Runner re-identification from single-view running video in the open-world setting

Tomohiro Suzuki[1*], Kazushi Tsutsui[1], Kazuya Takeda[1], Keisuke Fujii[1,2,3]

[1*]Graduate School of Informatics, Nagoya University, Chikusa-ku, Nagoya, Aichi, Japan.
[2]RIKEN Center for Advanced Intelligence Project, 1-5, Yamadaoka, Suita, Osaka, Japan.
[3]PRESTO, Japan Science and Technology Agency, Kawaguchi, Saitama,Japan.

*Corresponding author(s). E-mail(s): suzuki.tomohiro@g.sp.m.is.nagoya-u.ac.jp;
Contributing authors: tsutsui@g.sp.m.is.nagoya-u.ac.jp; kazuya.takeda@nagoya-u.jp; fujii@i.nagoya-u.ac.jp;

Abstract

In many sports, player re-identification is crucial for automatic video processing and analysis. However, most of the current studies on player re-identification in multi- or single-view sports videos focus on re-identification in the closed-world setting using labeled image dataset, and player re-identification in the open-world setting for automatic video analysis is not well developed. In this paper, we propose a runner re-identification system that directly processes single-view video to address the open-world setting. In the open-world setting, we cannot use labeled dataset and have to process video directly. The proposed system automatically processes raw video as input to identify runners, and it can identify runners even when they are framed out multiple times. For the automatic processing, we first detect the runners in the video using the pre-trained YOLOv8 and the fine-tuned EfficientNet. We then track the runners using ByteTrack and detect their shoes with the fine-tuned YOLOv8. Finally, we extract the image features of the runners using an unsupervised method with the gated recurrent unit autoencoder and global and local features mixing. To improve the accuracy of runner re-identification, we use shoe images as local image features and dynamic features of running sequence images. We evaluated the system on a running practice video dataset and showed that the proposed method identified runners with higher

accuracy than some state-of-the-art models in unsupervised re-identification. We also showed that our proposed local image feature and running dynamic feature were effective for runner re-identification. Our runner re-identification system can be useful for the automatic analysis of running videos.

Keywords: person re-identification, sports, computer vision, video processing

1 Introduction

In many sports, it is important to evaluate tactics and movements from videos, which allow us to collect and analyze useful information without interfering with the athlete's movements. Most of the previous studies in recent sports video processing have investigated methods for detecting and tracking players and balls [1–4], and have mainly focused on ball games such as soccer, basketball, ice hockey, and rugby. These approaches can be extended to detecting high-risk plays in rugby [5] and offsides in soccer [6], scoring in rhythmic gymnastics [7] and figure skating [8], and detecting race-walk faults [9, 10]. For such applications for each athlete in videos, athlete detection, tracking, and re-identification are important topics for the automation of the processes.

Player re-identification is the process of identifying each player in a video, which is crucial for automatic video processing in sports. Player re-identification uses the features of the player's image, such as jersey color and/or number [6, 11, 12], or the feature vector obtained from the feature extractor [12–16], to identify players in the multi- or single-video. These studies are not directly applicable to general sports video processing because they are re-identification in the closed-world setting [17] (Figure 1 Top), which focuses on improving the performance of the feature extractor using the prepared image dataset. In the closed-world setting, we can use the image dataset with sufficient labeled data to train the feature extractor. However, the cost of preparing the labeled image dataset from the video is enormous. In addition, real-world re-identification requires the identification of players who are not included in the dataset. Therefore, the closed-world setting is limited for player re-identification in general sports videos, and the open-world setting [17] (Figure 1 Bottom) should be considered. In the open-world setting, we need to directly process the raw image or video, including person image extraction and image feature extraction. It does not require manual data processing of the video. For these reasons, re-identification in the open-world setting is suitable for general sports video processing, especially daily practice videos. For example, in daily practice, many non-professional athletes need to consider the cost of video measurement and processing with a fixed camera. Our research motivation is to develop a runner re-identification system from single-view video in the open-world setting for daily practice use.

In track and field practice, especially in middle and long-distance running and race walking, runners make laps around the same place, such as a track, over and over again. In such situations, fixed-point video allows the runner to be captured many times. If only one runner appears in this video, for example, the runner's three-dimensional

