Person Re-Identification for Non-overlapping Cameras in Multimodal Person Localization

Thi Thanh Thuy Pham*†, Thi-Lan Le*, Trung-Kien Dao*
*International Research Institute MICA, HUST - CNRSUMI-2954 - GRENOBLE INP, Vietnam
†Faculty of Information Technology, University of Technology and Logistics, Bacninh, Vietnam
Email: {Thanh-Thuy.Pham;Thi-Lan.Le;Trung-Kien.Dao}@mica.edu.vn

Abstract—**Person re-identification is a crucial step in an indoor human localization system. It is a problem of person identity association at different locations and times. This paper presents a method for person re-identification in the context of multi-modal person localization using WiFi and camera. From the human region of interest determined by human detection, our method builds a visual signature of the person based on kernel descriptor and performs person re-identification by applying ranking Support Vector Machine. The evaluation results on several benchmark datasets as well as our dataset built in the context of multimodal person localization confirm the robustness of the proposed method.**

Keywords–Multimodal person localization; Person re-identification; Human detection.

I. Introduction

Person re-identification (Re-ID) and positioning are two key problems in a typical human localization system. In case of multi-object localization, we need to identify the person who is localized, therefore, we know the determined positions belong to which objects. Person Re-ID in camera network is a hard problem and has increasingly attracted many researchers. Three basic steps need to be done for vision-based person Re-ID problem. Human detection in consecutive frames is firstly executed, then feature extraction within the detected regions and feature descriptor is generated, finally object matching is done for Re-ID. Each step has its own challenges and these strongly affect to the system performance. In general, they include (1) illumination conditions that are different by time and space; (2) pose, scale and appearance variation of person at distinctive camera FOVs (Fields of View). This is considered as the most challenging, because human appearance features are mainly used in human re-identification systems; (3) occlusions in which people are obscured by each other or obstacles in the environment; (4) Re-ID scenarios involving closed set Re-ID (the identified objects are included in both gallery and probe sets) or open set Re-ID (the objects may not be contained in the gallery set).

Many approaches are proposed for vision-based person Re-ID problem, however, most of them are oriented to (1) build a distinctive feature descriptor for each object and then apply an effective object classifier for that or (2) design potential distance metrics from data. In our previous work [1], we concentrate on establishing a robust feature descriptor that improves the original KDES (Kernel Descriptor) of [2], and applying multi-class SVM as relative ranking for person Re-ID in camera networks. The proposed method is evaluated on two benchmark datasets (CAVIAR4REID and iLIDs) and our own dataset named MICA1. With these datasets, the person ROI (Region Of Interest) is manually determined. We extend our previous work with two main contributions. First, in order to make a fully-automated person Re-ID system, in the detection step, we propose to use a fusion method based on GMM (Gaussian Mixture Model) and HOG (Histogram of Oriented Gradients) along with SVM (Support Vector Machine). Second, we evaluate the proposed method on a dataset which is built in the context of multi-modal person localization, with both cases of automatic and manual person detection are considered.

The rest of the paper is organized as follow. In Section II, the related works on vision-based human Re-ID are presented. Section III indicates a combined system of WiFi and visual signals for human localization, in which appearance-based person Re-ID problem in camera network is solved by improved KDES. Some experimental results on benchmark datasets and our dataset are shown in Section IV. Conclusion and future directions will be finally denoted.

II. Related work

Design of a robust person descriptor is the most decisive step for vision-based person Re-ID problem. Many kinds of features are utilized for this, in which human body appearance is the simplest and the most popular one. Color, texture, and shape are features that can be extracted for human appearance. In [3][4], color histogram is used for feature descriptor. There are two ways to represent the image of detected people with color histogram: global color histogram and local color histogram. A single histogram is used in the first method for the whole image, while in the second way, image is divided into some parts and concatenating the part-based color histograms is done to give a final result. Most reported person Re-ID works pay attention on the second solution, such as in [5], a weighted color histogram derived from MSCR (Maximally Stable Colour Regions) and structured patches are combined for visual description. In [6][7], color histogram on different color models is calculated and syndicated with texture features to make person descriptor more robust. Shape features are also extracted for appearance model. However, they are unstable because of non-rigid objects as people; so, in [8], color and texture features are associated with shape feature to enhance the effectiveness of person descriptor. Local region descriptors, such as SIFT (Scale-Invariant Feature Transform), SURF (Speeded Up Robust Features) and GLOH (Gradient Location and Orientation Histogram) are evaluated in [9] for person Re-ID in image sequences. The results show that GLOH and SIFT outperform both shape context and SURF descriptor. Additionally, a large number of visual features are exploited for person Re-ID problem, such as Haar-like features, HOG

(Histogram of Oriented Gradients), edges, covariance, interest points, etc.

The next step in human Re-ID process is classification, with two scenarios of single-shot and multi-shot being reported. The first case is simpler with one-to-one matching between a pair of probe and gallery image for each person, whereas in the second scenario, each object has multiple images, either in the gallery or the probe set. In general, the purpose of classification in person Re-ID is finding out the most similar candidate for a target or ranking the candidates based on a standard distance minimization strategy, which is known as distance metric. This metric can be chosen independently (non-learning based method) [10] or learned from the data (learning-based method) [11] in order to minimize intra-class variation whilst maximize extra-class variation. They typically include histogram-based Bhattacharyya distance, K Nearest Neighbor classifiers, L1-Norm, diffusion distance [12]. Additionally, some later proposed methods, such as LMNN-R (Large Margin Nearest Neighbor) distance metric in [13] or PRDL (Probabilistic Relative Distance Learning) in [14] are more robust.

To get an ID ranking list, distance scores between true and wrong matches can be compared directly or relatively (ranking the scores that show the correspondence of each likely match to the probe image). The relative ranking treated by either Boosting as RankBoost in [15] or kernel-based learning, such as RankSVM [16], primal-based RankSVM [17] or Ensemble RankSVM [7].

III. Proposed system

In this section, we propose a fused system of WiFi and camera for person localization. In this system, the important role of person Re-ID based on the human appearance images extracted from a combined detector of adaptive GMM and HOG-SMV is analyzed. Thence, a new appearance-based model based on KDES is build for each individual and based on this, multi-class SVM is applied for person classification.

A. Overview of multimodal person localization system

Object localization is known as a problem of determining the object position in the environment. For each user in multi-user localization system, two problems of positioning (where the user is) and identifying (who the user is) must be solved simultaneously. A general diagram for object localization is illustrated in Figure 1. In this figure, the input cues can come from different sensors, such as optical, radio frequency, ultra sound, inertia, DC Electromagnetic sensors, etc. From the input cues, localization and Re-ID are executed simultaneously to give the output for object position and ID. Multimodal object localization is defined as a problem of multi-cue combination for input or fusion of different positioning methods. As proven by Vinyals et al. [18] and Dao et al. [19], compounding of different models gives better positioning results than applying a single model. Teixeira et al. [20] proposed to use the motion signature taken from wearable accelerometer for identifying people in camera network.

