Improving Person Re-identification via Pose-aware Multi-shot Matching

Yeong-Jun Cho and Kuk-Jin Yoon
Computer Vision Laboratory, GIST, South Korea
{yjcho, kjyoon}@gist.ac.kr

Abstract

Person re-identification is the problem of recognizing people across images or videos from non-overlapping views. Although there has been much progress in person re-identification for the last decade, it still remains a challenging task because of severe appearance changes of a person due to diverse camera viewpoints and person poses. In this paper, we propose a novel framework for person re-identification by analyzing camera viewpoints and person poses, so-called Pose-aware Multi-shot Matching (PaMM), which robustly estimates target poses and efficiently conducts multi-shot matching based on the target pose information. Experimental results using public person re-identification datasets show that the proposed methods are promising for person re-identification under diverse viewpoints and pose variances.

1. Introduction

In recent years, a huge number of surveillance cameras have been installed in public places (*e.g.* offices, stations, airports, and streets) in order to closely monitor the scene and to give early warning of events such as accidents and crimes. However, it requires a lot of human efforts for dealing with large camera networks. In order to reduce the human efforts, an automatic person re-identification, *i.e.* re-id, task that associates people across images from non-overlapping cameras has been widely utilized.

For re-identifying people, most previous works generally rely on people appearances such as color, shape, and texture, since there is no continuity between non-overlapping cameras in terms of time and space. For this reason, many works have been focused on appearance modeling and learning such as feature learning [7, 25], metric learning [10, 18], and saliency learning [24] for the efficient re-id task. Unfortunately, however, the appearance of a person can change considerably across images depending on camera viewpoints as well as the pose of a person as shown in Fig. 1; thus the person re-id task relying only on the appearances is very challenging. Nonetheless, many previous

Figure 1. Challenging in person re-identification due to person appearance changes. Person appearance changes depending on the camera viewpoint and the pose of a person.

re-id frameworks [10, 18, 24] commonly adopt single-shot matching for measuring similarity (*or* difference) between a pair of person image patches. However, it is still hard to identify people with single-shot appearance matching because of the aforementioned severe appearance changes of people. In order to overcome the limitation of single-shot matching, several multi-shot matching methods [7, 13, 19] have been proposed in recent years; however the ambiguity owing to the viewpoint and pose variations is still remained.

In real world surveillance scenarios, each target (*i.e.* person) provides multiple observations along with its moving path. Furthermore, surveillance videos contain scene structures as well as scene contexts; a ground plane of the scene, the moving trajectory of a person, etc. In practice, it is possible to estimate the camera viewpoint from the scene information via human height-based auto-calibrations [11, 16] and vanishing point-based auto-calibrations [17]; then the difficulties of person re-id become more tractable.

In this paper, we propose a novel framework for person re-identification by analyzing camera viewpoints and person poses, so-called Pose-aware Multi-shot Matching (PaMM). We first calibrate camera viewpoints and robustly estimate target poses based on the proposed target pose estimation method. We then generate a multi-pose model containing four representative features extracted from four image clusters grouped by person poses (*i.e.* *f*ront, *r*ight, *b*ack, *l*eft). After generating multi-pose models, we calculate matching scores between multi-pose models in a weighted summation manner based on the pre-trained matching weights. With the proposed person re-identification framework, we can exploit additional cues

such as person poses and 3D scene information so as to make person re-identification problem more tractable.

To validate our methods, we extensively evaluate the performance of the proposed methods using public person re-id datasets [3, 9, 19]. Experimental results show that the proposed framework is promising for person re-identification under diverse viewpoint and pose variations and outperform other state-of-the-art methods.

The main ideas of this work are simple but very effective. In addition, our method can flexibly adopt any existing person re-identification methods such as feature learning-based [7, 25] and metric learning-based [5, 10, 20] methods as the baseline of our re-id framework. To the best of our knowledge, this is the first attempt to exploit viewpoint and pose information for *multi-shot* person re-identification.

2. Previous Works

We classify previous person re-identification methods into single-shot matching-based methods and multi-shot matching-based methods and briefly review them.

2.1. Single-shot matching

In order to re-identifying people across non-overlapping cameras, most of works generally rely on the appearances of people since there are no spatio-temporal continuity; we cannot fully utilize the motion or spatial information of a target (*i.e*. person) for person re-identification. For this reason, most of works have focused on appearance-based techniques such as feature and metric learning for the efficient person re-identification.

For feature learning, M. Farenzena *et al*. [7] proposed symmetry-driven accumulation of local features which are extracted based on principles of symmetry and asymmetry of a human body. This method exploits the human body model which is robust to human pose variations. Feature learning methods that select or weight discriminative features have been proposed in [15, 25]. These methods enable us to adaptively exploit features depending on the person appearance. Regarding the metric learning, several methods have been proposed such as KISSME [10], LMNN-R [6], and applied to the re-identification problem. Some works [10, 18] extensively evaluated and compared several metric learning methods (*e.g.* ITML [5], KISSME [10], LMNN [20] and Mahalnobis [18]) and showed the effectiveness of metric learning for re-identification. Similar to the metric learning methods, a saliency learning method was also proposed by R. Zhao *et al*. [24] which learns saliency for handling severe appearance changes.

Recently, deep learning-based person re-identification using a Siamese convolutional network have been proposed [1, 22] for simultaneously learning both features and metrics. Also [14] proposed both feature extraction and metric learning methods for re-identification.

On the other hand, a few works [2, 21] using target pose priors (pose cues) for person re-identification have been proposed very recently. S.Bak *et al*. [2] proposed to learn a generic metric pool which consists of metrics, each one learned to match specific pairs of poses. Z.Wu. *et al*. [21] proposed person re-identification involving human appearance modeling using pose priors and person-specific feature learning. Although these methods utilized pose priors for person re-identification, they consider single-shot matching that recognizes people by using a single appearance, which has difficulties in handling diverse appearance changes. In this paper, we propose a person re-identification framework using pose cues for efficient *multi-shot matching*.

