SKEPRID: Pose and Illumination Change-Resistant Skeleton-Based Person Re-Identification

TUO YU, University of Illinois at Urbana-Champaign, USA

HAIMING JIN, Shanghai Jiao Tong University, China

WAI-TIAN TAN, Cisco Systems, Inc., USA

KLARA NAHRSTEDT, University of Illinois at Urbana-Champaign, USA

Currently, the surveillance camera-based person re-identification is still challenging because of diverse factors such as people's changing poses and various illumination. The various poses make it hard to conduct feature matching across images, and the illumination changes make color-based features unreliable. In this article, we present SKEPRID,¹ a skeleton-based person re-identification method that handles strong pose and illumination changes jointly. To reduce the impacts of pose changes on re-identification, we estimate the joints' positions of a person based on the deep learning technique and thus make it possible to extract features on specific body parts with high accuracy. Based on the skeleton information, we design a set of local color comparison-based cloth-type features, which are resistant to various lighting conditions. Moreover, to better evaluate SKEPRID, we build the PO&LI² dataset, which has large pose and illumination diversity. Our experimental results show that SKEPRID outperforms state-of-the-art approaches in the case of strong pose and illumination variation.

CCS Concepts: • **Computing methodologies** → **Computer vision tasks; Computer vision problems; Information systems** → *Retrieval models and ranking*;

Additional Key Words and Phrases: Person re-identification, pose recognition, cloth-type features

ACM Reference format:

Tuo Yu, Haiming Jin, Wai-Tian Tan, and Klara Nahrstedt. 2018. SKEPRID: Pose and Illumination Change-Resistant Skeleton-Based Person Re-Identification. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 4, Article 82 (October 2018), 24 pages.

<https://doi.org/10.1145/3243217>

¹SKEPRID refers to SKEleton-based Person Re-ID.

²PO&LI refers to POse & LIghting.

This research was funded by the National Science Foundation under Award No. CNS-1330491.

Authors' addresses: T. Yu, University of Illinois at Urbana-Champaign, Department of Computer Science, Urbana, IL, 61801, USA; email: tuoyu2@illinois.edu; H. Jin, Shanghai Jiao Tong University, John Hopcroft Center for Computer Science, Department of Electronic Engineering, Shanghai, 200240, China; email: jinhaiming@sjtu.edu.cn; W.-T. Tan, Cisco Systems, Inc. San Jose, CA, 95035, USA; email: dtan2@cisco.com; K. Nahrstedt, University of Illinois at Urbana-Champaign, Department of Computer Science, Urbana, IL, 61801, USA; email: klara@illinois.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Association for Computing Machinery.

1551-6857/2018/10-ART82 \$15.00

<https://doi.org/10.1145/3243217>

1 INTRODUCTION

Person re-identification (re-id) has been a hot research topic for years. Its task is to decide the similarity between two images of human bodies, so that for each input image, we can find the best match from an image pool. Usually, the images are taken by different cameras, or by the same camera but at different times. The technique of re-id can be potentially applied to surveillance systems, which are deployed in public places to provide security assurance or person tracking, e.g., tracking Alzheimer patients to avoid wandering in healthcare facilities.

However, currently person re-id is still challenging. One challenge is caused by the changing poses of people, which make it hard to find the corresponding points for comparison across two persons' images. For instance, for the people doing exercises in a gymnasium, the action of bending down makes it hard to automatically recognize their upper bodies. The other challenge is that the various lighting conditions make color-based features unreliable. For instance, when a person wearing a white shirt goes into a cafe, the shirt looks yellow under the warm white light. This makes it harder to re-identify the person when he is captured by an outdoor and an indoor camera, separately. Some other examples are shown in Figure 1.

For the problem caused by various poses, some previous approaches [6–9] estimate poses roughly but do not work well for significant pose variation, such as bending down and opening arms. The depth camera or 3D camera-based pose estimation [10, 11] achieves good accuracy, but it is impractical for surveillance systems, because these cameras are expensive and have not been widely deployed. Moreover, most of the previous re-id approaches focus on pedestrians, who have only relatively small pose changes such as swinging arms or legs. However, in many scenarios people make much more significant pose changes, such as the soccer fans jumping and cheering in stadiums. For the problem of various illumination, texture-based features are naturally resistant to lighting changes [12, 13]. However, the color information should still be used to better distinguish people. The interesting approach in Reference [14] shows that the high-level binary features (e.g., wearing skirt or not, wearing sunglasses or not), extracted from color and texture information, have good robustness to strong illumination variation. As other traditional methods, this approach does not work well for significant pose changes. Some previous approaches adopt deep neural networks [15–19], but they do not specifically focus on the impacts of pose and illumination changes simultaneously for re-id.

In this article, we present SKEPRID, a skeleton-based person re-identification method that is robustly resistant to strong pose changes (e.g., crouching, raising arms) and illumination changes (e.g., color changes of light sources). The main idea is that we first locate the joints' positions (e.g., elbows, knees) of a person, and then we extract the illumination-independent features that we have designed at specific body parts with high accuracy.

As mentioned above, the main challenge is that the various poses enhance the difficulty of feature extraction, and color-based features are unreliable in the face of strong illumination changes. With SKEPRID, we solve the two problems jointly. For a 2D person image, we use a deep learning method to estimate the person's *skeleton*, which consists of the coordinates of joints' positions. Based on the skeleton, we can extract features at any specific body parts accurately, which removes the consideration of poses for re-id. Moreover, relying on the joints' positions, we design new binary/ternary *cloth-type features*, which are based on local color comparison and thus insensitive to the illumination difference among images.

The design of SKEPRID is based on the assumption that the clothes of identities do not change among different camera views, which is common among most surveillance camera-based re-id approaches. Since facial recognition is often inapplicable because of non-ideal camera angles or low resolutions (as shown in Figure 1), the appearance information of clothes is mainly relied to re-identify people.

Fig. 1. Impacts of various poses and lighting conditions. The poses cause the inaccurate recognition for body parts, and the various illumination changes the colors of clothes in the images. The images come from the datasets [1–5].

The main contributions of our article are:

- We design a person re-id method that addresses the interference of strong pose and illumination changes jointly.
- We design novel skeleton and local color comparison-based cloth-type features to make SKEPRID robustly resistant to strong illumination variation.
- To evaluate SKEPRID, we build the PO&LI dataset, which includes strong pose and illumination changes.
- We implement a prototype system of SKEPRID and conduct evaluations based on it. Experiment results show that SKEPRID outperforms other state-of-the-art approaches on PO&LI and achieves comparable re-id accuracy on large pedestrian datasets such as CUHK03 [20] and Market-1501 [21].

Besides normal scenarios such as shopping malls, SKEPRID is specifically effective for the public places where people’s poses keep changing, such as gymnasiums, concert venues, and rehabilitation centers. SKEPRID is also effective for large facilities and campuses, such as hospital campuses and universities, where surveillance cameras are often deployed in both outdoor and indoor areas and thus suffer inconsistent illumination.

We organize the rest of our article as follows. After introducing the related work in Section 2, we present an overview of SKEPRID’s framework in Section 3. Then, we successively introduce the methodologies and algorithms used to perform person re-id in Sections 4–6. In Section 7, we introduce the PO&LI dataset. In Section 8, we evaluate the performance of SKEPRID with comprehensive experiments. We have further discussion in Section 9 and then conclude our article in Section 10.

2 RELATED WORK

Person re-identification: Person re-id problem has attracted a lot of attention in the community of computer vision. Typically, a re-id method reads two images containing two people’s full bodies as input, and it outputs a score to indicate the similarity between the two images [15]. Based on the scores, we can find the most similar image from an image pool for an input image. A typical re-id approach usually consists of two parts: a feature extraction method to extract features from input images and a matching method to compare the features of two images [15, 22]. For the feature

extraction method, researchers are engaged in finding the features that can separate each identity from others. For instance, color histogram-based features and local binary features have been used in many re-id approaches [23–25]. For the matching method, many metric learning approaches have been proposed, such as Local Fisher Discriminant Analysis [26], Marginal Fisher Analysis [27], saliency weighted distances [25], and Mahalanobis metric learning [24].

In this article, we focus on the robustness to different lighting conditions and human poses. Therefore, besides our newly designed cloth-type features, we also use texture based features [12, 13], which are invariant to lighting changes. Moreover, by incorporating the skeleton information, we design a new way to extract these features in spite of changing poses. We adopt the Large Margin Nearest-Neighbor Learning (LMNN) [28] to learn the matching metric, and further extend it with the metrics base on the L2-norm and Bhattacharyya distance [29].