Our research aims at developing a multimodal person localization system by using both WiFi and camera systems. This offers some benefits in comparison with single-method systems. (1) System setting cost is limited because of available WiFi infrastructure and uncrowded-deployed cameras. (2) Positioning range is easily broaden by simply adding more APs (Access Point) in the environment. (3) Computational expense is much lower for WiFi-based than vision-based positioning system. (4) The positioning accuracy is provided in accordance with the application-specific demands. Although the camera-based system brings more impressed positioning results, but not every where in building needs high localizing accuracy. (5) Sampling frequency is improved for the WiFi-based system, because it has lower sampling rate (about one signal measure per second) than vision-based system (approximately 15 fps). (6) The information for person Re-ID becomes richer. One object can be identified simultaneously by both WiFi and camera systems. These ID cues can be used in the way of supporting one another in multi-model object localization system. For example, at a certain time, one object is localized and identified by WiFi system, with the position of P_{WiFi} and the identity of ID_{WiFi} (the MAC address of mobile device), respectively. At the same time, this object is also determined by P_{cam} and ID_{cam} from camera system. However, P_{WiFi} is not as accurate as P_{cam}, whilst ID_{WiFi} is clearer than ID_{cam}. Therefore, by using both of these systems, the object can be localized by P_{cam} and identified by ID_{WiFi}.

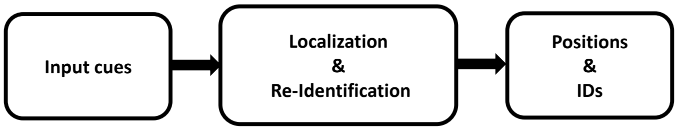


Figure 1. Flowchart of object localization system.

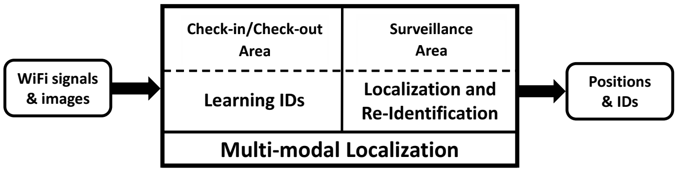


Figure 2. Multimodal localization system fusing WiFi signals and images.

Figure 2 shows a framework for our multi-model human localization system using both WiFi signals and camera network. The framework indicates that the proposed system is implemented in two subregions of the whole positioning area: check-in/check-out region and surveillance region. In the first region, learning ID cues is executed. Person holding a WiFi-integrated device will one by one come in and come out of the first region. At the entrance of the first region, the person's ID will be learned individually by the images captured from cameras and MAC address of WiFi-enable equipment held by that person. One camera, which is in front door of check-in gate, captures human face and then a face recognition program is executed. Another camera acquires human body images at different poses and learning phase of appearance-based ID is done for each person. In short, in the first region, we get three types of signature for each person (N_i): face-based ID (ID^i_F), WiFi-based ID (ID^i_{WF}), and appearance-based ID (ID^i_{Apr}). Depending on different circumstances, we can map among signatures of (ID^i_F, ID^i_{WF}), (ID^i_F, ID^i_{Apr}), (ID^i_{Apr}, ID^i_{WF}) and utilize them for person localization and identification in the surveillance region. The user will end up his route at the exit gate and he will be checked out by other camera. This camera acquires human face for person Re-ID, and based on this, the user will be removed from the localization system. By using check-in/check-out region, we can (1) control the human appearance changes (the difference in cloth colors) at each time people come in the positioning area, (2) decrease the computing cost by eliminating the checked-out users from the system, (3) map between different ID cues for the same person.

In the surveillance region, two problems of person localization and Re-ID will be solved concurrently by combining visual and WiFi information. Figure 3 demonstrates a surveillance region which contains WiFi range and camera FOVs. In this region, the WiFi range covers some visual ranges (the camera FOVs: FOV of C$_1$, FOV of C$_2$,.., FOV of C$_n$). This means the user always move within WiFi range but switch from one camera FOV to others.

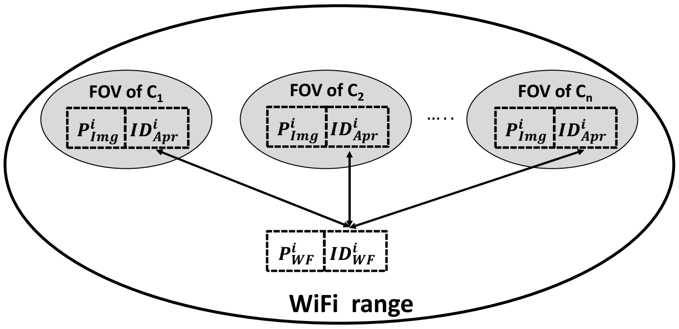


Figure 3. Surveillance region with WiFi range and disjoint cameras' FOVs.

In each camera FOV and for an individual, we calculate image-based and WiFi-based positions (P^i_{img}, P^i_{WF}) and ID^i_{Apr}. From ID^i_{Apr}, we know ID^i_{WF} correspondingly by ID mapping result taken from the first region. Outside the camera FOV, there only exits the information of P^i_{WiFi}, ID^i_{WF}, and ID^i_{Apr}, respectively. When people switch from one camera FOV to others, their positions and IDs will be updated in the WiFi-available region. The localization accuracy then be tuned by combination of WiFi-based and vision-based systems.

From the above analysis, we see that finding ID^i_{Apr} plays a key role in the proposed multimodal person localization system. It is used to link the object trajectories from one camera range to others through the intermediate positioning range of WiFi. Therefore, ID^i_{Apr} must be shown at each frame captured from different cameras in the surveillance area. That means the appearance-based person Re-ID problem needs to be solved. In this circumstance, it belongs to multi-shot person Re-ID problem, with multiple images for each detected person at different resolutions, lighting conditions, and poses are processed.

B. Vision-based person re-identification

1) The system overview: The flowchart of vision-based person Re-ID system is illustrated in Figure 4. It includes three stages of (1) person detection, (2) feature extraction, and (3) classification. Human ROIs (Region of Interest) are extracted from the first stage and based on this, in the second stage, feature extraction is done to build a feature descriptor for each individual. A classifier is then applied to learn the person model and predict the corresponding ID.

