Evaluation Setup

For the experimental evaluation, we adopt a setup in a relevant work [13]. The experimental environment is a common public hallway and lobby room on a floor of a building. The map of experimental environment is shown in Fig. 8. There are three parts: 02 hallways and 01 lobby. In this environment, there are four surveillance cameras. The videos are collected by Cam1 in lobby and Cam2 in left hallway. In most of situation, we assume that only one person moving. For a number of pedestrians, the person identifications is specified in advanced. Details of the evaluation sequences are listed in Table 1.

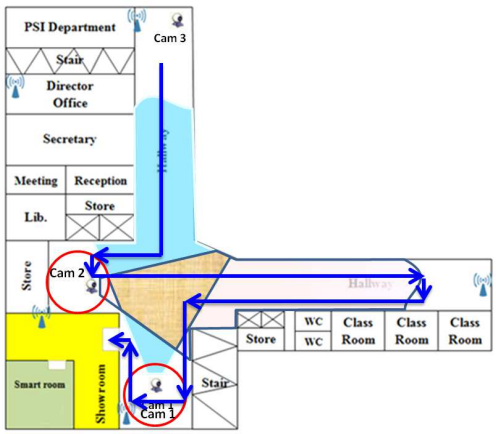


Figure 8: The setup for experimental evaluations.

To measure the general performances, we measure Root-Mean Square Errors (RMSE) of the detected foot-point vs. ground-truth one for each image sequence, that is defined as follows:

$$RMSE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(P_i - \hat{P}_i)^2} \tag{5}$$

where P_i and $\hat{P}_i$ are the detected foot-point and ground-truth, respectively. N is the number of frames of each evaluation sequence. Additionally, to evaluate the advantages of the proposed GP, we utilize two measurements ρ and σ, that measure a relative height between the estimated by GP and ground-truth, as well as the detected foot-point after applying the shadow removal, respectively, as follows:

$$\rho = \frac{1}{n}\sum_{i=1}^{N}\left|\frac{H^i_{est} - H^i_{gt}}{H^i_{gt}}\right| \text{ and } \sigma = \frac{1}{n}\sum_{i=1}^{N}\left|\frac{H^i_{det} - H^i_{gt}}{H^i_{gt}}\right| \tag{6}$$

The proposed techniques are wrapped by C++ and OpenCV 2.4.9 library in a PC Core i5, 4GB RAM.