Pose estimation-based re-id: Changing human poses makes it hard to locate the corresponding points for comparison between two images. To handle this problem, researchers have proposed to estimate person poses before the feature extraction. Some works provide rough pose estimation methods. For instance, the approach in Reference [6] estimates the symmetrical axes of different body parts. In Reference [7], the orientation of a person's upper body is roughly estimated with the HOG (Histograms of Oriented Gradients) detector. Some approaches apply view-point specific distance metrics and classify person poses into categories [8, 9]. Benefited by the RGB-D camera technique, accurate pose estimation is realized. For instance, in Reference [12] Microsoft Kinect is used to estimate the positions of joints, and texture features are extracted from these positions. However, to estimate the 3D skeletons, either dual cameras or infrared cameras are needed to obtain depth information, which requires expensive upgrade for current surveillance systems. Moreover, most of the commercial depth cameras only work within 10 meters, which is also impractical for the surveillance in public places.

The recent approaches in deep learning community show the possibility of estimating skeletons from 2D images only. For instance, Deeppose [30] takes a 2D image as input, and outputs an estimated skeleton with joints' positions labeled in the image. It is cheaper to apply these deep learning-based approaches to the commercial-camera-based surveillance systems. In this article, we adopt a state-of-the-art pose estimator, named convolutional pose machines (CPM) [31], based on which, we design novel skeleton-based features to obtain strong resistance to pose and illumination variation.

Deep learning-based re-id: Beyond using deep learning techniques to estimate poses only, researchers have also adopted deep neural networks to learn features and metrics. For example, in Reference [15], a deep convolutional neural network is designed to address the re-id problem. In Reference [16], the authors design a siamese deep neural network, and use Binomial deviance to evaluate the cost between similarities and labels. In Reference [32], a deep hybrid similarity learning (DHSL) method is designed based on a convolution neural network (CNN). A CNN learning feature pair is extracted from the input image pair, and a hybrid similarity function is designed to measure the similarity between the feature pair. A filter pairing neural network (FPNN) is proposed in Reference [20] to handle the misalignment caused by the unreliable pedestrian detector, as well as the impacts of background clutter, photometric and geometric transforms. The approaches in References [17, 18] also utilize CPM to locate body joints, and thus align pedestrians to a standard pose, or extract features from different body regions. However, these previous methods often fail to adapt to the significant lighting changes that are prevalent in different parts of a venue. In contrast, while SKEPRID also adopts deep learning techniques to estimate poses, we further design the cloth-type features that are resistant to various lighting conditions. Because the color-based comparison between images can be easily spoofed, we conduct local color comparison within the same image, and use the comparison results as the cloth-type features for identity matching. Our

Fig. 2. System architecture of SKEPRID. The image of a baseball player comes from LSP dataset [37].

experiment results show that SKEPRID outperforms the previous deep learning-based approaches in the face of distinct indoor/outdoor lighting changes.

3 SYSTEM ARCHITECTURE

In this section, we give an overview of SKEPRID's framework and the workflow of our method. By designing SKEPRID, we prove that the challenge of pose and illumination changes can be solved jointly with the help of the skeleton information. To handle the impacts of changing poses on re-id, we first locate the joints' positions of the person, and then crop patches at specific body positions. This implicitly reshapes the body to a basic standing posture, so that the following feature extraction is not impacted by any pose change. To handle the impacts of changing illumination on re-id, we further utilize the extracted joints' positions to design a set of novel color-based features. These features are mainly based on the color comparison among different body patches within the same image, which addresses the unreliability of normal color-based features when the lighting changes a lot across different images.

As shown in Figure 2, our system mainly consists of the following components.

Patch cropper: This component mainly handles the problem of pose changes. It receives video streams from security cameras, and treats them as frame sequences. The human detection sub-component first detects and crops each person from the frames. Then the skeleton extraction sub-component extracts a list of 16 joints' positions for each person. The subcomponent of body-part patch cropping further crops multiple patches according to the joints' positions for the convenience of feature extraction. The validity of each patch is also examined by the patch validation subcomponent.

Feature extractor: The impacts of illumination variation are mainly handled by this component. The input is a set of cropped patches. It extracts four types of features: *cloth-type features*, *cloth-pattern features*, *texture features*, and *color features*. The features are sent to the identity matcher for identity matching.

Identity dataset: We assume that this dataset has stored the extracted features of known identities.

Identity matcher: This component compares the input features with those stored in the identity dataset, and finds the set of stored features that match the input features best.

Next, we introduce these components in detail, respectively.

4 PATCH CROPPER

The key design point of this component is to crop patches from specific body positions and thus remove the consideration of various poses for the future feature extraction. Besides extracting the

Fig. 3. Cropped patches.

person's skeleton, we also need to carefully design the position profiles of patches so that they are fine-grained enough for the feature extraction. Moreover, in consideration of the occlusion among different body parts, the validity of the patches should also be examined. To achieve these goals, the patch cropper receives video streams and processes their frames with the following subcomponents.

4.1 Human Detection

If the camera is deployed in a region with low-light levels, then we first conduct the automatic color leveling for the image to handle the impact of shadows. For each RGB channel of the input image (frame), the bottom 5% of the pixels that have the lowest values are set to be 0, and the top 20% of the pixels with the highest values are set to be 255. Then, the values of the other pixels are re-mapped to the range [0, 255], which is similar to the idea of contrast stretching [33, 34]. This makes the dark regions in the image brighter, and thus reduces the impact of shadows.

We use the HOG detector [35] to detect each person in the image. For each person detected, this subcomponent crops the patch containing the whole body, and passes it to the subcomponent of skeleton extraction. We name this patch as the *whole-body patch*. One example is shown in the left part of Figure 3.

4.2 Skeleton Extraction

To use off-the-shelf security cameras only, which are highly cost-efficient and ubiquitously deployed, we apply CPM [31] to extract skeletons from 2D images. The approach of CPM applies multistage convolutional networks, which produce image features and belief maps in each stage. The belief maps provide a non-parametric encoding of the spatial uncertainty of location, and therefore enable CPM to obtain rich image-dependent spatial models of the relationships among body parts. The multistage architecture can be trained in an end-to-end fashion using backpropagation.

We train CPM using the training samples from both MPII Human Pose Dataset [36] and Leeds Sports Pose (LSP) Dataset [37]. We randomly select 19,363 images from the two datasets (17,928 from MPII and 1,435 from LSP). Note that MPII and LSP have different joint number profiles (16 versus 14 joints). For consistency, we use the 16-joint profile of MPII. For the images from LSP, we generate the two extra joints (upper neck and pelvis) by averaging their neighboring joints' coordinates. Our test results show that CPM has good robustness and works well enough with low resolution images.

Once the skeleton extraction subcomponent receives a whole-body patch from the human detection, it runs the CPM using the patch as input, and gets a list of 2D joint coordinates.

4.3 Body-Part Patch Cropping

Based on the joints' positions, we crop 16 square *joint patches*, denoted by P_0-P_{15} , as shown in Figure 3. To handle the various sizes of whole-body patches, the side length of each joint patch is set to be equal to 0.1 multiplied by the maximum of the four limbs' lengths. However, using the patches on joints' positions alone is not sufficient and would cause information loss. We further crop two types of patches according to the need of the following feature extraction.

For the purpose of cloth-pattern analysis (in Section 5.2), we crop the *body patch* surrounded by the joints 2, 3, 13, 12 in Figure 3, and reshape it to a 100×100 square patch (denoted by P_{body}) using perspective transformation.

For the convenience of cloth-type feature calculation (in Section 5.1), we also crop 8 more *limb patches* at the middle points between joint 11-12, 13-14, 10-11, 14-15, 1-2, 3-4, 0-1, and 4-5, denoted by P_{upperarm}^l , P_{upperarm}^r , P_{forearm}^l , P_{forearm}^r , P_{thigh}^l , P_{thigh}^r , P_{shank}^l , and P_{shank}^r , respectively, as illustrated by Figure 3.

4.4 Patch Validation

Because of various viewing angles and human poses, it is common that some body parts may be covered by others, and thus not all of them are visible. If we use RGB-D cameras, then we can easily figure out whether a body part is blocked or not with the depth information. However, in CPM the extracted joints from a single image have 2D coordinates only. This causes some issues such as it is hard for us to tell whether a person's swinging arm is covering his upper body, or the other way around. In practice, we regard a patch as covered or overlapped if more than 30% of its area overlaps other patches, despite whether the body part in the patch covers or is covered by the other body parts. The overlap ratio is easy to calculate, because all the patches are in the shape of rectangles.