2.2. Multi-shot matching

To overcome the limitation of single-shot matching-based methods, several multi-shot matching-based person re-identification methods have been proposed in recent years. Besides feature learning, Farenzena *et al*. [7] also provided multi-shot matching results by comparing each possible pair of histograms between different signatures (a set of appearances) and selecting the obtained lowest distance for the final score of matching. T. Wang *et al*. [19] proposed a video ranking method for multi-shot matching which automatically selects discriminative video fragments and learns a video ranking function. Y. Li *et al*. [13] also proposed a multi-shot person re-id method based on iterative appearance clustering and subspaces learning for effective multi-shot matching. Even though the multi-shot matching person re-identification methods overcome the limitations of single-shot matching to some extent, the ambiguity owing to the viewpoint and pose changes is still remained.

3. Motivation and Main Ideas

As shown in Fig. 1, person re-identification is quite challenging due to camera viewpoint and target pose variations. However, what if we know the camera viewpoint and the pose priors of targets in every non-overlapping camera in advance? In fact, the progress in auto-calibration techniques [11, 16] enable us to extract additional cues such as camera parameters, ground plane, 3D position of the targets without any off-line calibration tasks [23]. By exploiting those additional cues, we can also estimate target poses as described in Section 4.1. In this paper, we fully exploit those additional cues for multi-shot matching and propose the Pose-aware Multi-shot Matching (PaMM) for person re-identification.

Suppose that we estimate camera viewpoints and target poses, and there is a simple 2vs2 matching scenario containing one same-pose matching (*f*ront-*f*ront) and three different-pose matchings (*f*ront-*r*ight, *l*eft-*f*ront, *l*eft-*r*ight) as shown in Fig. 2. We can expect that the result of the

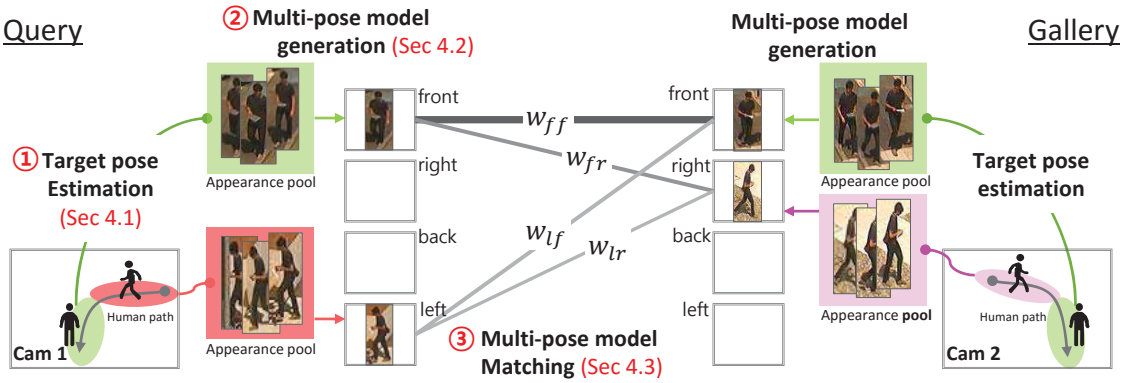


Figure 2. Proposed multi-shot matching framework for person re-identification.

same-pose matching is generally more reliable than those of different-pose matchings, since targets keep their appearances across cameras when the target poses are the same (In this work, we exclude photometric issues such as illumination changes and camera color response differences). Then, in this multi-shot matching scenario, it is desired that the same-pose matching (*f*ront-*f*ront) plays a more important role than different-pose matchings. Hence, in this work, we incorporate this matching idea by aggregating matching scores of all pose matchings in a weighted summation manner as shown in Fig. 2, where the thicknesses of lines indicate the matching weights (lines above ③ in Fig. 2). We also study how to efficiently match between multi-shot appearances using target pose information.

4. Proposed PaMM Framework

In the proposed person re-identification framework, we first estimate the camera viewpoint and target poses (Sec. 4.1), and then generate multi-pose models containing four representative features extracted from four image clusters obtained based on the target (*i.e.* person) poses (*i.e.* *f*ront, *r*ight, *b*ack, *l*eft) (Sec. 4.2). After generating multi-pose models, we calculate matching scores between multi-pose models in a weighted summation manner based on the pre-trained matching weights (Sec. 4.3). The matching weight training is describe in Sec. 4.4. Figure 2 illustrates an overall framework for person re-identification.

4.1. Target pose estimation

Before estimating target poses, we estimate camera intrinsic and extrinsic parameters (or camera pose) by using auto-calibration algorithms based on the human heights [11, 16]. Then, a relation between an image (pixel coordinates) and the real world (world coordinates) is described as

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K} \begin{bmatrix} \mathbf{R} & \mathbf{t} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{1}$$

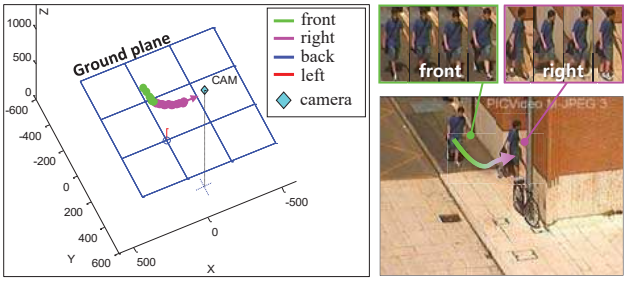


Figure 3. Target pose estimation: (left) estimated 3D structure and target poses along the path , (right) corresponding 2D images and appearances grouped by poses.

where $\mathbf{K}$, $\mathbf{R}$, and $\mathbf{t} = [X_{cam}, Y_{cam}, Z_{cam}]^\top$ represent a camera intrinsic matrix, a camera rotation matrix, and a camera position, respectively, In addition, $[u, v]$ and $[X, Y, Z]$ represent image and world coordinates, respectively. As we know the camera parameters, we can project every object in images onto the ground plane (world XY plane). Object k appearing at frame t in camera C is denoted by $\mathbf{O}_t^{C,k} = (\mathbf{P}_t^{C,k}, \mathbf{v}_t^{C,k}, \theta_t^{C,k})$, where $\mathbf{P}_t^{C,k} = \left[X_t^{C,k}, Y_t^{C,k}, 1\right]$, $\mathbf{v}_t^{C,k}, \theta_t^{C,k}$ are the position, velocity, and target pose angle w.r.t. the camera, respectively.