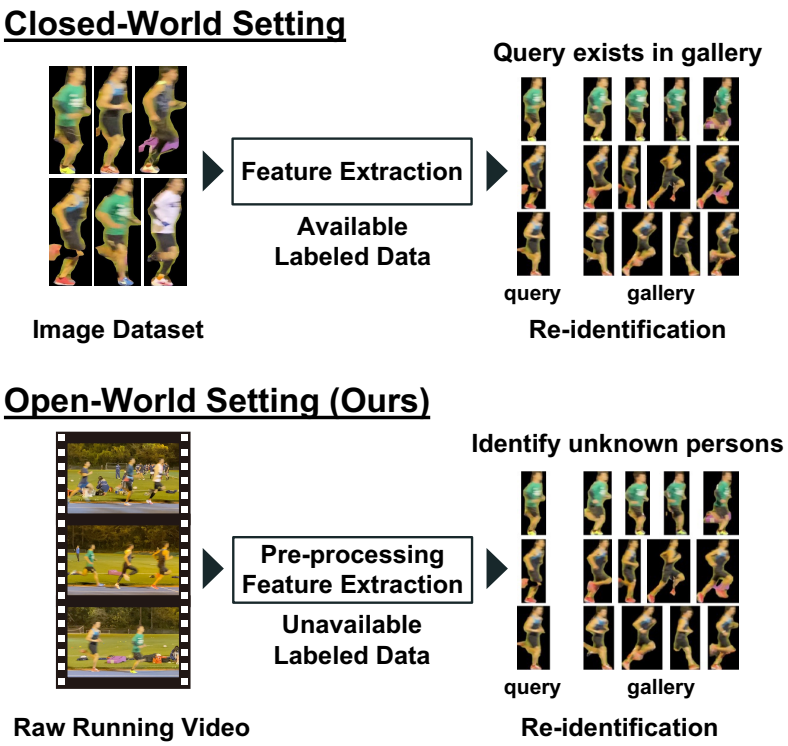

Fig. 1 Overview of the two re-identification tasks. Re-identification in the closed-world setting can use the labeled image dataset, but it is not directly applicable to real-world general sports video analysis. On the other hand, re-identification in the open-world setting, our proposed method, can directly process raw video and does not require labeled data for feature extraction.

joint position information and motion dynamics can be analyzed [18, 19]. However, these analysis techniques cannot distinguish between multiple runners in the video and analyze them simultaneously. To solve this problem, multi-object tracking can be used to identify runners as they pass through the capture area, but the tracking algorithm is unable to identify runners once they disappear from the screen. For this reason, runner re-identification in single-view videos captured at a fixed point plays an important role in the analysis of runner motion in practice videos.

In this paper, we propose a runner re-identification system (see Figure 2) that directly processes single-view video to address the open-world setting. In the system pipeline, we detect runners in the video using pre-trained YOLOv8 [20] and fine-tuned EfficientNet [21, 22]. We then track them using Bytetrack [23] and detect their shoes using fine-tuned YOLOv8. Finally, we extract the runner's image features using the Gated Recurrent Unit AutoEncoder (GRU AE) and global and local mixed features extracted by Hard-sample Guided Hybrid Contrast Learning for Unsupervised Person Re-identification (HHCL) [24]. Our research aims to realize runner re-identification in the open-world setting using single-view running practice video as input and evaluate the results. Since our method uses an unsupervised feature extraction method, it does not require the preparation of labeled data for feature extraction. In addition, our

system consistently and automatically processes runner detection, tracking, and re-identification in the video. These are great advantages for the open-world setting. We use a small amount of labeled data to fine-tune high-performance pre-processing models (runner classifier and shoe detector) required for the proposed system, such as runner detection and shoe detection. However, the labeled data is only needed for the initial fine-tuning of the pre-processing models, and once the pre-processing model is fine-tuned, the labeled data is no longer needed.

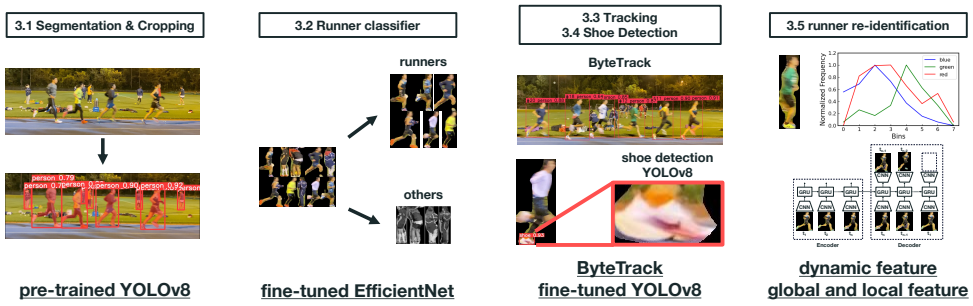

Fig. 2 Our proposed runner re-identification system flow. In Section 3.1, we describe the segmentation and cropping model, and in Section 3.2, we describe the EfficientNet fine-tuned for runner classification. In Sections 3.3 and 3.4, we describe the tracking and the shoe detection separately. In Section 3.5, we describe the runner re-identification.

Our main contributions are:

1. We propose a runner re-identification system from single-view video in the open-world setting, which supports automated video processing and analysis of daily practice. We evaluate the performance of the re-identification pipeline using raw video as input.
2. We propose dynamic features and global and local mixing features obtained by unsupervised method such as GRU AE. Since the training of feature extraction models does not require labeled data, we can reduce the cost of automatic video processing.
3. We show that our proposed method can identify runners in the practice videos with higher accuracy than some SOTA unsupervised re-identification methods. We also show that our proposed global and local feature mixing and dynamic features are effective for runner re-identification.

This paper is organized as follows. First, in Section 2, we provide an overview of the related work. Next, we describe the pipeline of our proposed system in Section 3. Then, we explain our dataset and implementation details in Section 4. Finally, we present the experimental results in Section 5 and conclude this paper in Section 6.

2 Related work

To achieve runner re-identification from video in the open-world setting, we first detect runner passing scenes in the practice video, followed by runner re-identification

between each detected scene. In other words, our research relates to the topics of sports video segmentation and player re-identification.

Segmentation of sports videos includes detecting specific actions (e.g., scoring and penalties) in the video [6, 25–28], detecting a series of actions [29], and extracting the active time in the game [12, 30]. In these approaches, the game situation is determined from the visual and optional auditory cues, or specific actions are detected by the movement of a single player. In contrast, detecting a runner's passing scene would be easier than in the above studies because runners in videos do not make complex movements as in other sports. For example, there is an example of detecting the passage of a runner by using text spotting to recognize the numbers in the image and matching them with the runner's "bib number" registered on the race event website [12]. However, our task is challenging in terms of re-identification from raw videos in the open-world setting without bib numbers.