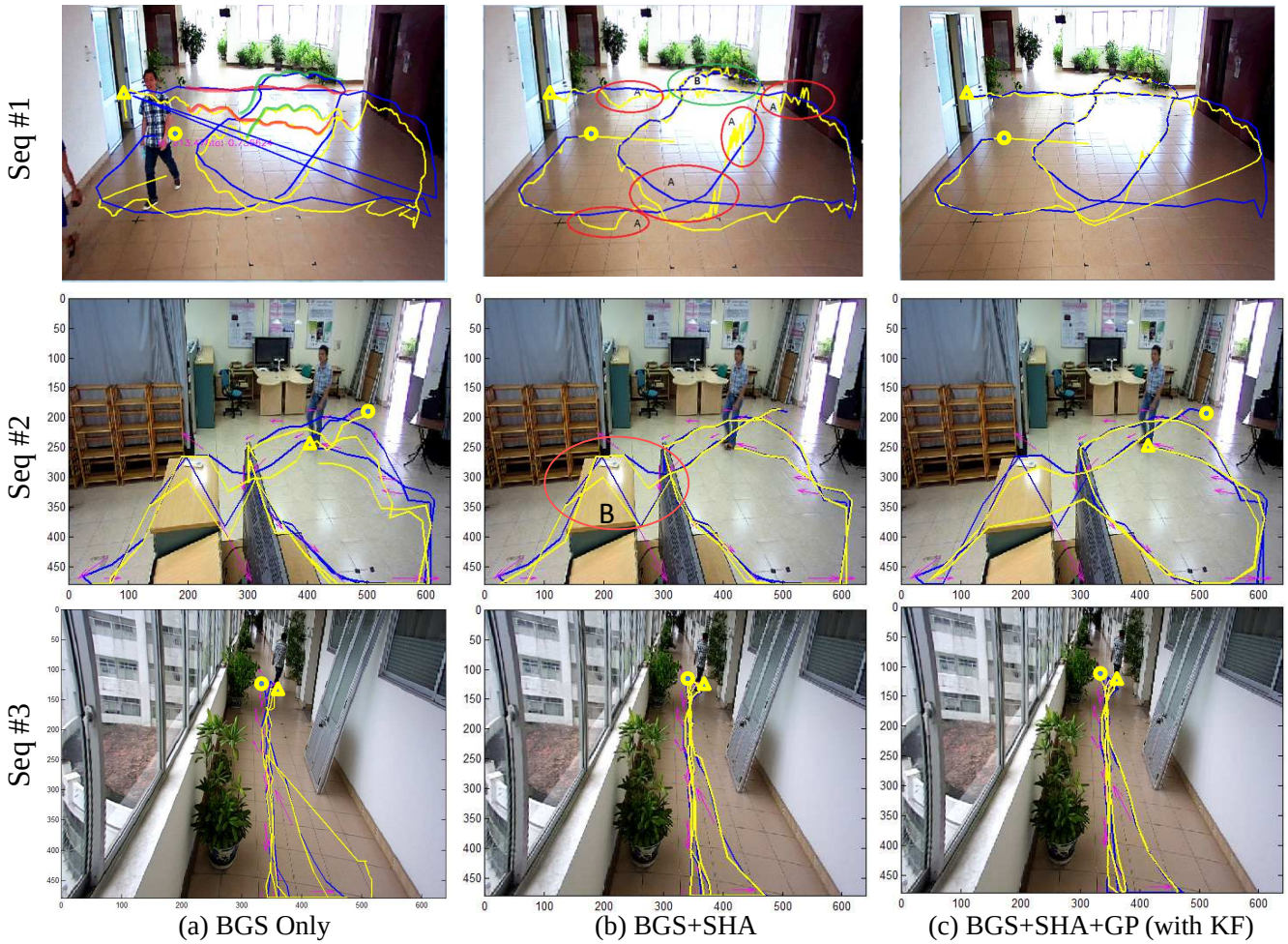
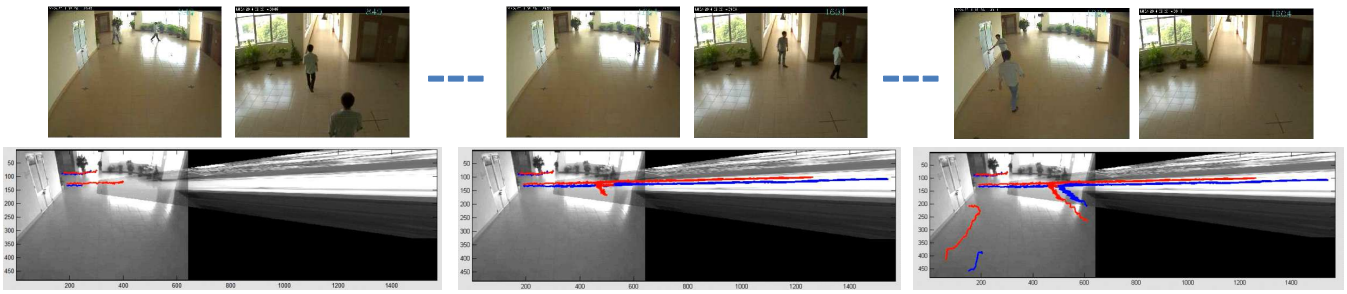


Figure 9: The results of full pedestrian trajectories constructed from the evaluation sequences. (a) Applying on Background Subtraction (BGS). (b) BGS and Shadow Removal (BGS+SHA). (c) BGS+SHA and Gaussian Processing (BGS+SHA+GP). On each panel, the extracted trajectory (yellow line) is overlaid on the ground-truth data (blue line) for visual comparisons. The pink arrows plot the moving directions of the subject with a starting point at yellow-circle and an ending point at yellow-rectangle.