Inspired by [21], we define the target velocity $\mathbf{v}_t^{C,k}$ and camera viewpoint vector $\mathbf{c}_t^{C,k}$ in order to estimate target poses as

$$\mathbf{v}_t^{C,k} = \left[(X_t^{C,k} - X_{t-1}^{C,k}), \quad (Y_t^{C,k} - Y_{t-1}^{C,k})\right], \tag{2}$$

$$\mathbf{c}_t^{C,k} = \left[(X_{cam}^{C} - X_{t-1}^{C,k}), \quad (Y_{cam}^{C} - Y_{t-1}^{C,k})\right]. \tag{3}$$

Assuming that pedestrians mostly walk forward, a target pose angle of the object k can be estimated by (for convenience we omit C from here),

$$\theta_t^k = \arccos\left(\frac{\mathbf{c}_t^{k\,T} \cdot \mathbf{v}_t^k}{\|\mathbf{c}_t^k\| \, \|\mathbf{v}_t^k\|}\right). \tag{4}$$

Figure. 3 shows the example of estimated target poses. However, initially estimated θ_t^k is noisy as in Fig. 4 (a). In

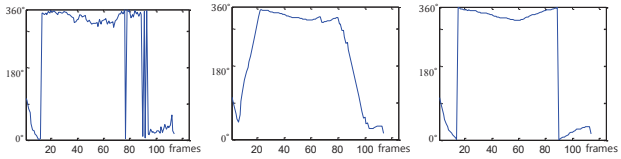


Figure 4. (left) initial pose angle, (middle) smoothing result in Cartesian coordinates, (right) smoothing result in polar coordinates.

order to reduce the noise, we smooth θ_t^k by using a moving average algorithm in the polar coordinate system as

$$\hat{\theta}_t^k = \arctan\left(\frac{\sum_{i=t-m}^{t+m}\sin\left(\theta_i^k\right)}{\sum_{i=t-m}^{t+m}\cos\left(\theta_i^k\right)}\right), \tag{5}$$

where m is a moving average parameter (we set $m = 10$). Although there are several discontinuities around $0°$ and $360°$, the smoothing result is reliable thanks to the smoothing process in polar coordinates, whereas the smoothing result in Cartesian coordinates is not reliable (Fig. 4 (b),(c)).

4.2. Multi-pose model generation

4.2.1 Sample selection based on sample confidence

For generating good multi-pose models, we ought to filter out the unreliable target samples having incorrect angles or polluted appearances along the moving trajectory. To this end, we define sample confidence to measure the reliability of target samples based on following requirements (R1-R3):

• **Variation of angle** (R1): We assume that the angle of walking person does not change abruptly between temporally neighboring frames. If there are rapid changes in angle across consecutive frames, we regard them as unreliable samples and filter them out. We observe that, inaccurate localization of a person generally causes large variation in angle. In order to consider it, we measure the angle variation as

$$\delta_t^k = \min\left(d\left(\hat{\theta}_t^k\right), \left|d\left(\hat{\theta}_t^k\right) - 360\right|\right), \tag{6}$$

where $d\left(\hat{\theta}_t^k\right) = \left|\hat{\theta}_{t-1}^k - \hat{\theta}_t^k\right|$. Even though there is an angle discontinuity between $0°$ and $360°$, δ_t^k is reliably calculated thanks to the second term of min function.

• **Magnitude of target velocity** (R2): When a target is stationary for several frames, the target velocity $\mathbf{v}_t^k$ is close to 0 and the estimated target angle based on Eq. (4) becomes unreliable[1]. To handle the problem, we measure the magnitude of the target velocity as $\mathcal{M}_t^k = \left\|\mathbf{v}_t^k\right\|_2$. A sample with the small velocity magnitude is regarded unreliable.

• **Occlusion rate** (R3): A target occluded by others is also not a reliable target sample since the appearance of target is polluted. To deal with the occluded target samples, we measure the occlusion rate of each target as

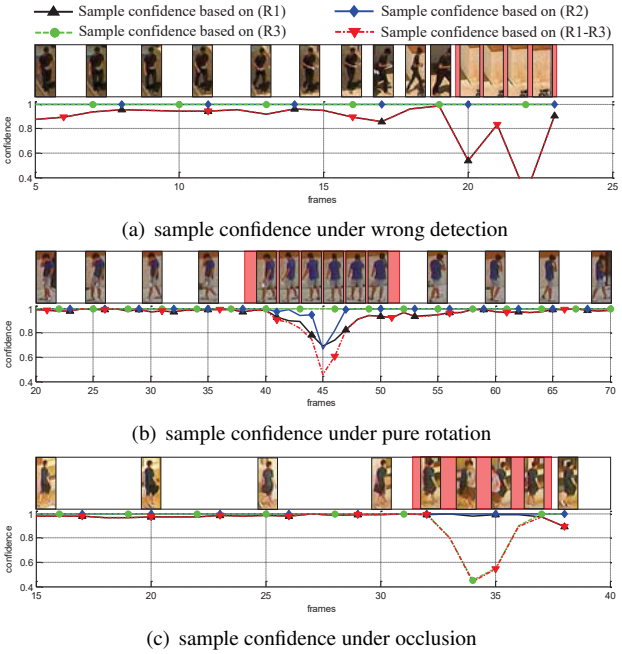
$$Occ_t^k = \max_{h\in\mathbf{H}^k}\left(\frac{area(B_t^k \cap B_t^h)}{area(B_t^k)}\right), \tag{7}$$

where B_t^k is a 2D bounding box of an object k at frame t, B_t^h is a 2D bounding box occluding B_t^k, $\mathbf{H}^k$ is a set of object indexes occluding object k. As we know the 3D position of each target $\mathbf{P}_t^k$, it is easy to find $\mathbf{H}^k$.