In player re-identification, many studies have used jersey color and number as features for re-identification [6, 11, 12]. For example, re-identification in running events [12] uses the bib number of the runner in the race to identify each runner. Although the bib numbers are used during the race, they are not available to identify runners in the daily practice videos. In general, the colors of an athlete's clothing and shoes can be more common features than the numbers to characterize them. However, since it is difficult to achieve high-performance re-identification with only these features, recent studies have used feature extractors based on metric learning or contrastive learning [12–16]. Furthermore, gait recognition [31, 32] is also effective for person re-identification, which captures unique features of gait dynamics. Such re-identification methods have contributed to the realization of high-performance player re-identification, but they have the problem of requiring large amounts of data and designing loss functions according to the data to be identified. In addition, many re-identification studies are the closed-world setting [17] that focuses on how to train effective feature extractors using player image datasets and does not consider the open-world setting [17] that identifies players directly from the video, which is important for automatic sports video analysis. Moreover, while some studies of re-identification for open-world settings have addressed unsupervised feature extraction methods [24, 33–36], training of feature extractors with the small number of images that can realistically be acquired [37], and retrieval methods that can handle large amounts of data [38], no studies have evaluated pipeline systems for re-identification that process video directly. To solve the above problems, we implement and evaluate a runner re-identification system that identifies runners directly from video, including pre-processing such as person detection. The feature extractor in our system does not require labeled data. Therefore, our system can directly identify unknown runners in the video.

3 Methods

Our goal is to identify runners in a usual running practice video, even if the runners are framed out multiple times. To identify framed-out runners, we detect the running scenes for each runner in the video. Here we consider a "running scene" as a sequence of frames during which each runner enters the screen from the left and exits from the

right. Videos of runners captured from the side are useful for analyzing the runner's form while running, including the trajectory of the runner's legs. Figure 2 shows the flow of our system in the following five steps. First, persons in each frame of the videos are segmented. We employ the YOLOv8 [20] segmentation model to segment persons. After the segmentation, each person's image is cropped for the next step. Second, cropped images are classified as runners or not using fine-tuned EfficientNet [21, 22]. Third, we track each runner using ByteTrack [23]. Moreover, each runner's shoes are detected using the fine-tuned YOLOv8 shoe detector. Finally, we identify the same runners from each tracked running scene. For the feature extractor, we use a Gated Recurrent Unit AutoEncoder (GRU AE) and global and local feature mixing by HHCL [24] as deep learning-based image features. In addition, we use simple color features, the RGB color histograms of the runner images. For comparison, we use four state-of-the-art (SOTA) unsupervised re-identification methods [24, 33–35] as baseline models. Since we only use labeled data for fine-tuning the runner classifier and shoe detector, our feature extractor is an unsupervised model.

3.1 Person Segmentation and Cropping

To process each person in each frame image individually, we segment and crop person images in each frame. We employ the segmentation model of YOLOv8 [20] trained on the MS COCO dataset [39] to segment persons. The segmented each person's image is cropped using bounding boxes and segmentation masks. The background color of the cropped image is undesired to identify runners. Therefore, we use instance segmentation instead of object detection. Note that, in the running video, the general background removal algorithm did not work well because the shapes of the runners were unclear due to motion blur.

3.2 Runner Classifier

The cropped images obtained from the above procedure include both runners and non-runners. Non-runner images should be removed because they would affect runner re-identification performance. Since the characteristics of runner images captured from the side are the same regardless of the video, a supervised approach does not compromise versatility. We train a runner classifier using labeled cropped person images. For the runner classification, we fine-tuned EfficientNet [21, 22] pre-trained on the ImageNet dataset [40].

3.3 Runner Tracking

In the tracking step, a unique ID is assigned to every runner in each running scene. We employ ByteTrack [23] for runner tracking. The coordinates of the runner's bounding box obtained by the runner classifier are input to ByteTrack. Then, the unique ID of each runner in each running scene is obtained. Note that in this step, the same runner has a different ID for each running scene, as shown in Figure 3, because the tracking algorithm considers the same person reappearing on the screen as a different person.

During tracking, continuous images of each runner's motion for two steps are stored with an ID. We used the correlation between the running motion periods and the width

Fig. 3 Example of tracking results. Representative images and a shoe image are obtained for each runner. Note that different IDs are assigned to the same person who reappears on the screen.

of the runner images to unify the running motion periods of the continuous images. The GRU AE is expected to extract the unique dynamic features of each runner's movements by unifying the period of the running movements.

3.4 Shoe Detection

In general, it is difficult to identify different runners wearing similar color clothes using only whole-body representative images. To address this issue, we detect the shoes of runners, one of the characteristic equipment for them, as a local image feature for runner re-identification.

In order to detect shoes, we employ a fine-tuned YOLOv8 object detection model. We used manually annotated images of runners whose shoes were clearly recognized for training. Figure 6 shows examples of shoes that are "clear" and "unclear". When multiple shoes are detected, the one with the highest confidence score is stored with each runner ID for the following runner re-identification.

3.5 Runner Re-Identification

In runner re-identification, we identify runners based on the similarity of various image features. We use the deep learning-based image feature vector and the color features of the image as image features.