Fortunately, usually the sleeves and the trouser legs of a person are symmetric, which implies we can still get reliable patches from at least one side of the body. For instance, if we find that the patches on the left arm are overlapped with the body patch, usually the right arm is uncovered, and we can still crop joint and limb patches from it. The patches on the left arm are labeled as *invalid* in this case. The same rule also applies to the patches of legs. If the symmetric patches on two arms or legs cover each other only, then we can regard both the patches as valid, since the cloth patterns on the symmetric body positions are usually similar. Even if the sleeves or the trouser legs of the person happen to have asymmetric lengths, in Section 9.1, we will show that SKEPRID can still maintain good performance.

We invalidate a body patch in two cases: the body patch is overlapped by other body parts, or the patch before reshaping is too narrow or too wide, i.e., the distance between joint 12, 13 is much shorter compared with that between joint 2, 12, and vice versa. If the body patch is invalid, then we use one of the valid upper arm patches as the body patch, which is based on the observation that the cloth pattern on sleeves is often similar to that on the body.

After labeling the validity of the cropped patches, the patch cropper sends the patches to the feature extractor.

5 FEATURE EXTRACTOR

The feature extractor receives the patches P_0-P_{15} , P_{body} , and the eight limb patches from the patch cropper, and it extracts a set of features from them.

Different from the previous approaches that extract features from roughly-estimated body regions, we extract features from the patches at known body positions, and thus achieve a higher

Fig. 4. Examples of local color comparison (we omit P_{upperarm}^l and P_{forearm}^l for brevity). Note that the first two rows indicate the same shirt. The dominant color of the shirt tends to be gray under the sunlight (the first row) but brown under the warm white light (the second row). Thus, the color-based comparison between images with different lighting conditions is unreliable. However, we can correctly recognize the cloth type (long/short sleeves) relying on the local color comparison within the same image.

accuracy. Moreover, we design a set of cloth-type features, which are more reliable than the traditional color-based features in the face of illumination changes.

The extracted features contain four types: cloth-type features, cloth-pattern features, texture features, and color features.

5.1 Cloth-Type Features

Besides the low-level texture features, cloth-type attribute-based modeling has been exploited to provide semantically meaningful features [14]. However, the approach in Reference [14] does not work well if the person has a large-range pose, such as bending down and opening arms. Based on the skeleton information, in SKEPRID, we design a new set of binary/ternary features, which are unambiguous and have better robustness. The intuition is that, since color-based comparison between images with different lighting conditions is unreliable (as shown in Figure 4), we should rely on the local color comparison in the same image. By incorporating the skeleton information, we can more accurately calculate the cloth-type features from the cropped patches. For instance, as illustrated by Figure 4, by comparing the color difference between the patches P_{upperarm}^l , P_{upperarm}^r and P_{forearm}^l , P_{forearm}^r , we can easily judge whether the person is wearing long sleeves or not.

During our experiments, we found that the Euclidean distance-based color comparison among the patches suffers from low precision, which is caused by the shadows on the body. We solve this problem by computing the included angle between two color vectors in different color spaces, and using the angle as a new color comparison metric, which will be shown in detail in the following subsections. This method removes the impacts of luminance changes on re-id and thus has better robustness to shadows.

Note that for cloth-type features, we focus on consistency instead of correctness. For example, for a person wearing a sleeveless shirt, our method recognizes him to be wearing long sleeves. However, since the person is always recognized as wearing long sleeves in different images, the consistency of the feature extraction results is maintained and the person can still be correctly re-identified.

We describe how to extract the cloth-type features in detail.

ALGORITHM 1: *dominant_color***Input:** Patches P_1, P_2 ;**Output:** Dominant RGB color vector $\mathbf{v}$;

```

1 if both  $P_1, P_2$  are valid then
2   |  $P_3 \leftarrow [P_1, P_2]$ ;
3 else if one of  $P_1, P_2$  is valid then
4   |  $P_3 \leftarrow$  the valid patch  $P_1$  or  $P_2$ ;
5 else
6   | return null;
7 Calculate the dominant color  $\mathbf{v}$  in  $P_3$  based on k-means clustering ( $k = 3$ ).  $\mathbf{v}$  is a vector in the RGB space
  and represents the color in the center of the largest cluster;
8 return  $\mathbf{v}$ ;

```

ALGORITHM 2: Feature Extraction for F_{sleeve} **Input:** Patches $P_{\text{upperarm}}^l, P_{\text{upperarm}}^r, P_{\text{forearm}}^l, P_{\text{forearm}}^r$;**Output:** Feature F_{sleeve} ;

// Calculate included angle between dominant colors

```

1  $\mathbf{v}_{\text{upperarm}} \leftarrow \text{dominant\_color}(P_{\text{upperarm}}^l, P_{\text{upperarm}}^r)$ ;
2  $\mathbf{v}_{\text{forearm}} \leftarrow \text{dominant\_color}(P_{\text{forearm}}^l, P_{\text{forearm}}^r)$ ;
3 if  $\mathbf{v}_{\text{upperarm}}$  or  $\mathbf{v}_{\text{forearm}}$  is null then
4   | return  $F_{\text{sleeve}} \leftarrow 0.5$ ;
5  $\theta_{\text{RGB}} \leftarrow \arccos(\frac{\mathbf{v}_{\text{upperarm}} \cdot \mathbf{v}_{\text{forearm}}}{\|\mathbf{v}_{\text{upperarm}}\|_F \|\mathbf{v}_{\text{forearm}}\|_F})$ ;
6 Convert  $\mathbf{v}_{\text{upperarm}}, \mathbf{v}_{\text{forearm}}$  to YCbCr space, and then remove Y elements. The results are labeled by
   $\mathbf{v}'_{\text{upperarm}}, \mathbf{v}'_{\text{forearm}}$ ;
7  $\theta_{\text{CbCr}} \leftarrow \arccos(\frac{\mathbf{v}'_{\text{upperarm}} \cdot \mathbf{v}'_{\text{forearm}}}{\|\mathbf{v}'_{\text{upperarm}}\|_F \|\mathbf{v}'_{\text{forearm}}\|_F})$ ;
8 // Calculate and output the feature
9 if  $\|\mathbf{v}_{\text{upperarm}}\|_F \geq v_T$  and  $\|\mathbf{v}_{\text{forearm}}\|_F \geq v_T$  then
10  | if  $\theta_{\text{RGB}} + \theta_{\text{CbCr}} < \theta_T$  then
11    |  $F_{\text{sleeve}} \leftarrow 1$ ;
12  | else
13    |  $F_{\text{sleeve}} \leftarrow 0$ ;
14 else
15  | if  $\theta_{\text{CbCr}} < \theta'_T$  then
16    |  $F_{\text{sleeve}} \leftarrow 1$ ;
17  | else
18    |  $F_{\text{sleeve}} \leftarrow 0$ ;
18 return  $F_{\text{sleeve}}$ ;

```

5.1.1 Long Sleeves or Not. The main idea is to compare the dominant color of the two upper arm patches $P_{\text{upperarm}}^l, P_{\text{upperarm}}^r$ with that of the two forearm patches $P_{\text{forearm}}^l, P_{\text{forearm}}^r$, and then determine whether the person has long sleeves or not.

The calculation procedure is shown in Algorithms 1 and 2. Algorithm 1 defines a function that combines two given patches as a single image (line 1-6), and then computes its dominant color in

the RGB space (line 7). Note that if one of the patches is invalid, we calculate the dominant color for the valid one (lines 3 and 4).

In Algorithm 2, we first compute the dominant colors $\mathbf{v}_{\text{upperarm}}$, $\mathbf{v}_{\text{forearm}}$ of the upper arm patches and the forearm patches, respectively (lines 1 and 2). Then, we calculate the included angle θ_{RGB} between the two vectors (line 5). To better handle the impact of shadows, we also calculate the included angle θ_{CbCr} after converting $\mathbf{v}_{\text{upperarm}}$, $\mathbf{v}_{\text{forearm}}$ to the YCbCr space and removing luma components (lines 6 and 7). We use F_{sleeve} as the binary feature that describes whether the person has long sleeves or not. We calculate F_{sleeve} in lines 8–17, where $\|\cdot\|_F$ denotes the Frobenius norm, and v_T , θ_T , θ'_T are three predefined constant thresholds. The main idea is that, if θ_{RGB} and θ_{CbCr} are small, the person is probably wearing long sleeves ($F_{\text{sleeve}} = 1$), otherwise short sleeves ($F_{\text{sleeve}} = 0$). Note that, if the cloth color is too dark, the error of θ_{RGB} is large. Therefore, if $\|\mathbf{v}_{\text{upperarm}}\|_F$ or $\|\mathbf{v}_{\text{forearm}}\|_F$ is too small (i.e., $< v_T$), then we use θ_{CbCr} alone to compute F_{sleeve} (lines 14–17). Moreover, if both P_{upperarm}^l and P_{upperarm}^r or both P_{forearm}^l and P_{forearm}^r are invalid, then we set $F_{\text{sleeve}} = 0.5$ (lines 3 and 4), so that the invisible sleeves' F_{sleeve} has equal distances to the “long sleeve” state and “short sleeve” state.