Based on the above requirements, we define the sample confidence as

$$conf\left(\mathbf{O}_t^k\right) = e^{-\alpha\delta_t^k}\cdot\tanh\left(\mathcal{M}_t^k\right)\cdot\left(1 - Occ_t^k\right), \tag{8}$$

where α is a scale parameter (we set $\alpha = 0.01$). The sample confidence lies in [0,1]. Figure 5 shows the sample confidences under various situations. We regard a target sample as a reliable one with high sample confidence when $conf\left(\mathbf{O}_t^k\right) > 0.8$.



(a) sample confidence under wrong detection

(b) sample confidence under pure rotation

(c) sample confidence under occlusion

Figure 5. Sample confidence under various conditions (best viewed in color and high resolution image).

4.2.2 Generating multi-pose model

After the sample selection, we divide target samples into four groups $\left\{\mathbf{g}_p^k\right\}_{p\in\{f,r,b,l\}}$ according to their pose angles (i.e. *f*ront, *r*ight, *b*ack, *l*eft). Each group covers $90°$. It is worthy to note that the proposed sample confidence efficiently filters unreliable samples out as shown in Fig. 6. After the clustering, we extract features from four groups and the multi-pose model of object k is defined as

$$\mathbf{M}^k = \left\{f\left(\mathbf{g}_p^k\right)\right\}_{p\in\{f,r,b,l\}}, \tag{9}$$

[1]To estimate target viewpoint angles, we assume that targets mostly move forward in Sec 4.1. However, in the case of stationary targets, the assumption is not satisfied. Note that the stationary targets are likely to have pure rotational motion which cannot be handled by Eq (4).

(a) average of each cluster without sample selection

(b) average of each cluster with sample selection

Figure 6. Clustering results according to target angels without and with sample selection. The clusters with sample selection repersent more clear directivity.

where $f(\cdot)$ is a function which extracts features from a set of images. Details of feature extraction is described below.

Feature extracion: We extract dColorSIFT which is a dense feature descriptor containing dense LAB-color histogram and dense SIFT as in [24]. The authors pointed out that the densely sampled local features have been widely applied to matching problems due to their robustness in matching. In our feature extraction process, each person image is normalized to 128×48 pixels and we extract dColorSIFT [24] descriptors from all images in each group. Then each groups $\left\{\mathbf{g}_p^k\right\}_{p\in\{f,r,b,l\}}$ has multiple feature descriptors, respectively. We then select a median dColorSIFT descriptor as the representative descriptor of each group. The selection of the median feature descriptor is reliable since it reflects the characteristics of each group robustly to outliers and furthermore it keeps details. Finally, the multi-pose model $\mathbf{M}^k$ of each person contains multiple representative dColorSIFT descriptors extracted from multiple groups. Our method can apply any kind of feature descriptors and feature extraction methods.

4.3. Multi-pose model matching

In this section, we describe the matching process of multi-pose models. Suppose that we have $\mathbf{M}^k$, $\mathbf{M}^l$ which are the multi-pose models of object k and l appeared in different cameras, respectively. In order to measure the similarity between two multi-pose models, we first calculate all pairwise feature distances of multi-pose models as $x_{pq} = dist\left(f\left(\mathbf{g}_p^k\right), f\left(\mathbf{g}_q^l\right)\right)$, where $p, q \in \{f, r, b, l\}$ and $dist(\cdot)$ is a distance function. For the distance function, we can use any metrics such as KISSME [10], ITML [5] and LMNN [20]. Then, the multi-pose model matching cost is computed in a weighted summation manner as

$$S\left(\mathbf{M}^k, \mathbf{M}^l\right) = \frac{\sum_{p,q} w_{pq} e_{pq} x_{pq}}{\sum_{p,q} w_{pq} e_{pq}}, \quad p, q \in \{f, r, b, l\}, \quad (10)$$

$$e_{pq} = \begin{cases} 1 & \text{if } (p,q) \text{ pair exists} \\ 0 & \text{otherwise} \end{cases}, \quad (11)$$

where w_{pq} is a matching weight that attaches importance to pairwise matching x_{pq}. Training matching weights is described in the next section.

4.4. Training matching weights

For training matching weights, we assume that every (p, q) pair exists and omit the normalization term $\sum_{p,q} w_{pq} e_{pq}$. Then Eq. (10) is rewritten as

$$\begin{aligned} S(\mathbf{x}) &= \mathbf{w}^T \mathbf{x}, \\ \mathbf{x} &= \{x_{ff}, x_{fr}, \ldots, x_{ll}\}^T, \\ \mathbf{w} &= \{w_{ff}, w_{fr}, \ldots, w_{ll}\}^T, \end{aligned} \quad (12)$$

where $\mathbf{x} \in \mathbb{R}^{16\times 1}$ is a vector of pairwise feature distances and $\mathbf{w} \in \mathbb{R}^{16\times 1}$ is a vector of matching weights. In order to train matching weights $\mathbf{w}$, we collect training samples $\mathcal{D} = \left\{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathbb{R}^{16\times 1}, y_i \in \{-1, 1\}\right\}_{i=1}^N$, where N is the number of training samples and y_i is a corresponding class of the sample. Given training set $\mathcal{D}$, we exploit Support Vector Machine (SVM) [4] to find the weights $\mathbf{w}$ by solving following optimization problem:

$$\begin{aligned} &\arg\min_{\mathbf{w},\xi} \left(\frac{1}{2}\|\mathbf{w}\|^2 + \lambda \sum_i^N \xi_i\right), \\ &s.t.\ \ y_i\left(\mathbf{w}^T \mathbf{x}_i\right) \geq 1 - \xi_i, \ \ \xi_i \geq 0, \ \ \text{for } 1 \leq i \leq N, \end{aligned} \quad (13)$$

where λ is a margin trade-off parameter and ξ_i is a slack variable. The solution given by SVM assures maximal margin. Details and the results of matching weight training are given in Sec 6.1.