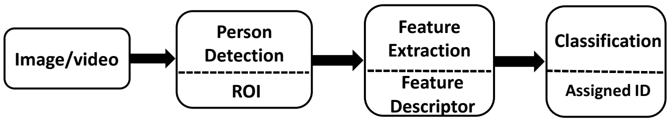


Figure 4. A diagram of vision-based person Re-ID system.

In the first stage, a popular background subtraction technique of adaptive GMM [21] and HOG-SVM [22] are combined for human detection. GMM is suitable method for real-time applications, but in case people are in close proximity or occlusion, it can not give the separated human ROIs for each individual. This can be partially solved by applying a HOG detector. However, the computation time for HOG-SVM is higher than most other background subtraction methods. The fusion of these two methods can help to achieve both accuracy and real-time demands for human detection, as proved in [23].

In adaptive GMM method, K Gaussian distributions are used to model the recent history $\{X_1, .., X_t\}$ of each pixel X. The probability of the pixel value is then defined by a sum of weighted Gaussian distributions as follows:

$$P(X_t) = \sum_{i=1}^{K} w_{i,t} * g(X_t|\mu_{i,t}, \Sigma_{i,t}) \tag{1}$$

where K is the number of distributions, $w_{i,t}$ are the mixture weights, $g(X_t|\mu_{i,t}, \Sigma_{i,t})$ are the component Gaussian densities, with mean vector $\mu_{i,t}$ and covariance matrix $\Sigma_{i,t}$ of the ith Gaussian component in the mixture at time t. When the new pixels are matched with the model, the mean and the variance values are updated using:

$$\sigma_t^2 = (1-\eta)\sigma_{t-1}^2 + \eta(X_t - \mu_t)^T(X_t - \mu_t) \tag{2}$$

$$\mu_t = (1-\eta)\mu_{t-1} + \eta X_t \tag{3}$$

where σ_{t-1}, μ_{t-1} are the previous mean and variance of the matched Gaussian, X_t is new pixel value and η is a learning rate. Conversely, when no ones are matched, the least probable component of the mixture is replaced by a new one modeling the incoming pixel.

In the proposed fusion method of human detection, from an input frame, if the human region is extracted by HOG-SVM, this region will be taken for final human detection result, otherwise the result of adaptive GMM is used. In case of having the human detection results from both methods, the area of the detected overlapping region between them will be checked. If it is bigger than a threshold τ, then the matching region is found and HOG-based result will be chosen as the final one for human detection, whilst a false positive of HOG-SVM or adaptive GMM is detected.

The second stage is done based on the results of human detection in the first stage. In the second stage, the features are extracted from the human ROIs and feature descriptors are created for each individual. The details of a new person appearance representation model based on KDES of [2] will be shown in the next subsection.

2) KDES-based person representation: The basic idea of the representation based on kernel methods is to compute the approximate explicit feature map for kernel match function (see Figure 5). In other words, the kernel match functions are approximated by explicit feature maps. This enables efficient learning methods for linear kernels to be applied to the non-linear kernels. This approach was introduced in [2]. Given a match kernel function $k(x, y)$, the feature map $\varphi(.)$ for the kernel $k(x, y)$ is a function mapping a vector **x** into a feature space so that $k(x, y) = \varphi(x)^\top \varphi(y)$. Given a set of basis vectors $B = \{\varphi(v_i)\}_{i=1}^{D}$, the approximation of feature map $\varphi(x)$ can be:

$$\phi(x) = Gk_B(x) \tag{4}$$

where G is defined by: $G^\top G = K_{BB}^{-1}$ and K_{BB} is a $D \times D$ matrix with $\{K_{BB}\}_{ij} = k(v_i, v_j)$, and k_B is a $D \times 1$ vector with $\{k_B\}_i = k(x, v_i)$.

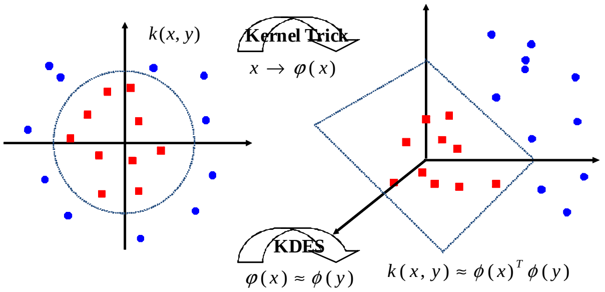


Figure 5. The basic idea of representation based on kernel methods.

Like in [2], in this work, three match kernel functions of gradient, color and shape are built from different pixel attributes of gradient, color and local binary pattern (LBP). For each match kernel, feature extraction is done at three levels: pixel, patch and whole detected human region.

First, gradient match kernel K_g is computed from three kernels of normalized gradient magnitude kernel $k_{\widetilde{m}}$, normalized orientation kernel k_o, and position kernel k_p. At pixel level, a normalized gradient vector is constructed for each pixel z. It is defined by magnitude $m(z)$ and normalized orientation $\omega(z) = \theta(z) - \overline{\theta}(P)$, where $\theta(z)$ is orientation of gradient vector at the pixel z, and $\overline{\theta}(P)$ is the dominant orientation of the patch P that is the vector sum of all the gradient vectors in the patch. This normalization is unlike the approach in [2] which presents $\theta(z)$ as orientation of gradient vector, and it will make patch-level features invariant to rotation. In practice, the normalized orientation of a gradient vector is:

$$\widetilde{\omega}(z) = [sin(\omega(z)) \; cos(\omega(z))] \tag{5}$$

At patch level, the image with different resolutions will be divided into a grid of a fix number of cells, instead of size-fixed cells as in [2]. A patch is then set by 2×2 cells and two adjacent patches along x axis or y axis are overlapped at two cells. This division results in size-adaptive patches to the different image resolutions, and nearly the same feature vectors for the scale-varied images of intraclass are created (see Figure 6). In this work, this technique is utilized for KDES extraction because of a large variation in human size caused by different distances from pedestrian to the stationary camera. From this remark, the gradient match kernel K_g is constructed as follows:

$$K_g(P, Q) = \sum_{z \in P} \sum_{z' \in Q} k_{\widetilde{m}}(z, z') k_o(\widetilde{\omega}(z), \widetilde{\omega}(z')) k_p(z, z') \tag{6}$$

where P and Q are the patches of two different images needed to measure the similarity, z and z' denote the 2D positions of a pixel in the image patches P and Q, respectively, $k_{\widetilde{m}}(z, z') = \widetilde{m}(z)\widetilde{m}(z')$ is a positive definite kernel with the normalized gradient magnitude $\widetilde{m}(z) = m(z)/\sqrt{\sum_{z \in P} m(z)^2 + \epsilon_g}$, ϵ_g is a small constant, $k_o(\widetilde{\omega}(z), \widetilde{\omega}(z')) = exp(-\gamma_o \|\widetilde{\omega}(z) - \widetilde{\omega}(z')\|^2)$ is Gaussian kernel over normalized orientation, $k_p(z, z') = exp(-\gamma_p \|z - z'\|^2)$ is a Gaussian position kernel.