4.2 Evaluation results

First, positioning information of a pedestrian extracted from consecutive image sequence is visually presented as continuous trajectory. Figure 9 draws the recovered trajectories collected by the cameras (as setup in Fig. 8) with different evaluation scenarios. While the Seq#1 consists scenes with heavy shadow appearances (at area nearby the windows regions). The trajectories recovered by the results of only BGS are far from ground-truth. Even applying the shadow removal, there are still existing many noise segments (e.g., marked in red-ellipses - A). By filtering outliers thanks the estimated results of the GP, one can observe the trajectory (yellow line) in BGS+SHA+GP scheme is more stable and tightly with the ground-truth one (blue line). Seq#2's scene is an indoor environment with a stable lighting condition. While BGS+SHA vs. BGS only is significantly improved, the results of BGS+SHA versus BGS+SHA+GP are not much different. But at occlusion area (nearby the cabinet/table - marked in B), thanks to GP, the outliers at these area are eliminated. The recovered trajectory therefore is more reasonable. The trajectory in Seq#3 is simply a straight forward and backward trajectory in several times. The BGS+SHA+GP scheme again confirms that a stable result is achievable.

For the quantitative evaluations, RMSE, ρ, σ are statistically measured. While RMSE measures errors (in pixels) of the positioning results, ρ and σ are to measure the errors in estimating a subject's height. Therefore, ρ and σ close to zero, better method is. Table 2 reports these indexes with three examining sequences. As expected from the visualization results, BGS+SHA+GP scheme achieves the lowest RMSE for three sequences, but the most significant differences can be observed between Only BGS vs. BGS+ SHA schemes. Moreover, as shown in Table 2, the evaluation results of adding SHA and GP schemes return smaller ρ (comparing with σ) for all sequences. In these evaluations, threshold τ in Eq. (2) is a pre-determined value ($\tau = 0.5$). This optimal τ is to preserve an enough number of tracked points that input to the Kalman filter.



(a) (b) (c)

Figure 10: Full trajectories are recovered from view-points of two cameras (Cam1 and Cam2). Top panel: Original frames; Bottom panel: the merging trajectories. (a) The moving subjects are observed by both cameras; (b) Only Cam2. (c) Only Cam1

.

Table 2: The statistical measurements (mean ± std) of the quantitative evaluations

Stat.	Only BGS		BGS+SHA		BGS+SHA+GP	
	RMSE	σ	*RMSE*	ρ	*RMSE*	ρ
Seq#1	59.2±45.1	2.3±0.5	53.5±47.9	1.6±0.5	35.8±34.2	0.5±0.5
Seq#2	22.2±16.6	4.3±2.3	15.0±6.3	3.0±1.9	8.4±5.1	2.5±1.5
Seq#3	55.3±110.1	1.5±0.2	13.6±3.0	0.5±0.0	12.9±1.5	0.5±0.0

4.3 Display the full trajectories

Utilizing the proposed technique, the trajectory on each camera is recovered. As a result, a full trajectory in a surveillance cameras is reconstructed. To do this, a transformation matrix that is described in Section 3.6 is applied to transform collected images from the Cam#2 to Cam#1 for references. In Fig. 10, we show some instances of two trajectories across two cameras. In this example, the full trajectories of the pedestrians in the experimental environments with the stable results are displayed. It is noticed that in this study, the person re-identification tasks are out-of-the scope. In the example shown in Fig.10, we manually marked person ID in advanced therefore the trajectory of each pedestrian is specified.

5 CONCLUSION

In this paper, we proposed a compact solution to estimate the pedestrian localization in a surveillance camera network. The proposed method utilizes simple but valuable cues such as a ground-plane in the scenes, human standing on the floor. Such types of cue appear commonly in public areas. The proposed method tackled advantages of the learning-based estimation to rectify the foot-point detection results. In additional, the trajectories are able to be constructed across camera views event without overlapping regions. The proposed techniques were evaluated in several evaluation scenarios. The experimental evaluations confirmed that the pedestrian positions achieved the high accurate and stable results. The full trajectories can be recovered without any limitations. In the future, we continue to evaluate and improve the proposed methods so that it adapts with tracking multiple users as well as the re-identification tasks.