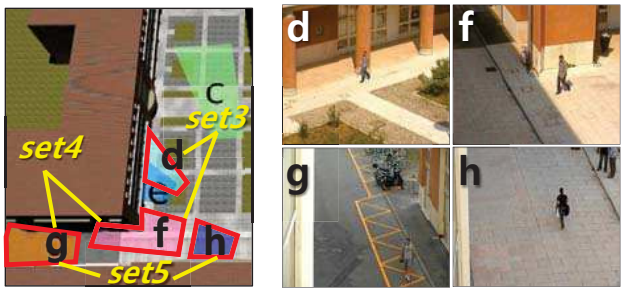
5. Datasets and Methodology

Datasets. For training matching weights, we use `CUHK02` [12] and `VIPeR` [8]. For test methods, we use `iLIDS-Vid` [19], `PRID 2011` [9] and `3DPeS` [3].

• `CUHK02` [12] contains 1,816 persons from five different outdoor camera pairs. Five camera pairs have 971, 306, 107, 193 and 239 persons with the size of 160×60 pixels, respectively. Each person has two images per camera which were taken in different times. Most of people are with burdens (*e.g.* backpack , handbag, strap bag, or baggage). For our experiments, we manually extract all pose angles of each person in four directions (*i.e.* *f*ront, *r*ight, *b*ack, *l*eft) since CUHK02 does not provide the pose angles. This dataset is used for training multi-shot weights $\mathbf{w}$.

• `VIPeR` [8] includes 632 persons and two outdoor cameras under different viewpoints and light conditions. Each person has one image per camera and each image has been scaled to be 128×48 pixels. It provides the pose angle of each person as 0°(*f*ront), 45°, 90°(*r*ight), 135°, 180°(*b*ack). We use both `CUHK02` and `VIPeR` datasets for training multi-shot matching weight $\mathbf{w}$.

• `iLIDS-Vid` [19] has been created from the pedestrians in 2 non-overlapping cameras, monitoring an airport arrival hall. It provides multiple cropped images for each 300 distinct individual and is very challenging due to clothing sim-



(a) camera layout of video sets (b) sample frames of each camera (d-h)

Figure 7. Test dataset: 3DPeS [3]

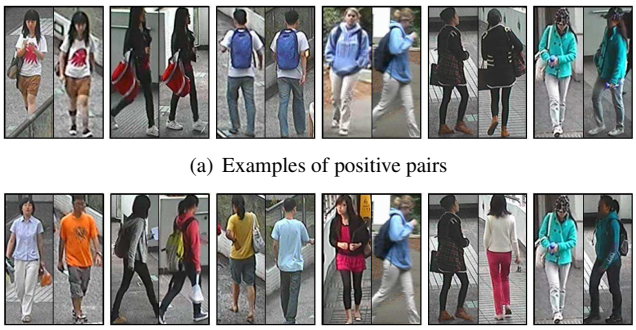
ilarities, lighting and viewpoint variations, cluttered background and severe occlusions.

• PRID 2011 [9] provides multiple person trajectories recored from 2 different static surveillance cameras, monitoring crosswalk and sidewalk. In the dataset, 200 persons appear in both views.

Since the datasets (iLIDS [19], PRID [9]) do not provide full surveillance video sequences but provide only cropped images, we could not automatically estimate camera viewpoints and poses of targets. Therefore, in order to evaluate our method with the datasets, we annotated the pose of each target manually.

• 3DPeS [3] has been collected by 8 non-overlapped outdoor cameras, monitoring different sections of the campus. Differently from other re-id datasets (iLIDS, PRID), it provides full surveillance video sequences: providing 6 sets of video pairs, uncompressed images with a resolution of 704x576 pixels at 15 frame rate, containing hundreds of people and calibration parameters. However, this dataset provides ground-truth person identity only for selected snapshots (*i.e.* no ground-truth for video sequences). For our experiments, we used 3 sets of video pairs (Set3,4,5) and manually extracted ground truth labels (identities, center points, widths, heights) of video Set3,4,5. The pose of each target was estimated as described in Sec. 4.1. The camera layout and sample frames are given in Fig. 7. Three video pairs contain 39, 24 and 36 identities, respectively.

Evaluation methodology. To compare person re-identification methods, we follow the evaluation steps described in [7]. First, we randomly split person-identities in video pairs into two sets with the equal number of identities, one set for training and the other set for testing. We learn several metrics such as LMNN [20], ITML [5], KISSME [10], and Mahal [18] for the baseline distance functions of our person re-identification framework. After training distance metrics, we calculate all possible matches between testing video pairs. We repeat the evaluation steps over 10 times. We plot the Cumulative Match Curve (CMC) [8] representing true match being found within the first n ranks for comparing performances of methods.



(a) Examples of positive pairs

(b) Examples of negative pairs

Figure 8. Examples of training sample pairs.