The approximate feature $\widetilde{F}_g(P)$ over the image patch P is constructed in Eq. (7) with normalized gradient magnitude $\widetilde{m}(z)$ and the feature maps of $\varphi_o(.)$ and $\varphi_p(.)$ for the gradient orientation kernel k_o and position kernel k_p, respectively.

$$\widetilde{F}_g(P) = \sum_{z \in P} \widetilde{m}(z) \phi_o(\widetilde{\omega}(z)) \otimes \phi_p(z) \tag{7}$$

where $\otimes$ is the Kronecker product, $\phi_o(\widetilde{\omega}(z))$ and $\phi_p(z)$ are approximate feature maps (Eq. (4)) for the kernel k_o and k_p, respectively.

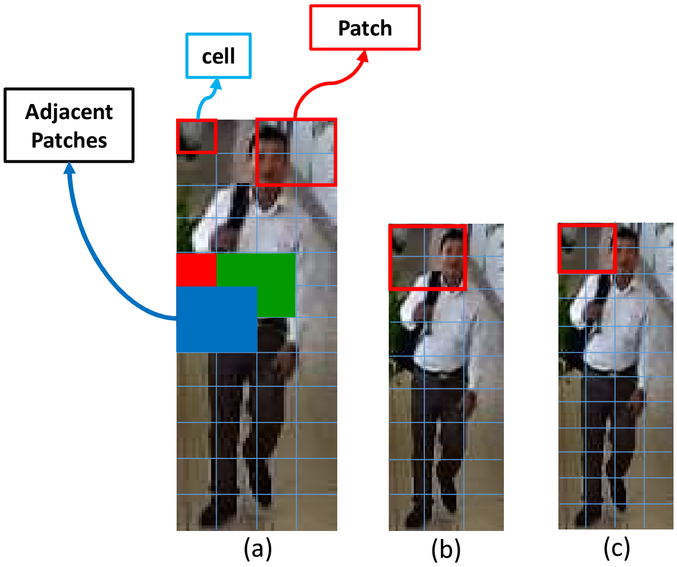


Figure 6. Illustration of size-adaptive patches (a, c) and size-fixed patches (a, b) which is mentioned in [2].

Second, the color match kernel K_c is computed over color pixels (RGB values) at position z as in [2] as follows:

$$K_c(P, Q) = \sum_{z \in P} \sum_{z' \in Q} k_c(c(z), c(z')) k_p(z, z') \tag{8}$$

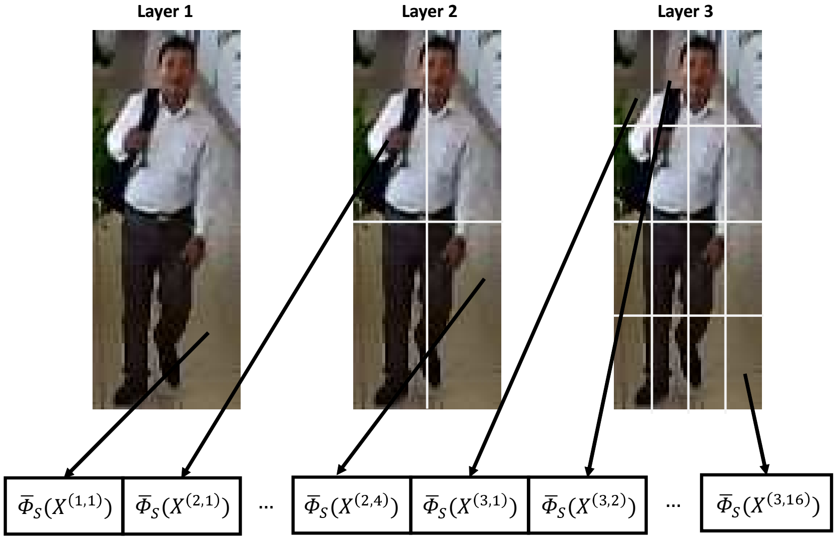

Layer 1
Layer 2
Layer 3
$\bar{\Phi}_S(X^{(1,1)})$
$\bar{\Phi}_S(X^{(2,1)})$
...
$\bar{\Phi}_S(X^{(2,4)})$
$\bar{\Phi}_S(X^{(3,1)})$
$\bar{\Phi}_S(X^{(3,2)})$
...
$\bar{\Phi}_S(X^{(3,16)})$

Figure 7. Image-level feature vector concatenated by feature vectors of blocks in the pyramid layers.

where $c(z)$ is the color value at pixel z, and the Gaussian color kernel $k_c(c(z), c(z')) = exp(-\gamma_c \|c(z) - c(z')\|^2)$ shows the similarity between two color pixels z and z'.

Similar to the calculation of $\widetilde{F}_g(P)$, the approximate color feature over patch P is defined by:

$$\widetilde{F}_c(P) = \sum_{z \in P} \phi_c(c(z)) \otimes \phi_p(z) \quad (9)$$

Finally, the shape match descriptor K_s is built from local binary pattern (LBP) attributes.

$$K_s(P,Q) = \sum_{z \in P} \sum_{z' \in Q} k_{\widetilde{s}}(z,z') k_b(b(z), b(z')) k_p(z,z') \quad (10)$$

where $k_{\widetilde{s}}(z,z') = \widetilde{s}(z)\widetilde{s}(z')$, and the standard deviation of pixel values in the 3×3 pixel block around z is $\widetilde{s}(z) = s(z)/\sqrt{\sum_{z \in P} s(z)^2 + \epsilon_s}$, and $b(z)$ is a vector of binary codes which label the pixel value differences in a local window around z. The similarity of LBP features is measured by the Gaussian kernel $k_b(b(z), b(z')) = exp(-\gamma_b \|b(z) - b(z')\|^2$.

The shape feature over image patch P is then derived as follows:

$$\widetilde{F}_s(P) = \sum_{z \in P} \widetilde{s}(z) \phi_b(b(z)) \otimes \phi_p(z) \quad (11)$$

The last level is achieved by creating a complete descriptor for the whole image. As in [24], a pyramid structure is used to combine patch features. Given an image, the final representation is built on the basis of features extracted from lower levels using EMK (Efficient Match Kernels) proposed in [2]. First, the feature vector for each cell of the pyramid structure is computed. The final descriptor is then the concatenation of feature vectors of all cells.