One concern would be that F_{sleeve} will change if the person happens to roll up the sleeves. We address this problem in Section 9.1 and prove that SKEPRID still maintains a good accuracy even if the cloth type is changed by rolling up sleeves.

5.1.2 Long Trouser Legs or Not. The calculation for trouser legs is similar to Algorithm 2, but based on P_{thigh}^l , P_{thigh}^r , P_{shank}^l , and P_{shank}^r instead. We use F_{leg} to denote this extracted feature, i.e., we set $F_{\text{leg}} = 1$ if we regard the person as wearing long trousers, otherwise $F_{\text{leg}} = 0$. Note that SKEPRID recognizes maxi skirts as long trousers, and short skirts as shorts. However, this does not impact the consistency of the binary attribute. Again, if both P_{thigh}^l and P_{thigh}^r or both P_{shank}^l and P_{shank}^r are invalid, we set $F_{\text{leg}} = 0.5$.

5.1.3 Color Similarity Level between Upper Garment and Trousers. We further consider the color difference between the upper garment and the trousers. We calculate the dominant color for the body patch P_{body} and compare it with that of P_{thigh}^l and P_{thigh}^r . As shown in Algorithm 3, with a similar setting as in Algorithm 2, we can evaluate the similarity between the two dominant colors by computing the included angle (lines 1–7). We use F_{gt} to denote the similarity level, which is calculated in lines 8–21. θ_{T2} and θ'_{T2} are two predefined constant thresholds. Therefore, $F_{\text{gt}} = 2$ means the colors of upper garment and trousers are similar, $F_{\text{gt}} = 1$ means the two colors are somehow similar, and $F_{\text{gt}} = 0$ denotes that they are obviously different.

Moreover, we use another feature $F_{\text{gt}}^{\text{sign}}$ to denote whether P_{body} is darker than P_{thigh}^l and P_{thigh}^r . We first convert the three patches to grayscale images. If P_{body} 's grayscale image has a smaller average pixel value than that of P_{thigh}^l and P_{thigh}^r , then we set $F_{\text{gt}}^{\text{sign}} = 1$. Otherwise, we set $F_{\text{gt}}^{\text{sign}} = 0$.

5.2 Cloth-Pattern Features

Sometimes it is insufficient to re-identify people using cloth-type features alone, because people would wear similar types of clothes due to certain weather. To handle this issue, SKEPRID also recognizes the patterns on people's clothes. In the previous approaches, it is hard to recognize cloth patterns, because the shapes of the patterns in images change with different poses. However, with the person's skeleton information, we can accurately locate the region of clothes and reshape it to a square patch, and thus obtain consistent patterns for recognition.

ALGORITHM 3: Feature Extraction for F_{gt}

Input: Patches P_{body} , P_{thigh}^l , P_{thigh}^r ;
Output: Feature F_{gt} ;

// Calculate included angle between dominant colors

1 $\mathbf{v}_{body} \leftarrow \text{dominant_color}(P_{body}, P_{body})$;
2 $\mathbf{v}_{thigh} \leftarrow \text{dominant_color}(P_{thigh}^l, P_{thigh}^r)$;
3 **if** $\mathbf{v}_{body}$ or $\mathbf{v}_{thigh}$ is **null** **then**
4 **return** $F_{gt} \leftarrow 1$;
5 $\theta_{RGB} \leftarrow \arccos(\frac{\mathbf{v}_{body} \cdot \mathbf{v}_{thigh}}{\|\mathbf{v}_{body}\|_F \|\mathbf{v}_{thigh}\|_F})$;
6 Convert $\mathbf{v}_{body}$, $\mathbf{v}_{thigh}$ to YCbCr space, and then remove Y elements. Label the results by $\mathbf{v}'_{body}$, $\mathbf{v}'_{thigh}$;
7 $\theta_{CbCr} \leftarrow \arccos(\frac{\mathbf{v}'_{body} \cdot \mathbf{v}'_{thigh}}{\|\mathbf{v}'_{body}\|_F \|\mathbf{v}'_{thigh}\|_F})$;
// Calculate and output the feature

8 **if** $\|\mathbf{v}_{body}\|_F \geq v_T$ and $\|\mathbf{v}_{thigh}\|_F \geq v_T$ **then**
9 **if** $\theta_{RGB} + \theta_{CbCr} < \theta_T$ **then**
10 $F_{gt} \leftarrow 2$;
11 **else if** $\theta_T \leq \theta_{RGB} + \theta_{CbCr} < \theta_{T2}$ **then**
12 $F_{gt} \leftarrow 1$;
13 **else**
14 $F_{gt} \leftarrow 0$;
15 **else**
16 **if** $\theta_{CbCr} < \theta'_T$ **then**
17 $F_{gt} \leftarrow 2$;
18 **else if** $\theta'_T \leq \theta_{CbCr} < \theta'_{T2}$ **then**
19 $F_{gt} \leftarrow 1$;
20 **else**
21 $F_{gt} \leftarrow 0$;
22 **return** F_{gt} ;

Since upper garments have better pattern diversity compared with trousers, we focus on the patterns on upper garments only. The input is the body patch P_{body} , and two types of features are extracted.

Plain or not: We first recognize whether P_{body} has a plain color only or not. Using k -means clustering ($k = 3$), we find the dominant color and the second-dominant color in P_{body} . With the same metric as in Algorithm 2, if the two colors are the same, P_{body} is plain. However, the shadows generated by the folds on clothes would also cause multiple colors. Therefore, if multiple dominant colors are detected, we further transfer P_{body} to grayscale image, and perform Gaussian smoothing and Canny edge detection [38] to detect the edges. For each pixel on an edge, we generate a 5×5 mask centered on the pixel. The edge then divides the mask to two regions. We compare the colors of the two regions. If the two colors are different, then an edge counter is increased by one. If the total count value is larger than a threshold that is related to the size of P_{body} , then P_{body} is not plain. We set a feature $F_{plain} = 0$ if P_{body} is not plain, otherwise $F_{plain} = 1$.

Pattern features: Whether P_{body} is plain or not, we extract pattern features from it. Since we crop P_{body} based on the skeleton, the rotation variation is naturally well handled. Like the method in [13], we conduct a two-level Haar wavelet transform on the grayscale image of P_{body} to obtain

three detail coefficient matrices. For each matrix, since each pixel has eight neighbor directions, we calculate the gray-level co-occurrence matrices for all the eight directions, as well as the final co-occurrence matrix as the sum of the eight co-occurrence matrices. Then, we calculate the features “mean, contrast, correlation, energy, homogeneity, entropy [39, 40],” based on the final co-occurrence matrix of each detail coefficient matrix. The total 18 features are labeled as $F_0^p - F_{17}^p$.

5.3 Texture Features

In Reference [12], it is shown that the texture features extracted from joint patches can help to re-identify people. In SKEPRID, we use SIFT (Scale-invariant feature transform) [41] as the feature descriptor to extract texture features.

We extract SIFT features for each joint patch. Because of the low resolution of surveillance images, even for the same joint’s position, the number of SIFT key points changes significantly with different poses. Therefore, we fix a single key point at the center of each patch. The extracted texture features are labeled as vectors $F_0^t - F_{15}^t$.

Moreover, although the SIFT-based image matching does not work well for body patches because of the varying folds caused by human poses, we find the number of SIFT key points still useful to indicate the complexity level of the patterns in body patches. Therefore, we use the number of SIFT key points in the body patch as a feature, denoted by F_{kp} .

5.4 Color Features

Although the color comparison between two images is unreliable because of various lighting conditions, it turns out that it is still useful to help distinguish extremely different colors. For instance, usually it is easy to distinguish a blue shirt and a yellow shirt even if lighting conditions vary. To further improve the accuracy of SKEPRID, we calculate the RGB color histograms of the body patch and 16 joint patches, stored as matrices $F_b^c, F_0^c - F_{15}^c$.