To obtain a deep learning-based image feature vector, we propose a GRU AE that aims to acquire dynamic features from sequential images of running motion. Figure 4 shows the model structure of the GRU AE. We refer to the model structure of the sequence to sequence autoencoder [41]. The encoder of the model transforms the sequence images of running motion into vectors using CNN, which are then fed into the GRU in time series order. The last hidden layer of the GRU is then output to the decoder as a latent variable. The decoder receives the latent variables and the time series images in reverse order and reconstructs the image one frame before the input image by deconvolution of the output of the GRU. In the re-identification step, the encoder is used as a feature extractor, and latent variables are used as feature vectors of the image.

In addition to the GRU AE, we propose a combination of runner and shoe features using HHCL [24]. HHCL [24] is a high-performance unsupervised re-identification

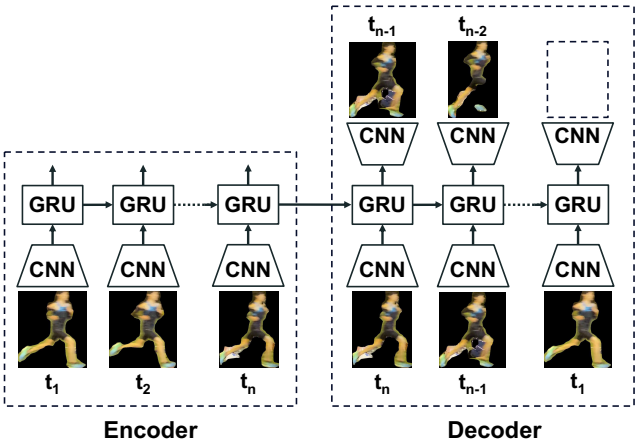

Fig. 4 Model structure of the GRU AE. The encoder receives sequential images of running motion and outputs a 128-dimensional latent variable. The decoder receives latent variable and inverse-ordered sequential images and outputs the reconstructed image one frame before each input image.

method that efficiently learns hard samples (i.e., images that are difficult to identify). The normal HHCL is trained unsupervised using person images. However, in this study, we also train the model using shoe images with the HHCL method. Although this study uses shoes as a local feature of the runner's image, the idea of using local features of the image could be extended to the other re-identification study. We use features from both models trained on runner images and models trained on shoe images for re-identification.

For the color features, the RGB color histograms obtained from the runner images and shoe images are used to identify the runner. The color histograms are obtained by dividing 8 bins for each channel. Note that we ignore the black pixel (background color) when calculating the histograms. The histograms of each channel were flattened to obtain 24 dimensional (= 3 channels $\times$8 bins) color features. For the runner images, since the lower body color features are similar for all runners, only the upper body color features are used.

After feature extraction, the similarity between each feature is calculated. We use it to evaluate the performance of re-identification. The similarity between the deep learning-based feature vectors is defined as the cosine similarity. The similarity between color features is defined as the correlation of histograms. There are two types of color similarity: the similarity of the histogram of the upper body image and the similarity of the histogram of the shoe image. When combining the above two color similarities, 90% of the color similarity of the upper body image and 10% of the color similarity of the shoe image are summed. Furthermore, the combined similarity of GRU AE features and color features is defined as follows:

$$sim_{comb} = \lambda sim_g + (1 - \lambda) sim_c, \tag{1}$$

where l, c denote GRU AE and color, sim denotes similarity, and λ denotes the hyper-parameter.

In the similarity calculation, we introduced lap time filtering as domain knowledge from running training. Once the runners pass in front of the camera, they do not reappear for a while. Therefore, the similarity with the features extracted from running scenes that satisfy the conditions of the following equation is calculated as zero.

$$|S_{c_i} - S_q| > th, \tag{2}$$

where c_i denotes one of the running scenes i, q denotes the query, S denotes the start frame of the running scene, and th is the threshold value of the lap time. Since it depends on situations or prior knowledge, here we set th to 3,600 frames (60 s).

4 Experiments

4.1 Dataset

To build our dataset for the experiments, we captured long-distance running practice videos on the track. The videos were captured at 4K and 60 fps, using an iPhone 11 Pro set up to capture the runner from the side. In each lap, the runners passed through a section of approximately 10 meters, which is the camera's capturing range. Since we were capturing regular practice videos, there were a lot of non-runners, such as other athletes and staff, in the videos.

We prepared three datasets for the experiments. The first dataset is for fine-tuning pre-processing models such as the runner classifier (EfficientNet) and shoe detector (YOLOv8) in the runner re-identification system. For this dataset, we used two videos of 10 minutes each, taken during daytime and nighttime. We cropped each person's image in the videos and annotated whether each person was a runner or not. We also annotated some of the runner images with bounding boxes of running shoes. As a result, we obtained 5,000 images of runners and 12,500 images of people who are not runners. In addition, 600 images of runners were annotated with shoe bounding boxes.

The second dataset is a running scene image dataset used to train GRU AE for unsupervised feature extraction. A running sequence is a series of images of each runner's running motion for two steps extracted from each running scene. We extracted running sequence images from a total of six videos, three each for daytime and nighttime, using the proposed system. As a result, we obtained 1,182 sets of running sequence images for daytime and 1,114 sets of running sequence images for nighttime. Some of these data were also used to train the baseline models [24, 33–35] (see Section 4.2.1 for details). Since the GRU AE and baseline models can be trained using an unsupervised method, no annotation is required for this dataset.