Let B be a block that has a set of patch-level features $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$, then the feature map on this set of vectors is defined as:

$$\overline{\phi}_S(X) = \frac{1}{|X|} \sum_{x \in X} \phi(x) \quad (12)$$

where $\phi(x)$ is the approximate feature map defined in Eq. (4) for the kernel $k(x,y)$. The feature vector $\overline{\phi}_S(X)$ on the set of patches is extracted explicitly. Given an image, let L be the number of spatial layers to be considered. In this case, $L = 3$ (see Figure 7). The number of blocks in the l^{th} layer is n_l, $X(l,t)$ is a set of patch-level features that fall within the spatial block (l,t) (the t^{th} block in the l^{th} layer). A patch falls in a block when its centroid belongs to the block. The feature map on the pyramid structure is:

$$\begin{aligned} \overline{\phi}_P(X) \;\; = \;\; & [w^{(1)}\overline{\phi}_S(X^{(1,1)}); ...; w^{(l)}\overline{\phi}_S(X^{(l,t)}); \\ & ...; w^{(L)}\overline{\phi}_S(X^{(L,n_L)})] \end{aligned} \quad (13)$$

In Eq. (13), $w^{(l)} = \frac{\frac{1}{n_l}}{\sum_{l=1}^{L} \frac{1}{n_l}}$ is the weight associated with level l.

The final feature vector is then concatenated from three image-level feature vectors of gradient, color and shape.

Once KDES descriptor is computed, a multi-class SVM is applied to train the model for each person. For each detected instance, a list of ranked objects will be generated on the basis of the class probabilities returned by SVM classification.

IV. Experimental results

This section presents the testing datasets and the comparative results of person Re-ID obtained by using the proposed method and some other state-of-the-art methods. The CMC (Cumulative Match Curve) is employed as the performance evaluation method for person Re-ID problem. The CMC curve represents the expectation of finding correct match in the top n matches.

A. *Testing datasets*

In our experiments, two multi-shot benchmark datasets of CAVIAR4REID and i-LIDS for person Re-ID are chosen for evaluating the proposed KDES descriptor. These datasets contain the manually-extracted human ROIs and most of the state-of-the-art methods for person Re-ID are evaluated on these. In order to evaluate the person Re-ID performance in real scenarios of vision-based human localization system, the automatically-detected human ROIs should be considered. A new dataset with the human ROIs resulted from auto human detector (as mentioned in Section III-B1) and manual cropping is built for the person Re-ID evaluation.

The CAVIAR4REID dataset includes 72 pedestrians, in which 50 of them are captured from two camera views and the remaining 22 from one camera view. i-LIDS dataset contains 119 individuals, with the images captured from multi-camera network. In ETH dataset, multiple images of diverse human appearances are captured by a moving camera.

Concerning to our dataset, we build it for multimodal person localization evaluation. A database for testing appearance-base person Re-ID is also established in this. Figure 8 shows a floor plan of the office building. It is set as the testing environment for our combined person localization system. At the entrance, people hold smart phones or tablets go one by one through the check-in gate, then move inside the surveillance area, and finish their routes by going out check-out gate.

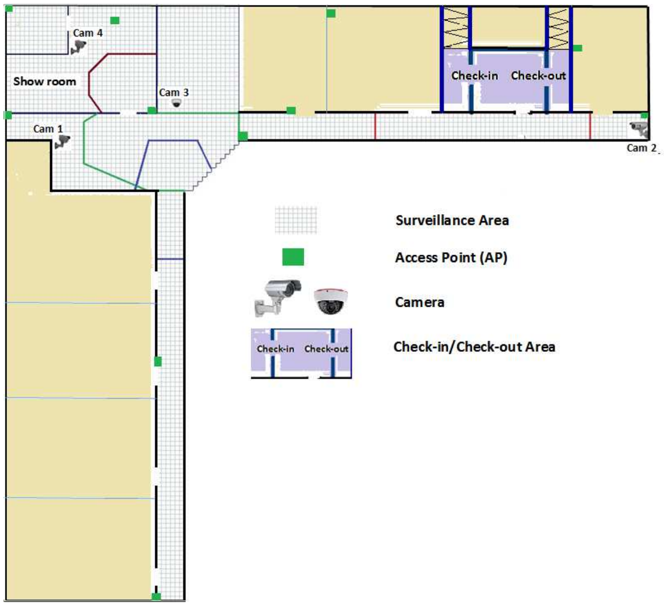


Figure 8. Testing environment.

In the check-in and check-out area, we set three cameras. Two of them are used at the entrance. One camera captures human face in order to check-in user by face recognition. The remaining camera acquires human body images at different poses. This will help the system learn appearance-based signature of the checked-in user. The third camera is used to capture human face at the exit, and based on this, the system will check out or release the user from its process. In the surveillance area, four cameras with non-overlapping FOVs are deployed along the hallway and in a room. People are detected, localized, and re-identified at each frame captured from these cameras. Besides this, 11 APs are established throughout the testing environment. RSSIs (Received Signal Strength Indicators) and the MAC address are consecutively scanned and sent from mobile device to the server to calculate the position and ID of the device holder.

In short, a total of seven AXIS IP cameras and eleven APs are deployed throughout the testing environment. These cameras and APs are fixed at certain distances from the floor ground (about 1.6 m - 2.2 m for cameras and 2 m - 2.8 m for APs). They are configured with static IP addresses. The camera frame rate is set to 20 fps and image resolution is 640x480.

Two scripts are proposed for building the human Re-ID datasets. The first script, namely MICA1 [25], includes 25 people and the second script called MICA2 [25] contains 40 people moving in different routes in the testing environment. Each person spends from 3 to 5 minutes for his route. An approximation of 800 values of RSSIs are scanned, about 3000 frames are captured for each camera in the surveillance area. All acquired frames are processed as real Re-ID scenario of multimodal pedestrian localization system.

In the MICA1 dataset, the human ROIs are manually extracted from video sequences acquired by the cameras. The training images are captured from the entrance camera so that they present the variety of human poses at different viewpoints. The images captured from four cameras (Cam1, Cam2, Cam3, Cam4 in Figure 8) in the surveillance area are processed for testing phase of person Re-ID. Examples in the MICA1 dataset are shown in Figure 9, with the images on the top are used for the training phase of the appearance-based human descriptor and the images for the testing phase are shown at the bottom.

In the MICA2 dataset, the training images are captured from Cam2 and the testing images are acquired from Cam1, Cam3, Cam4. Unlike the MICA1, the training images in the MICA2 dataset are nearly taken from two viewpoints of front and back (see the images on the top of Figure 10). The testing phase is done with the human ROIs extracted both manually and automatically from the images captured by Cam1 and Cam4. The examples of manually-cropped and automatically-detected human ROIs used for testing phase are shown respectively in the middle and at the bottom of Figure 10.