For each whole-body patch, all the extracted features are sent to the identity matcher for further re-identification.

6 IDENTITY MATCHER

The identity matcher first needs to learn the metric that quantifies the similarity between two sets of extracted features from two images. For cloth-type features and cloth-pattern features, we apply the widely used LMNN [28]. Based on the k-nearest-neighbors (kNN) classification rule, LMNN learns a Mahalanobis distance metric using semidefinite programming.

For texture features and color features, there exist mature metrics for SIFT key point matching and color histogram comparison. Thus, we calculate the L2-norm and the Bhattacharyya distance to evaluate the difference between the texture features and the color features from two images.

We use the CUHK03 dataset to train the LMNN-based metric. In Section 8.1, we will introduce this dataset and the training method in detail. LMNN returns a Mahalanobis matrix $\mathbf{M}$. We use two vectors $\mathbf{x}, \mathbf{y}$ to denote the feature vectors $[F_{\text{sleeve}}, F_{\text{leg}}, F_{\text{gt}}, F_{\text{gt}}^{\text{sign}}, F_{\text{plain}}, F_0^p, \dots, F_{17}^p, F_{kp}]$ of two images, and thus we can compute the distance between vectors $\mathbf{x}$ and $\mathbf{y}$ as $D_1 = (\mathbf{x} - \mathbf{y})\mathbf{M}(\mathbf{x} - \mathbf{y})^T$. Moreover, we compute the sum of the L2-norm distances for $F_0^t - F_{15}^t$ from two images, and the sum of the Bhattacharyya distances for all the RGB channels of $F_b^c, F_0^c - F_{15}^c$ from two images, denoted by D_2 and D_3 , respectively. Therefore, the total distance between two images is $D_{xy} = D_1 + \alpha D_2 + \beta D_3$, where α and β are two hyperparameters used to control the weight of texture features and color features. We fine-tune them based on the performance of SKEPRID on the validation dataset of CUHK03.

Fig. 5. Sample images from PO&LI dataset.

We assume that a set of ID-features tuples has been stored in the identity dataset. An ID is a unique label assigned to a person, and the features in the tuple are extracted by SKEPRID from the very person's image. Note that there would exist multiple tuples with the same ID, which are generated by the different images of the same person. For each input whole-body patch, the identity matcher computes D_{xy} between the patch's features and those of each tuple in the dataset. The ID of the tuple with the smallest D_{xy} is the re-id result. If the person in the input patch has the same ID with the tuple, then the re-id result is correct.

7 PO&LI DATASET

As we have mentioned, SKEPRID focuses on the scenarios where people's poses and illumination significantly change, such as in gymnasiums or concerts. However, currently we find that most of the datasets [4, 5, 20] for re-id tasks focus on pedestrian re-id only. Usually, pedestrians only have limited pose changes such as swinging arms or legs. Therefore, these datasets are not sufficient enough to fully evaluate SKEPRID. However, some sport or pose datasets [1–3] contain strong pose changes such as running and crouching. However, the records for an identity are usually collected in the same location where the illumination do not change. Therefore, to better compare the performance of SKEPRID with that of other approaches in the case of strong significant pose changes and illumination variation, we build our own PO&LI dataset.

The PO&LI dataset contains 600 images collected from 50 identities by four pairs of camera views. Note that a person wearing two different suits of clothes is considered as two different identities, which will be discussed in Section 9.1. PO&LI contains two subsets, PO and LI. The PO dataset places emphasis on pose changes, while the LI dataset focuses on illumination changes. Examples from the two subsets are shown in Figure 5.

PO dataset: In this subset, we collect 288 images containing different person poses. The data are collected in four scenarios: indoor with warm white light, indoor with cool white light, outdoor with sunlight, and outdoor with cloudy sky. In each scenario, we use one camera to record the motion of 24 volunteers. The location of the camera would change among different records. During the recording, the volunteers are required to keep changing their poses, such as waving arms, kicking and bending down (examples are shown in Figure 5). We randomly capture 12 images from each record, and thus collect 288 images in total. The resolution of the whole-body patches is around 110×220 pixels, which is practical for normal security cameras. Note that while the PO dataset focuses on pose changes, it also contains the illumination diversity, because the images are captured in the four scenarios with different illumination conditions.

Fig. 6. Sample images from CUHK03 [20], Market-1501 [21], and VIPeR [42].

LI dataset: In this subset, we collect 312 images from pedestrians with different illumination conditions. With the same camera and scenario setting as that in the PO dataset, we collect 312 images from 26 pedestrians, and each of them is captured in two of the four aforementioned scenarios. For each record in each scenario, we randomly capture 6 images. In other words, each identity is captured by two cameras, and each camera captures 6 images. The resolution of the whole-body patches is also around 110×220 pixels. While the LI dataset emphasizes illumination change, it also contains basic pedestrian poses such as walking and turning body.

Therefore, compared with the previous datasets, PO&LI is more suitable for evaluating the robustness of re-id approaches in the face of strong pose and illumination changes, because it includes more distinct pose variation (e.g., crouching, raising arms) and lighting variation (e.g., color changes of light sources).

8 EXPERIMENTS

We have implemented a prototype system of SKEPRID on a desktop equipped with a NVIDIA GTX 1080 GPU. We fine-tune the system parameters based on the performance on the validation dataset during training. Specifically, we set $\theta_T = 0.2$, $\theta'_T = 0.1$, $\theta_{T2} = 0.6$, $\theta'_{T2} = 0.3$, and $v_T = 17$ while RGB value ranges in $[0, 255]$. In this section, we evaluate our prototype system by comparing it with the state-of-the-art methods on four datasets: PO&LI, CUHK03, Market-1501, and VIPeR [42].

8.1 Datasets

As we have mentioned in the previous section, to better compare the performance of SKEPRID with that of other approaches, we mainly use our own PO&LI dataset. Moreover, we also evaluate SKEPRID in some commonly-used pedestrian re-id datasets, i.e., CUHK03, Market-1501, and VIPeR.

CUHK03 contains 13,164 images of 1,360 pedestrians collected from five pairs of camera views. In CUHK03, each pedestrian is observed by two disjoint cameras, and each camera captures 4.8 images for each pedestrian on average. The dataset provides manually labeled bounding boxes and automatically labeled bounding boxes. Given that previous approaches achieve their best performance with manually labeled bounding boxes [15], we choose to use the manually labeled ones only. Examples from CUHK03 are shown in Figure 6.

Market-1501 contains 32,668 images of 1,501 identities collected from 6 cameras. Each camera captured 3.6 images of each identity on average, and each annotated identity is guaranteed to be captured by at least two cameras. The Deformable Part Model (DPM) detector is applied to provide

Fig. 7. CMC curves on PO dataset.

Fig. 8. CMC curves on LI dataset.

bounding boxes, while the bounding boxes of query images are hand-drawn. In Figure 6, we also show some examples from Market-1501.

VIPeR contains 1,264 images of 632 identities. For each identity, two images are captured from two arbitrary viewpoints. The dataset contains a significant amount of viewpoint changes. Some image pairs from VIPeR are shown in Figure 6.

Compared with the other pedestrian re-id datasets, PO&LI is more suitable for evaluating the robustness of re-id approaches in the face of pose and illumination changes, because it includes distincter pose variation (e.g., crouching, raising arms) and lighting variation (e.g., color changes of light sources).

Since PO&LI is a small dataset, it is hard to train deep neural networks without overfitting. For fairness, we train SKEPRID and the other methods with CUHK03. As the same protocol used in [15, 20], we randomly select 1,160 identities for training, and 100 identities for validation. The training process takes about 15h to converge. The remaining 100 identities are left as the test set. Note that CPM is still trained with MPII and LSP datasets.

We first conduct experiments on PO&LI to show that SKEPRID outperforms state-of-the-art approaches in the face of distinct illumination and pose variation. Then, we show that SKEPRID can also achieve good re-id accuracy on the other pedestrian datasets, where illumination and person poses do not change a lot.

8.2 Experiments on PO

To evaluate the robustness of SKEPRID to changing person poses, we conduct experiment with the PO dataset. We divide the PO dataset into two subsets, each of which contains 6 images for each identity. Since the goal is to evaluate the impacts of poses, we treat the two subsets as the images captured by two cameras. The images from one camera are used to generate the identity dataset, while those from the other one are used for the re-id test. Since PO&LI does not provide bounding boxes, we use the human detection subcomponent to crop whole-body patches for all the tested methods. The baselines are KISSME [24], eSDC [43], SDALF [6], FPNN [20], Spindle Net [18], and the approach in Reference [15] (denoted by IPLA), among which FPNN, Spindle Net, and IPLA are deep learning-based methods. Note that besides SKEPRID, these baselines are also trained following the protocol used in References [15, 20].