The third dataset was used to evaluate the performance of the proposed runner re-identification system, and was generated from 20 minutes of video for each daytime and nighttime. We extracted the sequence images of each running scene from these videos using the proposed system. We then annotated a unique runner ID for each running scene. Table 1 shows the number of running scenes and unique runners in each video obtained by annotation.

Note that the annotation data was only used for fine-tuning the runner classifier and the shoe detector, and for evaluating the proposed system. We did not use the

Table 1 The number of running scenes and unique runners in each video.

Video type	running scenes	unique runners
Daytime	331	30
Nighttime	330	26

annotation data for training the re-identification model to address the open-world setting. While runner image datasets from race events exist [12, 42], to the best of our knowledge, this is the only dataset for runner re-identification for daily practice. Our dataset does not contain features that can easily identify runners, such as bib numbers used in races. Therefore, runner re-identification in this paper would be more difficult than the race event dataset [12], even though it is in the closed-world setting (note that our task is re-identification in the open-world setting).

4.2 Implementation and Evaluation

We evaluated the performance of the runner classifier, the shoe detector, and the runner re-identification. Our system was implemented in Python 3 with Pytorch. Our code, evaluation dataset, and demo video will be available at https://github.com/SZucchini/runner-reid.

For the fine-tuning of the runner classifier, 14,000 annotated images were used as training data and 3,500 images as test data. In addition, 20% of the training data was used as validation data. The number of runner and non-runner images in the training data was 4000 and 10000, while the number of runner and non-runner images in the test data was 1000 and 2500. We trained the model with a binary cross entropy loss function, employing the Adam optimizer with a learning rate of 3×10^{-5}. Due to the bias in the number of data for runners and non-runners, the F1-score was used to evaluate the classification performance.

For the fine-tuning of the shoe detector, 500 of the images of the runners annotated with shoe bounding boxes were used for training and 100 for testing. We fine-tuned two different pre-trained models of YOLOv8 (yolov8n and yolov8m) and chose the one that showed better detection performance on the test data (yolov8n). The training setting of the fine-tuning was the same as the pre-trained model. Detection performance was evaluated using mean average precision ($\text{mAP}_{50\text{-}95}$) on the test data.

For runner re-identification, we compare the performance of four high-performance unsupervised re-identification models [24, 33–35] (as baseline models) and five proposed feature extractors using mean average precision (mAP) and Cumulative Matching Characteristic (CMC) rank-n accuracy. These evaluation metrics are used in many re-identification studies [13, 16, 17, 24, 33–35, 43]. Since our dataset is smaller than the major re-identification dataset [44–46], we use $n = 1, 5$ in the rank-n accuracy. The feature extractors used to compare performance are baseline models [24, 33–35], color histograms of upper body images (Color hist. w/o shoes), color histograms of upper body images and shoe images (Color hist. w/ shoes), autoencoder with input sequence images of running (GRU AE w/o color hist.), a combination of autoencoder

and color histogram (GRU AE w/ color hist.), and global and local fetures mixing by HHCL (HHCL w/ shoes). The details of each feature extractor are described below.

4.2.1 Baseline models

We used some SOTA unsupervised re-identification models [24, 33–35], which can extract effective image features with unsupervised learning. These model perform well on some re-identification datasets such as Market-1501 [44], MSMT17 [45], and VeRi-776 [46]. In the experiments, we used models trained on our running scene images dataset as a baseline. Since we could not use all images from the dataset due to GPU memory limitations, we used 12,059 runner images from the daytime dataset and 10,229 runner images from the nighttime dataset, respectively. The models were trained separately for each dataset. In the re-identification performance evaluation phase, running scene images from the evaluation dataset were input to the model, and the 2,048-dimensional feature vectors from each image were averaged for each sequence. The cosine similarity of each averaged vector was then calculated as the similarity between each running scene.

4.2.2 Color histogram without shoes

We used only the RGB color histograms of the runner's upper body images as the image feature vector. We resized all images to $(W, H) = (64, 64)$ and defined the upper half of the image as the upper body.

4.2.3 Color histogram with shoes

In addition to the upper body image histograms, the RGB color histograms of each runner's shoe images were used as feature vectors. Each shoe image was resized to $(W, H) = (200, 100)$. The similarity of the upper body vectors and the shoe vectors was calculated separately. We averaged the body color similarity and the shoe color similarity after calculation.

4.2.4 GRU AE without color histogram

We trained the GRU AE using the running scene images dataset. Since the runner in the video looks very different during daytime and nighttime, we trained different models separately to evaluate each video. In both cases, we used 80% of the image sets for training and the rest for validation. We trained the model with mean square error as the loss function, employing the Adam optimizer with a learning rate of 1×10^{-3}. The dimension of the latent variable was 128.

4.2.5 GRU AE with color histogram

We combined the similarity of the image feature vectors in Color w/ shoes with that in GRU AE according to Equation (1). We used mAP as a comparison metric between different lambda. Then we chose a lambda that gives the highest mAP as 0.85.

4.2.6 HHCL w/ shoes

We trained HHCL [24], the SOTA unsupervised re-identification model, on images of runners and shoes, respectively. In other words, we obtained a model that identifies the runner and a model that identifies the shoe separately. The similarity calculated from the features of each model between the runner images and the shoe images was then weighted and averaged at a ratio of 3 : 1, and this similarity was used for re-identification.

5 Results

In this section, we present the results of the performance of the runner classifier, the shoe detector, and the runner re-identification.