REFERENCES

[1] Y. Benezeth, P.-M. Jodoin, B. Emile, H. Laurent, and C. Rosenberger. 2010. Comparative study of background subtraction algorithms. *J. Electron. Imaging* 19 (March 2010), 033003–033003–12. Issue 3.

[2] T.J. Ellis, D. Makris, and J. Black. 2003. Learning a multicamera topology.. In *in Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. 165–171.

[3] O. Faugeras. 1993. *Three Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, New York.

[4] R. Hartley and A. Zisserman. 2003. *Multiple View Geometry in Computer Vision,*. Cambridge University Press, London.

[5] J.-B. Huang and C.-S. Chen. 2009. Moving cast shadow detection using physics-based features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2310–2317.

[6] R.E. Kalman. 1960. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*. 82 (1960). Issue 35.

[7] P. Kriz, F. Maly, and T. Kozel. 2016. Improving Indoor Localization Using Bluetooth Low Energy Beacons. *Mobile Information Systems* 2016 (March 2016), 1–11.

[8] M. Liem and D. M. Gavrila. 2009. Multi-person tracking with overlapping cameras in complex, dynamic environments. In *Proceedings of British Machine Vision Conference*. BMVC, 1–10.

[9] C. Loy, T. Xiang, and S. Gong. 2009. Multi-camera activity correlation analysis.. In *IEEE Internat. Conf. Computer Vision and Pattern Recognition*. IEEE.

[10] S. Matthias. 2004. Gaussian Processes for Machine Learning. *International Journal of Neural Systems*. 14 (2004), 69–104. Issue 2.

[11] L. M. Ni, D. Zhang, and M. R. Souryal. 2011. RFID-based localization and tracking technologies. *IEEE Wireless Communications* 18, 2 (April 2011), 45–51.

[12] R. Pflugfelder and H. Bischof. 2010. Localization and trajectory reconstruction in surveillance cameras with nonoverlapping views. *IEEE Trans. Pattern Anal. Machine Intell.* 32 (2010), 709–721.

[13] T. T. T. Pham, T.L. Le, H. Vu, T. K. Dao, and V. T. Nguyen. 2017. Fully-automated person re-identification in multi-camera surveillance system with a robust kernel descriptor and effective shadow removal method. *Image Vision Comput.* 49 (2017), 44–62.

[14] T. T. T. Pham, H. Vu, and A. T. Pham. 2015. A Robust Shadow Removal Technique Applying For Person Localization in a Surveillance Environment. In *Proceedings of the Sixth Int. Symp. on Infor. and Comm. Tech.* ACM, 268–275.

[15] M. Piccardi. 2004. Background subtraction techniques: a review. In *IEEE Int. Conf. Syst. Man Cybern.*, Vol. 4. IEEE, 3099–3104.

[16] J. A. Puertolas-Montanes, A. Mendoza-Rodriguez, and I. Sanz-Prieto. 2013. Smart Indoor Positioning/Location and Navigation: A Lightweight Approach. *Int. J. Interact. Multimed. Artif. Intell.* 2, 2 (2013), 43–50.

[17] C. Stauffer and K. Tieu. 2003. Automated multi-camera planar tracking correspondence modeling.. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*. IEEE.

[18] T.H. Tran, T.L. Le, V.N. Hoang, and H. Vu. 2017. Continuous detection of human fall using multimodal features from Kinect sensors in scalable environment. *Computer Methods and Programs in Biomedicine* 146 (2017), 151 – 165.

[19] X. Wang. 2013. Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters* 34, 1 (2013), 3 – 19.

[20] W. Xiaogang, M. K. Teck, and Ng.Gee-Wah. 2011. Trajectory Analysis and Semantic Region Modeling Using Nonparametric Hierarchical Bayesian Models. *International Journal of Computer Vision* 95, 3 (Dec 2011), 287–312.

[21] Z. Zivkovic. 2004. Improved adaptive Gaussian mixture model for background subtraction. In *Proceedings of the 17th ICPR 2004*, Vol. 2. 28–31.