6. Experimental Results

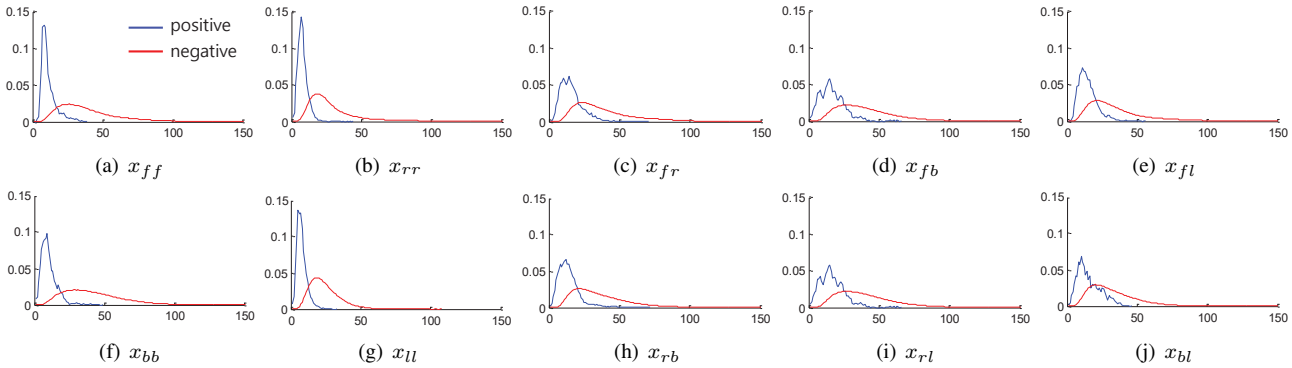
6.1. Training multi-shot matching weights

In practice, we need to consider only 10 weights rather than 16 weights due to the weight symmetry: we let $w_{pq} = w_{qp}$, where $p \neq q$. Consequentially, we learn four same-pose matching weights $(w_{ff}, w_{rr}, w_{bb}, w_{ll})$ and six different - pose matching weights $(w_{fr}, w_{fb}, w_{rb}, w_{rl}, w_{bl}, w_{fl})$.

As mentiond in Sec. 5, for training the weights $\mathbf{w} \in \mathbb{R}^{10\times 1}$, we use two datasets, CUHK02 [12] and VIPeR [8]. By using the datasets, we generate 3,520 positive image pairs and 35,200 negative image pairs that cover diverse pose combinations as shown in Fig. 8. Here, a positive image pair is a pair of images of the same person and a negative image pair is a pair of images of the different persons regardless of the poses of persons. We then extract pairwise feature distances $\{x_{ff}, x_{rr}, \ldots, x_{fl}\}$ for all images pairs by following metric learning steps described in Sec. 5. Distributions of feature distances[2] $\{x_{pq}\}$ are plotted in Fig. 9. For example, Fig. 9 (a) shows the feature distance distribution of $f$ront image pairs of the same person (positive) and difference persons (negative). Note that, a large statistical distance between positive and negative distributions implies the high discriminating power. We observe that the same-pose matchings (Fig. 9 (a,b,f,g)) are more discriminative than different-pose matchings (Fig. 9 (c-e,h-j)).

After obtaining distributions of feature distances, we generate training samples $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i \in \mathbb{R}^{10\times 1}$, $y_i \in \{1, -1\}$ by randomly selecting each x_{pq} from each distribution. Figure 10 shows the result of weight training using the KISSME [10] distance metric. The result represents that the weights of the same-pose matchings (ff, rr, bb, ll) are generally larger than those of the different-pose matchings (fr, fb, rb, rl, bl, fl). For consecutive experiments, we train each matching weight for each metric learning method, individually. The training results do not depend on the metric learning methods and show similar tendencies.

[2]Unfortunately, we could not make (r, l) pairs using training datasets CUHK02, VIPeR since they do not have such pairs. In order to make the distribution of x_{rl}, we regard x_{rl} follows the similar distribution with x_{fb}.



(a) x_{ff} (b) x_{rr} (c) x_{fr} (d) x_{fb} (e) x_{fl}

(f) x_{bb} (g) x_{ll} (h) x_{rb} (i) x_{rl} (j) x_{bl}

Figure 9. Distributions of pairwise feature distances $\{x_{pq}\}$ extracted from training data.

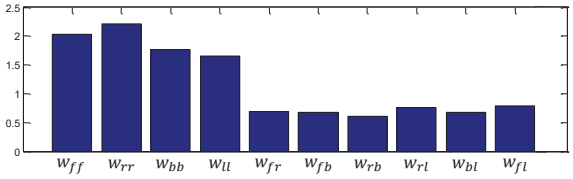


Figure 10. Trained weights (distance metric: KISSME [10])

6.2. Performance enhancements via PaMM

In experiments, we denote our method 'Pose-aware Multi-shot Matching' as PaMM. As a baseline method for PaMM, we can use any conventional single-shot matching-based methods (*e.g.* feature learning methods, metric learning methods) — any single-shot matching-based method can be used for computing pairwise feature distance in our framework. In this work, we choose several single-shot person re-identifications[3] based on various metric learnings such as LMNN [20], ITML [5], KISSME [10], and Mahalanobis [18] for the baselines of PaMM.

For validating the performance enhancement via PaMM, we compare the person re-identifications with and without PaMM using the dataset `3DPeS-Set3`. To evaluate each performance, we follow evaluation steps explained in Sec. 5. As shown in Table. 1, all baselines [5, 10, 18, 20] are improved considerably for all ranks (r=1,3,5,10) thanks to the proposed PaMM. Especially, the performance enhancement at r=1 is remarkable (achieving 2.7%∼47.4% enhancement). The results imply that proposed PaMM can improve any kind of single-shot matching-based person re-identification. Compared to conventional single-shot matching-based methods, PaMM exploits a plenty of appearances selected based on the sample confidence and matches multiple appearances efficiently using pose cues. In the consecutive experiments, we use KISSME [10] as the baseline for PaMM.

[3] As the appearance of each identity for the single-shot matching-based methods, we randomly select a single appearance for each identity. For unbiased selections, we repeat the appearance selection over 10 times and calculate the average performance for the final result.

Table 1. Performance enhancement over single-shot matching-based methods via PaMM. † denotes a multi-shot matching method.