For the runner classifier, the F1-score was 99.8%. An example of images misclassified as runners is shown in Figure 5. Most of the misclassified images were of players running in other sports. Misclassification was addressed by filtering the results after runner tracking using frame length and image shape. Filtering resulted in the removal of approximately 98% of non-runner images and incomplete runner images not suitable for re-identification.

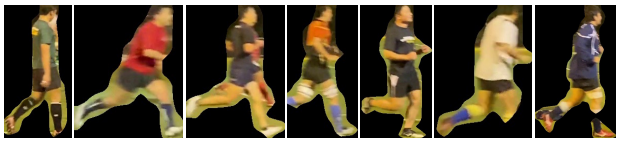

Fig. 5 Examples of images misclassified as runners. Although we accomplished a 99.8% F1-score, most of the misclassifications were images of players from other athletes running.

For the shoe detector performance, the $mAP_{50\text{-}95}$ was 86.2%. As an example of an error, the system detected unclear shoes which is difficult to use to compare the color features for re-identification, as shown in Figure 6. However, since the pre-trained model without fine-tuning could not detect "shoe", which are objects defined in the COCO dataset, fine-tuning was effective for shoe detection in our research.

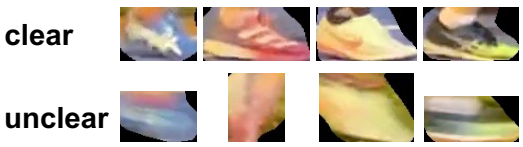

Fig. 6 The shoe detection results. Although we accomplished a higher detection performance (upper), the model sometimes detected unclear shoes (lower).

For the runner re-identification, Table 2 shows the mAP and Rank-n accuracy for each model. In the evaluation for daytime video, the proposed method "GRU AE w/ hist." outperformed all the high performance re-identification methods in the previous studies except HHCL. For daytime videos, the images of runners obtained

by segmentation are clearer and dynamic features are considered easier to extract. In contrast, in the evaluation for night videos, the performance of "GRU AE w/o hist." is significantly lower, indicating that dynamic features are not well extracted. Since the image quality of nighttime video is poor, motion blur appears in the images, and the segmentation accuracy of the runner image extraction is decreased. This makes it difficult for the proposed method to extract the color and dynamic features of the runner images, resulting in poor performance. For this reason, our proposed method may be unstable to changes in the video recording environment compared to previous studies. Based on the above results, we conclude that our proposed dynamic features of running motion are effective in daytime conditions when the images are clear.

While "GRU AE w/ hist." performed well for daytime videos, it could not outperform HHCL, a method that effectively uses hard samples for training. Since the runner dataset in our study has a lot of data from the same runner and many combinations of hard samples, we consider that HHCL was effective. Therefore, our proposed "HHCL w/ shoes", which learns HHCL on runner and shoe images and combines their features, also works well. "HHCL w/ shoes" outperformed all of the proposed and comparative methods, while also showing stable results for both daytime and nighttime videos. This method successfully utilizes the shoe image, a local feature in the image, to achieve more accurate re-identification. The combination of global and local image features is also considered to be effective for general re-identification.

Table 2 Comparison of the re-identification performance with feature extraction models.

Method	Daytime			Nighttime		
	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
SpCL [33]	0.874	0.954	0.985	0.884	0.971	0.974
ICE [34]	0.869	0.960	0.978	0.863	0.965	0.971
HHCL [24]	0.940	**0.985**	**0.988**	0.911	**0.974**	**0.981**
PPLR [35]	0.810	0.957	**0.988**	0.785	0.968	0.974
Color hist. w/o shoes	0.765	0.920	0.954	0.664	0.904	0.942
Color hist. w/ shoes	0.830	0.938	0.963	0.761	0.923	0.955
GRU AE w/o color hist.	0.852	0.957	0.981	0.636	0.875	0.958
GRU AE w/ color hist.	0.900	0.975	**0.988**	0.788	0.946	0.968
HHCL w/ shoes	**0.944**	**0.985**	**0.988**	**0.915**	**0.974**	0.978

To evaluate the effectiveness of our proposed methods, we show a typical example of the re-identification results. Figure 7 shows the top 10 similarity runner images for each of our proposed methods for an example query. First, in Figure 7a, since only the upper body color feature is used for re-identification, 7 of the top 10 similarity runners are different runners from the query wearing the same clothing as the query. In contrast, in Figure 7b, the combination of clothing and shoe color features works effectively, and 8 of the top 10 similarity runners are the same runner as the query. On the other hand, misidentified runners wearing shoes similar to the query are ranked in the top 10 in similarity, such as 6th and 8th similarity. In Figure 7c, the top 8 runners are the same as the query. This result is more accurate than the results using color features (a and b), and the higher Average Precision (AP) shows that more

Fig. 7 Re-identification result examples of our proposed methods. In each row, the similarity is higher for the left image. Runners identical to the query are marked with blue boxes and different runners are marked with red boxes. **(a)** "Color hist. w/o shoes" cannot identify runners wearing clothing similar to the query. **(b)** While an improvement over (a), "Color hist. w/ shoes" cannot identify runners wearing shoes similar to the query. **(c)** "GRU AE w/o color hist." can identify runners with higher similarity than (b). **(d)** "GRU AE w/ color hist." provides more accurate identification by combining dynamic and color features. **(e)** "HHCL w/ shoes" had all images of the same runner with the example query ranked higher in similarity than the images of the other runners.

accurate results are obtained for images outside the similarity range shown in the figure. This indicates that GRU AE's dynamic features other than appearance are more effective. Moreover, in Figure 7d, the combination of the GRU AE features and the color features results in the top 9 out of the top 10 similarities being the same runners as the query. Finally, in Figure 7e, the combination of runner and shoe image features from each trained HHCL model ("HHCL w/ shoes") results in all runners in the top 10 being correct. In addition, "HHCL w/ shoes" has all images of the same runner with the example query ranked higher in similarity than the images of the other runners ($AP = 1.0$).