In comparison with other person Re-ID datasets, such as iLIDS, CAVIAR4REID, our datasets have more variations for intra-class images in terms of resolution, illumination, pose, scale, and occlusion. In our datasets, the human ROIs are not only extracted manually, but also automatically from the frames captured by many cameras at distinctive FOVs. Table I shows the summary of these datasets.

B. *Person re-identification results*

The person Re-ID performance of the proposed descriptor is compared with some other state-of-the-art methods in two situations of human detection: manually-cropped and automatically-detected human ROIs.

In the first situation, the benchmark datasets of CAVIAR4REID and i-LIDS are used. These datasets contain the human ROIs extracted manually from the captured frames. In addition, the manually-cropped images in the MICA1 and MICA2 datasets are also utilized in this evaluation. The

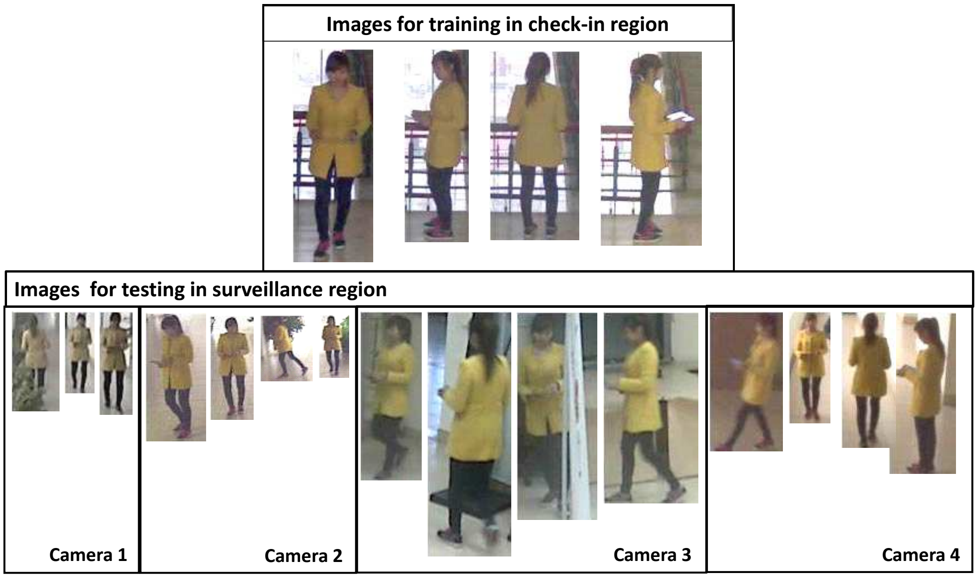

Images for training in check-in region
Images for testing in surveillance region
Camera 1
Camera 2
Camera 3
Camera 4

Figure 9. Examples in the MICA1 dataset. The images on the top are captured from a camera at check-in region and used for training phase. The images at the bottom are the testing images which acquired from 4 other cameras (Cam1, Cam2, Cam3, Cam4) in surveillance region.

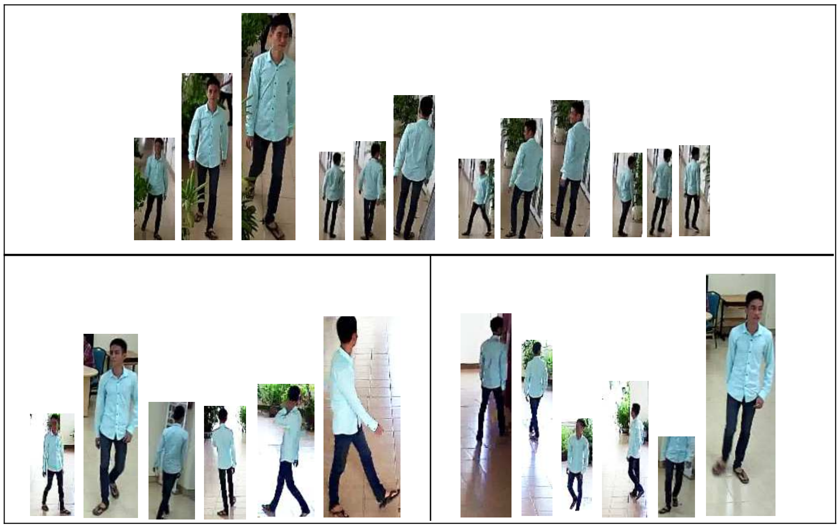


Figure 10. Examples in the MICA2 dataset. The images on the top are captured from Cam2 and used for training phase. Images of human ROI with manual and automatic detection are shown on the left and right groups, respectively.

TABLE I. Datasets for person re-identification testing. In the last column, the number of sign (√) shows the ranking for intra-class variation of the datasets.

Dataset	Release time	# identities	# cameras	Label method	Crop size	Multi-shot	Tracking sequences	Intra-class Variation
iLIDS	2009	69	1	Hand	Vary	Yes	Yes	√
CAVIAR4ReID	2011	72	2	Hand	Vary	Yes	No	√√√
MICA 1, 2	2015	25, 40	5, 3	Hand, Auto	Vary	Yes	Yes	√√√√

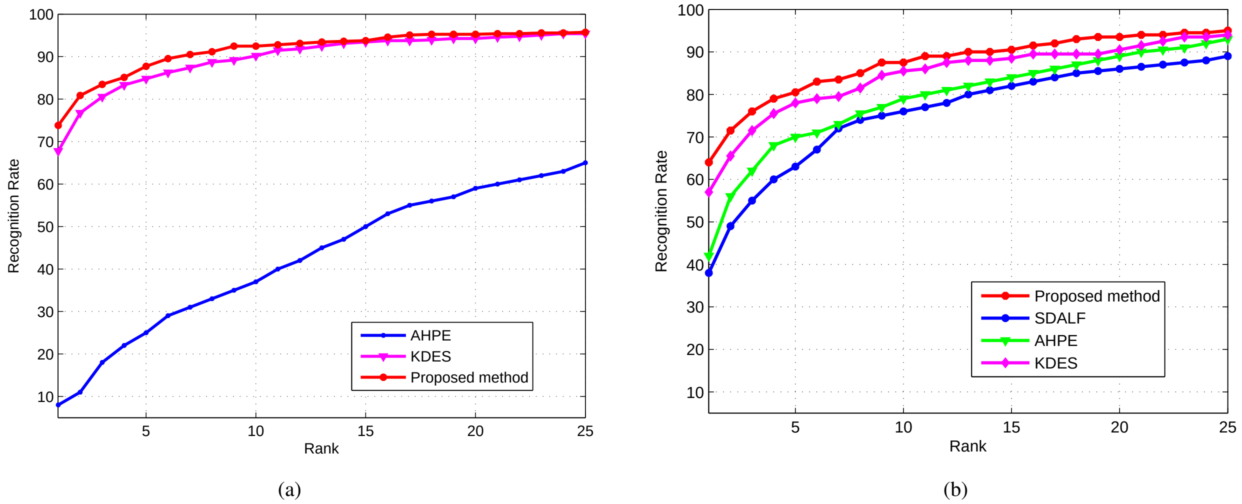


Figure 11. The results of proposed method against AHPE [26], SDALF [27] and KDES [2] on (a) CAVIAR4REID dataset and (b) iLIDS dataset.

automatically-detected human ROIs in the MICA2 dataset are used in the evaluations of the second situation to give the comparative results for person Re-ID in real scenarios of human detection and localization.