	3DPeS – Set 3			
Modality \ Rank	r = 1	r = 3	r = 5	r = 10
L_2 (Euclidean)	31.5	52.6	63.1	86.8
L_2-PaMM†	34.2	57.9	73.7	89.5
LMNN [20]	26.3	52.6	68.4	89.4
LMNN-PaMM†	52.6	79.0	86.8	94.7
ITML [5]	31.5	57.8	76.3	89.4
ITML-PaMM†	39.5	68.4	81.6	92.1
KISSME [10]	21.0	50.0	63.1	84.2
KISSME-PaMM†	68.4	89.5	94.7	100
Mahal [18]	31.5	63.1	73.6	89.4
Mahal-PaMM†	60.5	78.6	89.5	97.4

6.3. Performance comparisons with other methods

For performance comparisons with other methods, we first evaluate methods using `3DPeS` dataset providing full surveillance videos. In this section we denote different versions of our person re-identification framework as follows:

- PaMM-nss: PaMM without sample selection.
- PaMM-nw: PaMM without weighted multi-pose matching (we use uniform weights $\mathbf{w} = \mathbf{1}$ for PaMM-nw).
- PaMM: PaMM with all proposed methods.

We also implemented other multi-shot matching methods for comparison as follows:

- Full Match-avg: matching all possible pairs between multiple appearances and averaging all matching scores.
- Full Match-min: matching all possible pairs between multiple appearances and selecting the smallest matching score for the final score as used in [7].

Table. 2 represents that our methods outperform other multi-shot matchings (Full Match-avg, Full Match-min) and all single-shot matching methods. Even though the multi-shot matchings 'Full Match-avg' and 'Full Match-min' exploit all multiple appearances of targets, the performances of both methods are lower than PaMM. It supports that the proposed PaMM reasonably extract representative features among multiple appearances (Sec. 4.2) and

Table 2. Performance comparison. † denotes a multi-shot matching method. The best and second best scores in each rank are marked with **red** and blue. For VaMM-nW, we use uniform weights $\mathbf{w} = \mathbf{1}$. We use the same feature descriptor (dColorSIFT) for all methods. Full Match and PaMM use KISSME [10] for their metrics. **AUC** is an area under curve of CMC.

	3DPeS – Set 3				3DPeS – Set 4				3DPeS – Set 5				3DPeS – Set All				
Modality \ Rank	r = 1	r = 3	r = 5	**AUC**	r = 1	r = 3	r = 5	**AUC**	r = 1	r = 3	r = 5	**AUC**	r=1	r=5	r=10	r=15	**AUC**
L_2 (Euclidean)	31.5	52.6	63.1	79.4	33.3	41.6	54.1	64.6	11.1	27.7	44.4	61.6	19.3	30.6	38.7	50.0	66.8
LMNN [20]	26.3	52.6	68.4	80.8	33.3	58.3	66.6	74.3	22.2	47.2	72.2	79.2	24.4	48.7	65.3	74.4	80.8
ITML [5]	31.5	57.8	76.3	82.7	41.6	66.7	70.8	76.0	33.3	63.8	72.2	80.9	23.4	52.0	71.4	77.5	82.5
KISSME [10]	21.0	50.0	63.1	78.8	25.0	58.3	66.6	71.9	22.2	47.2	63.8	76.5	26.5	52.0	66.3	79.5	82.4
Mahal [18]	31.5	63.1	73.6	83.5	33.3	58.3	66.6	73.3	25.0	47.2	69.4	78.4	28.5	50.0	67.3	78.5	82.8
Full Match-avg.†	42.1	63.2	73.7	85.0	50.0	62.5	75.0	79.5	33.3	55.7	69.4	79.6	23.5	49.0	61.2	70.4	78.1
Full Match-min†	47.4	78.9	89.4	92.1	45.8	75.0	**91.7**	85.4	44.4	66.7	**83.3**	88.6	35.7	73.5	83.7	89.8	91.0
PaMM-nss†(ours)	57.9	**89.5**	**94.7**	95.3	**58.3**	75.0	83.3	87.2	**55.6**	**77.8**	**83.3**	**89.5**	52.0	79.6	83.7	91.8	92.7
PaMM-nw†(ours)	**68.4**	**89.5**	**94.7**	95.6	**58.3**	**83.3**	83.3	**88.9**	**55.6**	**77.8**	**83.3**	87.4	56.1	78.6	**89.8**	91.8	92.4
PaMM†(ours)	**68.4**	**89.5**	**94.7**	**96.4**	**58.3**	**83.3**	**91.7**	**88.9**	**55.6**	**77.8**	**83.3**	88.6	**59.2**	**82.7**	**89.8**	**94.9**	**94.1**

Table 3. Performance comparison with other methods. † denotes a multi-shot matching and aR an average rank. We implemented the multi-shot version of L+XQDA by following [7].

	iLIDS-Vid					PRID 2011					3DPeS – Set All				
Modality \ Rank	r=1	r=5	r=10	r=20	aR	r=1	r=5	r=10	r=20	aR	r=1	r=5	r=10	r=15	aR
(S1) SDALF-SS [7]	5.1	14.9	20.7	31.3	8.0	4.9	21.5	30.9	45.2	7.7	10.8	24.6	35.7	42.4	6.0
(S2) Salience [24]	10.2	24.8	35.5	52.9	5.5	25.8	43.6	52.6	62.0	5.5	–	–	–	–	–
(S3) L+XQDA-SS [14]	18.0	41.2	54.7	67.0	4.5	39.0	68.0	83.0	91.0	3.2	35.5	69.5	80.1	87.6	3.0
(S4) KISSME [10]	11.3	27.3	37.3	49.7	5.0	22.3	43.2	55.1	70.4	5.5	26.5	52.0	66.3	79.5	4.0
(M1) SDALF-MS† [7]	6.3	18.8	27.1	37.3	7.0	5.2	20.7	32.0	47.9	7.2	23.2	44.4	56.6	65.7	5.0
(M2) Salience+DVR† [19]	**30.9**	54.4	65.1	77.1	1.7	41.7	64.5	77.5	88.8	3.7	–	–	–	–	–
(M3) L+XQDA-MS† [14]	21.7	49.1	61.8	75.3	3.0	**56.5**	**85.7**	**96.3**	**97.7**	**1.0**	42.1	70.0	84.5	91.4	2.0
(M4) PaMM† (Ours)	30.3	**56.3**	**70.3**	**82.7**	**1.2**	45.0	72.0	85.0	92.5	2.0	**59.2**	**82.7**	**89.8**	**94.9**	**1.0**

efficiently matches multi-pose models (Sec. 4.3).