In Figure 7, we plot the commonly used Cumulative Matching Characteristics (CMC) curves (single-shot) for the baselines and our approach. The CMC curve shows how the probability of correct re-id (y-axis) changes with the number of returned candidates (x-axis) [44], i.e., the rank-accuracy represents the probability that the correct re-id result is included in the top-k best matches returned by the re-id method.

The results show that our approach outperforms the other methods significantly, with the rank-1 rate of 71.5%. The rank-1 rates for Spindle Net, IPLA, FPNN, KISSME, eSDC, and SDALF are 51.4%, 31.3%, 20.8%, 14.7%, 12.5%, and 6.9%, respectively. This is because the baselines do not handle the large-range or complex human poses well, which often lead to the failure of feature extraction. In contrast, SKEPRID has adopted the skeleton extraction method to locate body parts, which performs consistently even when the person is doing strenuous exercises. Spindle Net also utilizes the skeleton information, but it only divides the body to seven rectangle regions. We find that if the person has complex poses such as crouching and curling arms, the roughly-cropped rectangle regions lead to low accuracy. SKEPRID avoids this problem by cropping patches at a more fine-grained level.

8.3 Experiments on LI

We further use the LI dataset to evaluate the performance of SKEPRID for various illumination. As mentioned before, in LI each identity is captured by two cameras with different illumination conditions. The images from one camera is included in the identity dataset, and those from the other one are used for the test. Using the same baselines and setting as in Section 8.2, we repeat the test on the LI dataset.

The test results are shown in Figure 8. The rank-1 accuracy of SKEPRID achieves 48.7%, while those for Spindle Net, IPLA, FPNN, KISSME, eSDC, and SDALF are 37.8%, 34.6%, 30.8%, 27.6%, 25.6%, and 28.2%, respectively. Therefore, it is confirmed that, since we mainly rely on lighting-independent features, SKEPRID has better resistance to lighting condition changes.

8.4 Performance of Cloth-Type Features

To better understand the contribution of the binary/ternary cloth-type features, we further design an experiment. We collect images from four volunteers with eight different types of clothes, who are considered as eight different identities. The cloth types of the eight identities are: #1–2 long sleeves and long trouser legs, #3–7 short sleeves and long trouser legs, #8 long sleeves and short trouser legs. Note that although some volunteers are wearing the same type of clothes, the patterns on their clothes are still different from each other, so that they are still differentiable. In each of the four scenarios as mentioned before, we take four images for each identity, during which the person is free to make different poses.

We first show the overall performance of SKEPRID, before we analyze the contribution of each cloth-type feature. Figure 9 shows the distance map for all the images. Each pixel denotes the distance D_{xy} given by SKEPRID between two images. Red color stands for small distance and green color stands for large distance. The pixels in each white cell denote the comparison between the images from two scenarios. The pixels in each black cell denote the comparison between the images of two identities. The detailed examples are shown in Figure 9. Note that the black cells along the diagonal show the comparison between the images of the same identity. For convenience, we label the minimum-distance pixel in each row with white color (besides the pixels on the diagonal, which always have zero distance). Figure 9 also shows an example of the white pixel. It is obvious that almost all the white pixels are within the black cells along the diagonal. This shows that SKEPRID can correctly re-identify most of the images. Moreover, the white pixels are not tightly blocked by the diagonal white cells, but they are mostly bounded by the diagonal black cells. Since the images are taken in four scenarios with different lighting conditions, it is shown that SKEPRID can correctly re-identify a person even with various illumination.

To show the contribution of cloth-type features, we increase the weight of each feature, respectively. The results are shown by Figure 10. Although there exist some misjudgements caused by extremely dark shadows, it is obvious that the identities with the same cloth type usually have

Fig. 9. Distance map. The example of the white cell shows the comparison results between the images of identity 2 in scenario 2 and the images of identity 8 in scenario 3. The example of the black cell shows the comparison results between the images of identity 4 and those of identity 8 in all the four scenarios. The example of the white pixel shows that the fourth image of identity 7 in scenario 1 has the shortest distance to the fourth image of the same identity in scenario 4.

shorter distances to each other. Therefore, by using each cloth-type feature, we can effectively recognize the type of clothes that the person is wearing. The extracted cloth-type information can help to distinguish identities with high accuracy, which confirms the contribution of our newly designed features.

8.5 Experiments on CUHK03, Market-1501 and VIPeR

Although the design of SKEPRID focuses on handling the variation of illumination and poses, we show that SKEPRID also performs well for pedestrian re-id in normal scenarios where illumination and people's poses do not change a lot. We take the 100 identities from the test set in CUHK03 and conduct the same experiment as in Section 8.2. Since some of the baselines' performance on CUHK03 has been reported in Reference [15], and SKEPRID is trained and tested based on the same protocol, we cite their test results.

As shown in Figure 11, while SKEPRID does not outperform Spindle Net, it achieves higher rank-1 accuracy (61.5%) than the other baselines. The reason why SKEPRID does not achieve the best performance on CUHK03 is that CUHK03 mainly contains the surveillance images from a campus, where many students carried bags. The bags would block the body or joint patches, and thus impact the corresponding features. We leave it to our future work to design strategies to eliminate the impacts of bags.

We further evaluate the performance of SKEPRID on Market-1501 and VIPeR. Market-1501 has 19,732 images for testing, and provides 3,368 query images of 751 identities. The same test protocol in Reference [21] is followed. In VIPeR, we randomly select the images of 316 identities for testing, which follows the same protocol in Reference [44]. The test results as well as the reported performance of some other baselines are shown in Tables 1 and 2. Note that the baselines in the tables still follow their own training protocols. Similar to the test results on CUHK03, SKEPRID

Fig. 10. Distance maps relying on cloth-type features only.

outperforms most of the baselines, and achieves an accuracy that is comparable with those of the most recent approaches.

As mentioned before, SKEPRID is designed to re-id people who make significant pose changes, and work in the scenarios where illumination changes a lot. Thus, SKEPRID is not designed for pedestrian re-id only. CUHK03, Market-1501, and VIPeR are pedestrian re-id datasets, in which the identities have walking poses only, such as swinging arms or legs. However, in the scenarios such as athletic contests and concerts, people have much stronger pose changes. For instance, soccer fans jump and cheer during a match, and fans wave arms during a concert. The three pedestrian datasets do not contain such strong pose changes. In contrast, the PO dataset contains various poses and thus provides a better benchmark to evaluate the performance of SKEPRID. However, although the images in VIPeR contain illumination changes, each image pair is taken in the same scenario with the same light source. The illumination change within an image pair is mainly caused by shadows. In the LI dataset, the images of an identity are collected from different scenarios with different light sources. Thus, LI provides better illumination diversity to evaluate SKEPRID's performance. Although on the three pedestrian datasets SKEPRID does not outperform some baselines that are specifically designed for pedestrian re-id, it has better performance on the PO&LI dataset, which is more suitable to evaluate approaches' robustness to stronger pose and illumination changes. Therefore, according to the test results in Sections 8.2 and 8.3, we confirm that SKEPRID has better performance in the face of changing poses and lighting conditions.

Fig. 11. CMC curves on CUHK03 dataset.