The person Re-ID results of the proposed method are compared with the original KDES [2] and other state-of-the-art approaches reported in [26]. In [26], multi-shot datasets of CAVIAR4REID and a modified version of iLIDS are used. The same experimental settings as in [26] are used in this paper for evaluation the performance of person Re-ID. The outperforming results of the proposed method are shown in Figure 11. For CAVIAR4REID dataset, the recognition rate of Rank 1 for AHPE (Asymmetry-based Histogram Plus Epitome) in [26] is much lower than the proposed method. It is only about 8 % compared with 67.76 % of the original KDES in [2] and 73.81 % of the proposed method. However, both KDES and the proposed methods gain nearly the same figures from Rank 13 and backward.

For iLIDS dataset, the gap of Rank 1 between the proposed method with AHPE in [26] or SDALF (Symmetry-Driven Accumulation of Local Features) in [27] is approximately 20 %, and about 7 % with the original KDES in [2]. This gap is slightly decreased for KDES but significantly reduced for AHPE and SDALF after Rank 15.

Other experiments with iLIDs dataset are presented in Figure 12-a in comparison with other methods reported in [28]. In [28], the highest result for Rank 1 belongs to RDC (Relational Divergence Classification), but it is roughly 14% lower than the proposed method with 66.18%. The method of KDES in [2] is tested on this dataset with 61.76% for Rank 1, which is approximately 5% smaller than the proposed method at the first 7 ranks.

The state-of-the-art SDALF reported in [26] and the proposed method for person Re-ID are also tested on MICA1 dataset. Figure 12-b shows the testing results, with 73.13% at Rank 1 for the proposed method compared with the original KDES of 67.16% and 30% of SDALF. The deviation between two recognition rates of the proposed method and the original KDES gradually declines and almost reaches to the same value as SDALF after Rank 21.

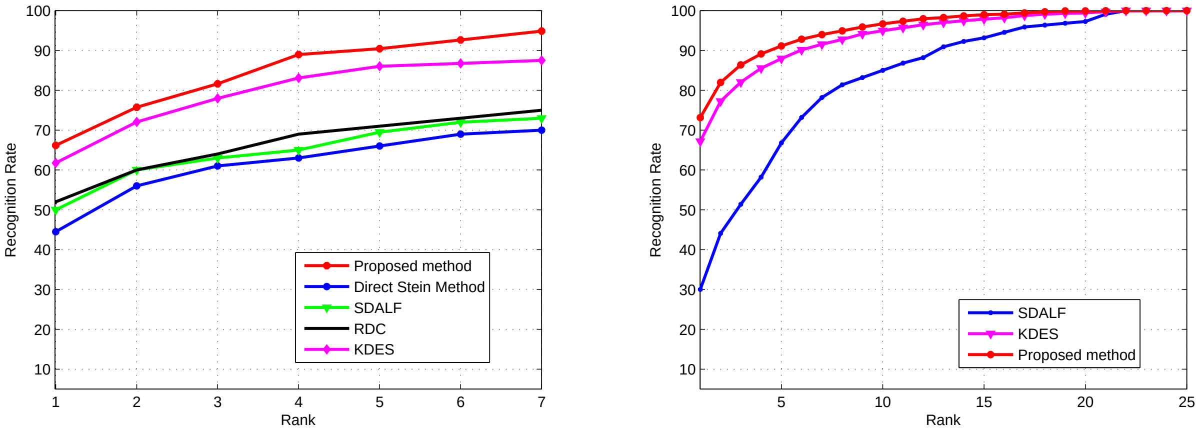
All of the above person Re-ID evaluations are done on the manually-cropped human ROIs, with the outperforming results obtained from the proposed KDES descriptor in comparison with the original KDES and some other state-of-the-art methods.

In order to show the comparative results between the manually-cropped and automatically-detected human ROIs, an experiment on MICA2 dataset is done. As indicated in Figure 13, when the human ROI images are gained by manual cropping, the recognition rate of Rank 1 is 33.25% by applying the proposed KDES descriptor. However, this figure falls sharply to 18.47% when the human ROIs are detected automatically. Clearly, the person Re-ID performance based on the proposed descriptor depends strongly on the human detection results. The well-aligned bounding boxes of the manually-extracted human ROIs give the better person Re-ID results than the automatically-detected ones because of occlusion, shadow phenomenon or background clutter appeared in the phase of human detection.

Some examples of automatic human detection results are shown in Fig. 14. We can see that the human detection is far from perfect. The human ROIs are sometimes false positive, mis-aligned or contain multiple persons. These can strongly affect the results of human re-identification.

V. Conclusion and Future Work

In this paper, person Re-ID problem in camera network achieves state-of-the-art performance on the benchmark datasets and our dataset by applying a robust person appearance representation based on KDES. Unlike some other state-of-the-art methods which are tested on the benchmark datasets with the manually-cropped human ROIs, in this paper, the person Re-ID performance is evaluated on the results of automatic human detector. Our experiments show that the proposed method gives an acceptable result for person re-identification in case of perfect human detection (74% for



(a)

(b)

Figure 12. The comparative results with (a) reported methods in [28] with iLIDs dataset and (b) the results are tested on our MICA1 dataset.