From the above results, the shoes, which are the unique equipment of runners, would be effective for re-identification. In addition, in the case of daytime video with clear images, the dynamic features extracted by GRU AE functioned as more effective features than the color features. While some runners wear similar clothes and shoes, the dynamic features of the runners (i.e., their running form) are unique to each runner and could be an important feature for runner re-identification. We also found that HHCL is more effective for datasets with many combinations of hard samples.

6 Conclusion

In this paper, we proposed and evaluated a runner re-identification system for a single-view video in the open-world setting. The evaluation results showed that our proposed method could identify runners in running practice videos with higher accuracy than some high performance unsupervised re-identification methods. We also showed that extracting dynamic features of running using the GRU AE and runner-specific local features of the image (e.g., shoes) were effective for runner re-identification in the single-view video. Our runner re-identification system can be useful for the automatic analysis of running videos.

However, we found that the proposed GRU AE was not robust against differences in brightness during video recording. In some cases, it was not possible to distinguish between runners with similar appearances, even when GRU AE dynamic features were used. For future work, a better pre-processing model, such as one that allows accurate runner segmentation exploiting novel segmentation methods [47–49], could be useful to address this issue. Besides, features that are not affected by appearance, such as joint location information, can be effective. In addition, although the GRU AE in this study used a simple model architecture combining CNN and GRU, we consider that more accurate re-identification could be achieved by using more sophisticated representation learning methods such as space-time correspondance learning methods [50, 51], Transformer-based autoencoder [52], or video autoencoder [53] with high-memory GPUs.

Acknowledgments

This work was financially supported by JSPS Grant Number 20H04075 and JST PRESTO Grant Number JPMJPR20CA.

Declarations

Data availability statements

The evaluation datasets generated during and/or analyzed during the current study will be available at https://github.com/SZucchini/runner-reid. However, data from participants who only agreed to participate in this study and did not consent to the release of their data may not be released.

Compliance with Ethical Standards

The participants were fully informed about the study and their consent was obtained in advance. All the experimental procedures were performed after obtaining prior approval for "Experiments on Human Subjects" from the Graduate School of Informatics, Nagoya University.

References

[1] Vandeghen, R., Cioppa, A., Van Droogenbroeck, M.: Semi-supervised training to improve player and ball detection in soccer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3481–3490 (2022)

[2] Cioppa, A., Giancola, S., Deliège, A., Kang, L., Zhou, X., Cheng, Z., Ghanem, B., Van Droogenbroeck, M.: Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3491–3502 (2022)

[3] Scott, A., Uchida, I., Onishi, M., Kameda, Y., Fukui, K., Fujii, K.: Soccertrack: A dataset and tracking algorithm for soccer with fish-eye and drone videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3569–3579 (2022)

[4] Wang, J., Peng, Y., Yang, X., Wang, T., Zhang, Y.: Sportstrack: An innovative method for tracking athletes in sports scenes. arXiv preprint arXiv:2211.07173 (2022)

[5] Nonaka, N., Fujihira, R., Nishio, M., Murakami, H., Tajima, T., Yamada, M., Maeda, A., Seita, J.: End-to-end high-risk tackle detection system for rugby. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 3550–3559 (2022)

[6] Uchida, I., Scott, A., Shishido, H., Kameda, Y.: Automated offside detection by Spatio-Temporal analysis of football videos. In: Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports. MMSports'21, pp. 17–24. Association for Computing Machinery, New York, NY, USA (2021)

[7] Díaz-Pereira, M.P., Gomez-Conde, I., Escalona, M., Olivieri, D.N.: Automatic recognition and scoring of olympic rhythmic gymnastic movements. Human movement science **34**, 63–80 (2014)

[8] Xu, C., Fu, Y., Zhang, B., Chen, Z., Jiang, Y.-G., Xue, X.: Learning to score figure skating sport videos. IEEE transactions on circuits and systems for video technology **30**(12), 4578–4590 (2019)

[9] Suzuki, T., Takeda, K., Fujii, K.: Automatic fault detection in race walking from a smartphone camera via fine-tuning pose estimation. In: 2022 IEEE 11th Global Conference on Consumer Electronics (GCCE), pp. 631–632 (2022). IEEE

[10] Suzuki, T., Takeda, K., Fujii, K.: Automatic detection of faults in race walking from a smartphone camera: a comparison of an olympic medalist and university athletes. arXiv preprint arXiv:2208.12646 (2022)

[11] Ivankovic, Z., Rackovic, M., Ivkovic, M.: Automatic player position detection in basketball games. Multimedia tools and applications **72**, 2741–2767 (2014)

[12] Napolean, Y., Wibowo, P.T., Gemert, J.C.: Running event visualization using videos from multiple cameras. In: Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports. MMSports '19, pp. 82–90. Association for Computing Machinery, New York, NY, USA (2019)