Table 1. Comparison on Market-1501 Dataset

Methods	Rank 1	Rank 5	Rank 10	Rank 20
LOMO + XQDA [45]	43.79	—	—	—
BoW+Kissme [21]	44.42	63.90	72.18	78.95
WARCA [46]	45.16	68.12	76.00	84.00
TMA [47]	47.92	—	—	—
LDNS [48]	55.43	—	—	—
HVIL [49]	78.00	—	—	—
PersonNet [50]	37.21	—	—	—
DGDropout [51]	59.53	—	—	—
Gate S-CNN [52]	65.88	—	—	—
LSTM S-CNN [53]	61.60	—	—	—
PIE [17]	79.33	90.76	94.41	96.65
Spindle Net [18]	76.90	91.50	94.60	96.70
PDC [19]	84.14	92.73	94.92	96.82
SKEPRID	72.56	84.43	89.76	92.19

Table 2. Comparison on VIPeR Dataset

Methods	Rank 1	Rank 5	Rank 10	Rank 20
MLAPG [54]	40.73	—	82.34	92.37
LOMO + XQDA [45]	40.00	67.40	80.51	91.08
BoW [21]	21.74	—	—	—
WARCA [46]	40.22	68.16	80.70	91.14
LDNS [48]	42.28	71.46	82.94	92.06
IDLA [15]	34.81	76.12	—	—
DGDropout [51]	38.6	—	—	—
SI+CI [55]	35.80	67.40	83.50	—
Gate S-CNN [52]	37.80	66.90	77.40	—
LSTM S-CNN [53]	42.40	68.70	79.40	—
MTL-LORAE [56]	42.30	72.20	81.60	89.60
Spindle Net[18]	53.80	74.10	83.20	92.10
PDC [19]	51.27	74.05	84.18	91.46
SKEPRID	49.87	71.92	83.11	90.82

8.6 Impacts of Low Resolution

For the application scenario of surveillance systems, it is common that only low-resolution images are available. Even if high-quality cameras are deployed, the resolution of a person's image is still limited if the person is far away from the cameras. Moreover, the resolution for the same person may also vary among different cameras. Therefore, it is important that our method is resistant to various resolutions. To evaluate the performance of SKEPRID in the case of different resolutions, we repeat the test in Section 8.2, but with different resolution levels: 110×220 pixels and 50×100 pixels for the whole-body patches. Also, we consider the case that the first camera captures people with 50×100 pixels while the second one with 110×220 pixels.

The experimental results in Figure 12 show that the performance of SKEPRID is not seriously impacted by the various resolution. The slight decrease in rank-1 rate is inevitable, because some cloth-pattern details are lost due to the low resolution. However, since the cloth-type features still work well, SKEPRID can achieve the rank-1 accuracy of 63.2% even in the worst case, which still outperforms the baselines in Section 8.2.

Fig. 12. Impacts of different resolutions.

9 DISCUSSION

9.1 Person Appearance-Based Re-Identification

A common implicit assumption of most surveillance camera-based re-id approaches is that the clothes of identities do not change among different camera views. Since a person's face image is often unavailable, because the person is not facing the camera or the resolution is low, the facial recognition approaches [57, 58] usually do not apply to the surveillance systems. Therefore, many re-id methods only rely on appearance information (e.g., color, texture). SKEPRID is also based on this assumption, and thus we focus on appearance-based re-id only.

To address the concern that SKEPRID would be impacted by the action of rolling up sleeves, we repeat the experiments on PO and LI in Sections 8.2 and 8.3, and set $F_{\text{sleeve}} \leftarrow 1 - F_{\text{sleeve}}$ for all the test data. This is equivalent to the extreme case that each person changes the cloth type by rolling up or putting down sleeves. The results show that the rank-1 accuracy of SKEPRID achieves 58.8% on PO and 41.6% on LI, which are still higher than those of the best baseline (51.4% and 37.8%). Therefore, SKEPRID has the robustness to the action of rolling up sleeves. The action of rolling up sleeves would impact the feature of F_{sleeve} . However, despite the failure of a single feature, as long as the other features still work, SKEPRID still maintains better performance than that of the baselines. Moreover, we repeat the test by setting $F_{\text{leg}} \leftarrow 1 - F_{\text{leg}}$. The rank-1 accuracy is 64.6% on PO and 44.3% on LI, which are also higher than those of the baselines.

Furthermore, the previous tests naturally cover the case that the sleeves or the trouser legs have asymmetric length. Consider the unequal-length sleeves for example. If the left short sleeve is blocked in one image and then the right long sleeve is blocked in another image, then F_{sleeve} would be impacted. However, in the previous test, we have shown that even if F_{sleeve} is flipped, SKEPRID still works well. This further confirms that SKEPRID can maintain a good performance even if the cloth types are slightly changed among different images.

9.2 Accuracy of Human Detection

It is important to note that the current HOG descriptor still suffers from high false negative and false positive rate. During the experiment, we observe that if a person is raising his arm or leg, the HOG would fail to detect the person. Meanwhile, sometimes the other objects in the background would also be detected as humans, which accounts for around 10% of all the cropped whole-body patches. To handle this high false positive rate, we check the validity of the skeleton extracted from each whole-body patch. For instance, if the size of body patch is much smaller compared

with the sizes of arms and legs, we remove this cropped patch from the dataset. In fact, most re-id approaches do not focus on the accuracy of human detection and work on cropped images directly. Since there are many improved human detection approaches [59–61] that can be applied to SKEPRID, we consider it out of our article’s scope to further improve the accuracy of human detection.

9.3 Impacts of Carried Objects

As we have mentioned, currently SKEPRID has no specific algorithm to handle the impacts of the objects carried by people. For instance, bags would block patches and thus impact corresponding features. However, in Section 8.5 it is shown that SKEPRID still achieves good performance on CUHK03, Market-1501 and VIPeR, where a large number of pedestrians carry bags or backpacks. This is because the carried objects only block limited number of patches, and the features extracted from other patches still work well. Moreover, there have been techniques to detect the existence of carried objects [62–64]. Based on these techniques, we can locate the bags and reduce their impacts by invalidating the blocked patches. In future, we will design new object detection strategies for the skeleton-based re-id to further improve the performance of SKEPRID.

9.4 Gait-Based Individual Recognition

It has been shown that humans’ walking properties can help to distinguish different people [65, 66]. Since we can estimate skeletons from surveillance images, it is possible to monitor the gait of a person and thus help to re-identify the person. However, the challenge is that the extracted skeleton is in 2D space instead of 3D space, which causes some vagueness, e.g., if a camera captures a person from the side view, it is hard to judge whether the person is facing left or right. Additional features are needed to rule out this vagueness. However, the computational complexity of the current skeleton estimation algorithm makes it hard to track the gaits online in real time. We leave it to our future work to further explore the usage of gait-based features in surveillance systems.

10 CONCLUSION

In this article, we present SKEPRID, a re-id method that is resistant to strong pose and lighting changes. By incorporating the skeleton information, we reduce the impact of various poses and design a set of skeleton-based illumination-independent features, which significantly improve the re-id accuracy. The experimental results show that, SKEPRID outperforms state-of-the-art approaches on our PO&LI dataset and has good performance on CUHK03, Market-1501, and VIPeR, which confirms the benefit of handling complex poses and various illumination jointly. SKEPRID can be easily applied to the current surveillance systems and is specifically effective for the scenarios such as gymnasiums and campuses, where people’s poses keep changing and outdoor/indoor cameras suffer inconsistent lighting conditions.

ACKNOWLEDGMENT

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the sponsors.

REFERENCES

- [1] S. M. Safdarnejad, X. Liu, L. Udpa, B. Andrus, J. Wood, and D. Craven. 2015. Sports videos in the wild (SVW): A video dataset for sports analysis. In *Proceedings of the IEEE FG*, vol. 1. 1–7.
- [2] J. Liu, J. Luo, and M. Shah. 2009. Recognizing realistic actions from videos in the wild. In *Proceedings of the IEEE CVPR*. 1996–2003.