Even though the test datasets 3DPeS-Set3,4,5 contain people having various appearances and poses, they contain a few number of identities (Set3:39, Set4:24, Set5:36). When the number of identities is small, the re-identification task becomes much easier because of the small pool of comparison targets. To show the person re-identification performance under more large scale camera networks, we concatenate all datasets and generate 3DPeS-Set All containing 99 identity pairs. It is reasonable since each dataset (3DPeS-Set3,4,5) does not share identities with each other. Table. 2 shows the comparison results with dataset 3DPeS-Set All and represents our methods still show promising performance compared to others.

We also provide evaluation results and comparisons with other state-of-the art multi-shot matching methods such as SDALF [7], Salience+DVR [19], and L+XQDA [14] with more public datasets (iLIDS [19], PRID [9]) in Table. 3. In Table 3, S1, S2, S3, and S4 (single-shot re-id methods) are used as the baselines of M1, M2, M3, and M4 (multi-shot re-id methods), respectively. Overall, PaMM shows the best performance while significantly enhancing its baseline performance (19%∼32.7% enhancement at r=1).

Although L+XQDA-MS [14] shows better performance for PRID, it is mainly because the superior performance of its baseline (S3). Actually, the performance enhancement of M3 from S3 is less (3.7%∼17.5% at r=1) than ours — for iLIDS which is much more challenging than PRID, the improvement of ours and [14] are 19% and 3.7%, respectively. PaMM shows promising performance regardless of datasets. It should be also noted that PaMM can achieve better performance by adopting better baseline methods.

7. Conclusions

In this paper, we proposed a novel framework for person re-identification, so called Pose-aware Multi-shot Matching (PaMM) that robustly estimates target poses and efficiently conducts multi-shot matching based on the target pose information. We extensively evaluated and compared the performance of the proposed method using public person re-identification datasets. The results showed that proposed methods are promising for person re-identification under diverse target pose variances. The proposed methods can flexibly adopt any existing person re-identification methods for computing pairwise feature distance in our framework.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) (No. NRF-2015R1A2A1A01005455) and Institute for Information & communications Technology Promotion(IITP) (No.B0101-16-0552, Development of Predictive Visual Intelligence Technology) grants funded by the Korea government(MSIP).

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. *Differences*, 5:25, 2015. 2

[2] S. Bak, F. Martins, and F. Bremond. Person re-identification by pose priors. In *IS&T/SPIE Electronic Imaging*, pages 93990H–93990H. International Society for Optics and Photonics, 2015. 2

[3] D. Baltieri, R. Vezzani, and R. Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *MA3HO*, pages 59–64, Scottsdale, Arizona, USA, Nov. 2011. 2, 5, 6

[4] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 5

[5] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216. ACM, 2007. 2, 5, 6, 7, 8

[6] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*, pages 501–512. Springer, 2011. 2

[7] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367. IEEE, 2010. 1, 2, 6, 7, 8

[8] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *PETS*, volume 3. Citeseer, 2007. 5, 6

[9] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, 2011. 2, 5, 6, 8

[10] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295. IEEE, 2012. 1, 2, 5, 6, 7, 8

[11] W. Kusakunniran, H. Li, and J. Zhang. A direct method to self-calibrate a surveillance camera by observing a walking pedestrian. In *DICTA*, pages 250–255. IEEE, 2009. 1, 2, 3

[12] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, pages 3594–3601. IEEE, 2013. 5, 6

[13] Y. Li, Z. Wu, S. Karanam, R. J. Radke, and N. Troy. Multi-shot human re-identification using adaptive fisher discriminant analysis. *BMVC*, 2015. 1, 2

[14] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*. IEEE, 2015. 2, 8

[15] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important? In *ECCV*, pages 391–401. Springer, 2012. 2

[16] F. Lv, T. Zhao, and R. Nevatia. Camera calibration from video of a walking human. *TPAMI*, (9):1513–1518, 2006. 1, 2, 3

[17] R. Orghidan, J. Salvi, M. Gordan, and B. Orza. Camera calibration using two or three vanishing points. In *FedCSIS*, 2012. 1

[18] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof. Mahalanobis distance learning for person re-identification. In *Person Re-Identification*, pages 247–267. Springer, 2014. 1, 2, 6, 7, 8

[19] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703. Springer, 2014. 1, 2, 5, 6, 8

[20] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, pages 1473–1480, 2005. 2, 5, 6, 7, 8

[21] Z. Wu, Y. Li, and R. J. Radke. Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features. *TPAMI*, 37(5):1095–1108, 2015. 2, 3

[22] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *ICPR*, pages 34–39. IEEE, 2014. 2

[23] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *ICCV*, volume 1, pages 666–673. IEEE, 1999. 2

[24] R. Zhao, W. Ouyang, and X. Wang. Unsupervised salience learning for person re-identification. In *CVPR*, pages 3586–3593. IEEE, 2013. 1, 2, 5, 8

[25] R. Zhao, W. Ouyang, and X. Wang. Learning mid-level filters for person re-identification. In *CVPR*, pages 144–151. IEEE, 2014. 1, 2