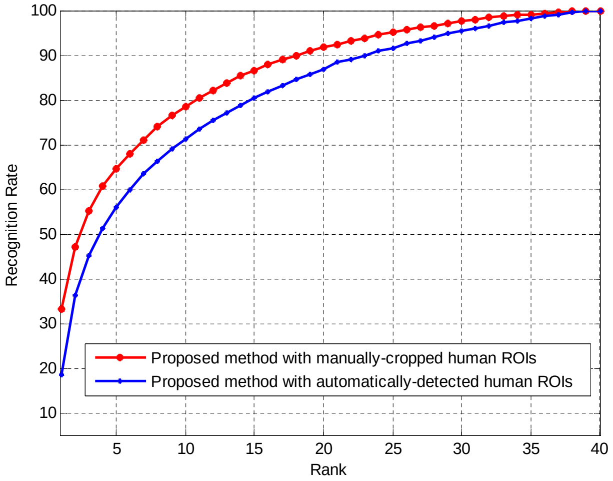


Figure 13. The recognition rates of the proposed KDES on the MICA2 dataset with manually-cropped and automatically-detected human ROIs.

CAVIAR4REID dataset). However, with the automatic human detection, it needs more works in order to reach the requirement. The vision-based person ID can be used in connective and complementing manner of different types of information in the proposed multimodal pedestrian localization system. The experimental results are promising, and based on this, a multimodal method, which uses particle filter and integrated data association algorithm, will be promoted in the future work to increase the performance of the combined person Re-ID and localization system.

Acknowledgement

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.04-2013.32.

References

[1] T. T. T. Pham, T. L. Le, T. K. Dao, and D. H. Le, “A robust model for person re-identification in multimodal person localization,” in The Ninth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM), 2015, pp. 38–43.

[2] L. Bo, X. Ren, and D. Fox, “Kernel descriptors for visual recognition,” in Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada, 2010, pp. 244–252.

[3] L. F. Teixeira and L. Corte-Real, “Video object matching across multiple independent views using local descriptors and adaptive learning,” Pattern Recognition Letters, vol. 30, no. 2, 2009, pp. 157–167.

[4] D. N. T. Cong, C. Achard, L. Khoudour, and L. Douadi, “Video sequences association for people re-identification across multiple non-overlapping cameras,” in Image Analysis and Processing (ICIAP). Springer, 2009, pp. 179–189.

[5] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, “Custom pictorial structures for re-identification,” in The British Machine Vision Conference (BMVC), vol. 2, no. 5, 2011, p. 6.

[6] D. Figueira, L. Bazzani, H. Q. Minh, M. Cristani, A. Bernardino, and V. Murino, “Semi-supervised multi-feature learning for person re-identification,” in 10th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2013, pp. 111–116.

[7] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, “Person re-identification by support vector ranking.” in The British Machine Vision

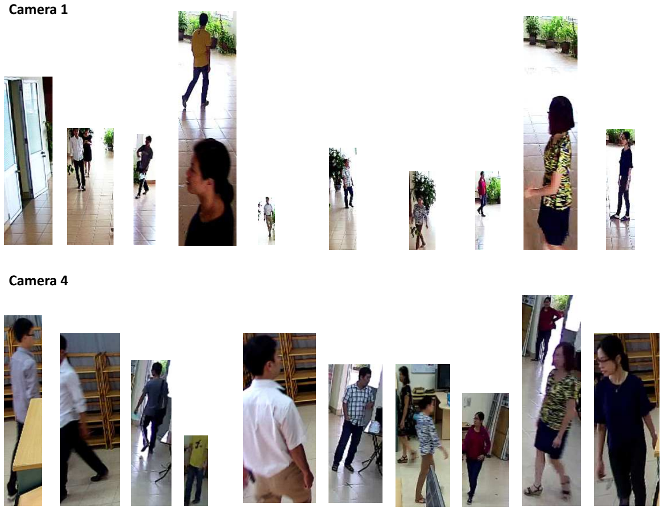


Figure 14. Results of automatically-detected ROIs for 10 people in MICA2 dataset.

Conference (BMVC), vol. 2, no. 5, 2010, p. 6.

[8] N. Martinel, C. Micheloni, and C. Piciarelli, "Learning pairwise feature dissimilarities for person re-identification," in Distributed Smart Cameras (ICDSC), 2013 Seventh International Conference on. IEEE, 2013, pp. 1–6.

[9] M. Bauml and R. Stiefelhagen, "Evaluation of local features for person re-identification in image sequences," in 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), 2011, pp. 291–296.

[10] S. Bak, E. Corvee, F. Brémond, and M. Thonnat, "Person re-identification using spatial covariance regions of human body parts," in Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2010, pp. 435–440.

[11] W. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 649–656.

[12] D. Figueira and A. Bernardino, "Re-identification of visual targets in camera networks: A comparison of techniques," in Image Analysis and Recognition. Springer, 2011, pp. 294–303.

[13] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in Asian Conference on Computer Vision (ACCV), 2011, pp. 501–512.

[14] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 3, 2013, pp. 653–668.

[15] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," The Journal of machine learning research, vol. 4, 2003, pp. 933–969.

[16] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler, "Re-identification of pedestrians in crowds using dynamic time warping," in European Conference on Computer Vision (ECCV), 2012, pp. 423–432.

[17] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with svms," Information Retrieval, vol. 13, no. 3, 2010, pp. 201–215.

[18] O. Vinyals, E. Martin, and G. Friedland, "Multimodal indoor localization: An audio-wireless-based approach," in Fourth International Conference on Semantic Computing (ICSC), 2010, pp. 120–125.

[19] T. K. Dao, H. L. Nguyen, T. T. Pham, E. Castelli, V. T. Nguyen, and D. V. Nguyen, "User localization in complex environments by multimodal combination of gps, wifi, rfid, and pedometer technologies," The Scientific World Journal, vol. 2014, 2014.

[20] T. Teixeira, D. Jung, G. Dublon, and A. Savvides, "Identifying people in camera networks using wearable accelerometers," in Proceedings of the 2nd International Conference on PErvasive Technologies Related to Assistive Environments, 2009, p. 20.

[21] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in Proceedings of the 17th International Conference on Pattern Recognition (ICPR), vol. 2, 2004, pp. 28–31.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2005, pp. 886–893.

[23] W. Bing-Bing, C. Zhi-Xin, W. Jia, and Z. Liquan, "Pedestrian detection based on the combination of hog and background subtraction method," in International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE), 2011, pp. 527–531.

[24] L. Bo and C. Sminchisescu, "Efficient match kernel between sets of features for visual recognition," in Advances in neural information processing systems, 2009, pp. 135–143.

[25] "http://mica.edu.vn/perso/Le-Thi-Lan/ReID.html, last access date is 4th may 2016."

[26] L. Bazzani, M. Cristani, A. Perina, and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses," Pattern Recognition Letters, vol. 33, no. 7, 2012, pp. 898–903.

[27] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 2360–2367.

[28] A. Alavi, Y. Yang, M. Harandi, and C. Sanderson, "Multi-shot person re-identification via relational stein divergence," in 20th IEEE International Conference on Image Processing (ICIP), 2013, pp. 3542–3546.