- [3] S. Ramagiri, R. Kavi, and V. Kulathumani. 2011. Real-time multi-view human action recognition using a wireless camera network. In *Proceedings of the ACM/IEEE ICDSC*. 1–6.
- [4] W. Li, R. Zhao, and X. Wang. 2012. Human re-identification with transferred metric learning. In *Proceedings of the ACCV*. 31–44.
- [5] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. 2011. Custom pictorial structures for re-identification. In *Proceedings of the BMVC*, vol. 5. 6.
- [6] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. 2010. Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the IEEE CVPR*. 2360–2367.
- [7] C. Weinrich, M. Volkhardt, and H.-M. Gross. 2013. Appearance-based 3D upper-body pose estimation and person re-identification on mobile robots. In *Proceedings of the IEEE SMC*. 4384–4390.
- [8] S. Bak, F. Martins, and F. Bremond. 2015. Person re-identification by pose priors. In *Proceedings of the SPIE/IS&T Electronic Imaging, International Society for Optics and Photonics*. 93990H–93990H.
- [9] Y.-J. Cho and K.-J. Yoon. 2016. Improving person re-identification via pose-aware multi-shot matching. In *Proceedings of the IEEE CVPR*. 1354–1362.
- [10] R. Satta, F. Pala, G. Fumera, and F. Roli. 2013. Real-time appearance-based person re-identification over multiple KinectTM cameras. In *Proceedings of the VISAPP*, vol. 2. 407–410.
- [11] I. B. Barbosa, M. Cristani, A. D. Bue, L. Bazzani, and V. Murino. 2012. Re-identification with RGB-D sensors. In *Proceedings of the ECCV*. 433–442.
- [12] M. Munaro, S. Ghidoni, D. T. Dizmen, and E. Menegatti. 2014. A feature-based approach to people re-identification using skeleton keypoints. In *Proceedings of the IEEE ICRA*. 5644–5651.
- [13] S. Yuan, Y. Tian, and A. Arditi. 2011. Clothing matching for visually impaired persons. *Technol. Disabil.* 23, 2, 75–85.
- [14] R. Layne, T. M. Hospedales, S. Gong, and Q. Mary. 2012. Person re-identification by attributes. In *Proceedings of the BMVC*, vol. 2(3), p. 8.
- [15] E. Ahmed, M. Jones, and T. K. Marks. 2015. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE CVPR*. 3908–3916.
- [16] D. Yi, Z. Lei, S. Liao, and S. Z. Li. 2014. Deep metric learning for person re-identification. In *Proceedings of the IEEE ICPR*. 34–39.
- [17] L. Zheng, Y. Huang, H. Lu, and Y. Yang. 2017. Pose invariant embedding for deep person re-identification, arXiv preprint arXiv:1701.07732.
- [18] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. 2017. Spindle Net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the CVPR*.
- [19] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. 2017. Pose-driven Deep Convolutional Model for Person Re-identification. In *Proceedings of the IEEE ICCV*. 3980–3989.
- [20] W. Li, R. Zhao, T. Xiao, and X. Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE CVPR*. 152–159.
- [21] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE ICCV*. 1116–1124.
- [22] S. Khamis, C.-H. Kuo, V. K. Singh, V. D. Shet, and Larry S. Davis. 2014. Joint learning for attribute-consistent person re-identification. In *Proceedings of the ECCV*. 134–146.
- [23] F. Xiong, M. Gou, O. Camps, and M. Szaier. 2014. Person re-identification using kernel-based metric learning methods. In *Proceedings of the ECCV*. 1–16.
- [24] M. Koestinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. 2012. Large scale metric learning from equivalence constraints. In *Proceedings of the IEEE CVPR*. 2288–2295.
- [25] N. Martinel, C. Micheloni, and G. Feresti. 2014. Saliency weighted features for person re-identification. In *Proceedings of the ECCV Workshop on Visual Surveillance and Re-identification*.
- [26] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian. 2013. Local Fisher discriminant analysis for pedestrian re-identification. In *Proceedings of the IEEE CVPR*. 3318–3325.
- [27] S. Yan, D. Xu, B. Zhang, H. J. Zhang, Q. Yang, and S. Lin. 2007. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE TPAMI* 29, 1, 40–51.
- [28] K. Q. Weinberger, J. Blitzer, and L. K. Saul. 2006. Distance metric learning for large margin nearest-neighbor classification. In *Proceedings of the NIPS*, vol. 18. 1473.
- [29] A. Bhattacharyya. 1943. On a measure of divergence between two statistical population defined by their population distributions. *Bull. Calcutta Math. Soc.* 35, 99–109.
- [30] A. Toshev and C. Szegedy. 2014. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE CVPR*. 1653–1660.
- [31] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. 2016. Convolutional pose machines. In *Proceedings of the IEEE CVPR*. 4724–4732.

- [32] J. Zhu, H. Zeng, S. Liao, Z. Lei, C. Cai, and L. Zheng. 2017. Deep hybrid similarity learning for person re-identification. arXiv preprint arXiv:1702.04858.
- [33] S. S. Negi and Y. S. Bhandari. 2014. A hybrid approach to image enhancement using contrast stretching on image sharpening and the analysis of various cases arising using histogram. In *Proceedings of the IEEE ICRAIE*. 1–6.
- [34] B. Xu, Y. Zhuang, H. Tang, and L. Zhang. 2010. Object-based multilevel contrast stretching method for image enhancement. *IEEE Trans. Consum. Electron.* 56, 3.
- [35] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE CVPR*, vol. 1. 886–893.
- [36] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2014. 2D human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE CVPR*. 3686–3693.
- [37] S. Johnson and M. Everingham. 2010. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the BMVC*.
- [38] J. Canny. 1986. A computational approach to edge detection. *IEEE TPAMI* 6, 679–698.
- [39] R. M. Haralick, K. Shanmugam, and I. Dinstein. 1973. Textural features of image classification. *IEEE Trans. Syst., Man Cybernetics*. SMC-3, no. 6.
- [40] H. B. Kekre, S. D. Thepade, T. K. Sarode, and V. Suryawanshi. 2010. Image retrieval using texture features extracted from GLCM, LBG and KPE. *IJCTE* 2, 5, 695.
- [41] D. G. Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the IEEE ICCV*, vol. 2. 1150–1157.
- [42] D. Gray, S. Brennan, and H. Tao. 2007. Evaluating appearance models for recognition, reacquisition, and tracking. *IEEE PETS* 3, 5, 1–7.
- [43] R. Zhao, W. Ouyang, and X. Wang. 2013. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE CVPR*. 3586–3593.
- [44] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. 2015. Learning to rank in person re-identification with metric ensembles. In *Proceedings of the IEEE CVPR*. 1846–1855.
- [45] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the CVPR*. 2197–2206.
- [46] C. Jose and F. Fleuret. 2016. Scalable metric learning via weighted approximate rank component analysis. In *Proceedings of the ECCV*. 875–890.
- [47] N. Martinel, A. Das, C. Micheloni, and A. K. RoyChowdhury. 2016. Temporal model adaptation for person re-identification. In *Proceedings of the ECCV*. 858–877.
- [48] L. Zhang, T. Xiang, and S. Gong. 2016. Learning a discriminative null space for person re-identification. In *Proceedings of the CVPR*. 1239–1248.
- [49] H. Wang, S. Gong, X. Zhu, and T. Xiang. 2016. Human-in-the-loop person re-identification. In *Proceedings of the ECCV*. 405–422.
- [50] L. Wu, C. Shen, and A. V. D. Hengel. 2016. Personnet: Person re-identification with deep convolutional neural networks. arXiv preprint arXiv:1601.07255.
- [51] T. Xiao, H. Li, W. Ouyang, and X. Wang. 2016. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE CVPR*. 1249–1258.
- [52] R. R. Varior, M. Haloi, and G. Wang. 2016. Gated siamese convolutional neural network architecture for human re-identification. In *Proceedings of the ECCV*. 791–808.
- [53] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. 2016. A siamese long short-term memory architecture for human re-identification. In *Proceedings of the ECCV*. 135–153.
- [54] S. Liao and S. Z. Li. 2015. Efficient PSD constrained asymmetric metric learning for person re-identification. In *Proceedings of the ICCV*. 3685–3693.
- [55] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. 2016. Joint learning of single-image and cross-image representations for person re-identification. In *Proceedings of the CVPR*.
- [56] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. 2017. Multi-task learning with low rank attribute embedding for multi-camera person re-identification. In *Proceedings of the IEEE TPAMI*.
- [57] R. N. Rodrigues, G. N. Schroeder, J. J. Corso, and V. Govindaraju. 2009. Unconstrained face recognition using MRF priors and manifold traversing. In *Proceedings of the IEEE BTAS*. 1–6.
- [58] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. 2005. Face recognition using Laplacian faces. *IEEE TPAMI* 27, 3, 328–340.
- [59] X. Wang, Tony X. Han, and S. Yan. 2009. An HOG-LBP human detector with partial occlusion handling. In *Proceedings of the IEEE ICCV*. 32–39.
- [60] W. Chen and J. J. Corso. 2015. Action detection by implicit intentional motion clustering. In *Proceedings of the IEEE ICCV*. 3298–3306.

- [61] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. 2010. Object detection with discriminatively trained part-based models. *IEEE TPAMI* 32, 9, 1627–1645.
- [62] D. Tao, X. Li, S. J. Maybank, and X. Wu. 2006. Human carrying status in visual surveillance. In *Proceedings of the IEEE CVPR*, vol. 2. 1670–1677.
- [63] E. Zhang, Y. Zhao, and W. Xiong. 2010. Active energy image plus 2DLPP for gait recognition. *Signal Process.* 90, 7, 2295–2302.
- [64] D. Damen and D. Hogg. 2012. Detecting carried objects from sequences of walking pedestrians. *IEEE TPAMI* 34, 6, 1056–1067.
- [65] J. Man and B. Bhanu. 2006. Individual recognition using gait energy image. *IEEE TPAMI* 28, 2, 316–322.
- [66] L. Lee and W. Eric L. Grimson. 2002. Gait analysis for recognition and classification. In *Proceedings of the IEEE FG*. 155–162.

Received January 2018; revised May 2018; accepted July 2018