[13] Comandur, B.: Sports re-id: Improving re-identification of players in broadcast videos of team sports. arXiv preprint arXiv:2206.02373 (2022)

[14] Koshkina, M., Pidaparthy, H., Elder, J.H.: Contrastive learning for sports video: Unsupervised player classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4528–4536 (2021)

[15] Vats, K., McNally, W., Walters, P., Clausi, D.A., Zelek, J.S.: Ice hockey player identification via transformers and weakly supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3451–3460 (2022)

[16] Habel, K., Deuser, F., Oswald, N.: CLIP-ReIdent: Contrastive training for player Re-Identification. In: Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports. MMSports '22, pp. 129–135. Association for Computing Machinery, New York, NY, USA (2022)

[17] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.: Deep learning for person re-identification: A survey and outlook. IEEE transactions on pattern analysis and machine intelligence **44**(6), 2872–2893 (2021)

[18] Li, W., Liu, H., Ding, R., Liu, M., Wang, P., Yang, W.: Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Transactions on Multimedia **25**, 1282–1293 (2022)

[19] Uhlrich, S.D., Falisse, A., Kidziński, Ł., Muccini, J., Ko, M., Chaudhari, A.S., Hicks, J.L., Delp, S.L.: Opencap: 3d human movement dynamics from smartphone videos. BioRxiv, 2022–07 (2022)

[20] Jocher, G., Chaurasia, A., Qiu, J.: YOLO by Ultralytics (2023). https://github.com/ultralytics/ultralytics

[21] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR

[22] Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: International Conference on Machine Learning, pp. 10096–10106 (2021). PMLR

[23] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: European Conference on Computer Vision, pp. 1–21 (2022). Springer

[24] Hu, Z., Zhu, C., He, G.: Hard-sample guided hybrid contrast learning for unsupervised person re-identification. arXiv preprint arXiv:2109.12333 (2021)

[25] Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: Soccernet: A scalable dataset for action spotting in soccer videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1711–1721 (2018)

[26] Cioppa, A., Deliege, A., Giancola, S., Ghanem, B., Van Droogenbroeck, M.: Scaling up soccernet with multi-view spatial localization and re-identification. Scientific data **9**(1), 355 (2022)

[27] Zhu, H., Liang, J., Lin, C., Zhang, J., Hu, J.: A transformer-based system for action spotting in soccer videos. In: Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports, pp. 103–109 (2022)

[28] Vats, K., Fani, M., Walters, P., Clausi, D.A., Zelek, J.: Event detection in coarsely annotated sports videos via parallel multi-receptive field 1d convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 882–883 (2020)

[29] McNally, W., Vats, K., Pinto, T., Dulhanty, C., McPhee, J., Wong, A.: Golfdb: A video database for golf swing sequencing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019)

[30] Pidaparthy, H., Dowling, M.H., Elder, J.H.: Automatic play segmentation of hockey videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4585–4593 (2021)

[31] Wan, C., Wang, L., Phoha, V.V.: A survey on gait recognition. ACM Computing Surveys (CSUR) **51**(5), 1–35 (2018)

[32] Chao, H., Wang, K., He, Y., Zhang, J., Feng, J.: Gaitset: Cross-view gait recognition through utilizing gait as a deep set. IEEE transactions on pattern analysis and machine intelligence **44**(7), 3467–3478 (2021)

[33] Ge, Y., Zhu, F., Chen, D., Zhao, R., Li, H.: Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In: Advances in Neural Information Processing Systems (2020)

[34] Chen, H., Lagadec, B., Bremond, F.: Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14960–14969 (2021)

[35] Cho, Y., Kim, W.J., Hong, S., Yoon, S.-E.: Part-based pseudo label refinement for unsupervised person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7308–7318 (2022)

[36] Li, M., Zhu, X., Gong, S.: Unsupervised tracklet person Re-Identification. IEEE Trans. Pattern Anal. Mach. Intell. **42**(7), 1770–1782 (2020)

[37] Zheng, W.-S., Gong, S., Xiang, T.: Towards Open-World person Re-Identification by One-Shot Group-Based verification. IEEE Trans. Pattern Anal. Mach. Intell. **38**(3), 591–606 (2016)

[38] Zhu, X., Wu, B., Huang, D., Zheng, W.-S.: Fast Open-World person Re-Identification. IEEE Trans. Image Process. **27**(5), 2286–2300 (2018)

[39] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pp. 740–755 (2014). Springer

[40] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). Ieee

[41] Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: International Conference on Machine Learning, pp. 843–852 (2015). PMLR

[42] Penate-Sanchez, A., Freire-Obregon, D., Lorenzo-Melian, A., Lorenzo-Navarro, J., Castrillon-Santana, M.: Tgc20reid: A dataset for sport event re-identification in the wild. Pattern Recognition Letters **138**, 355–361 (2020)

[43] Wieczorek, M., Rychalska, B., Dąbrowski, J.: On the unreasonable effectiveness of centroids in image retrieval. In: Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part IV 28, pp. 212–223 (2021). Springer

[44] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1116–1124 (2015)

[45] Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 79–88 (2018)

[46] Liu, X., Liu, W., Mei, T., Ma, H.: A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14, pp. 869–884 (2016). Springer

[47] Qin, Z., Lu, X., Nie, X., Liu, D., Yin, Y., Wang, W.: Coarse-to-fine video instance segmentation with factorized conditional appearance flows. IEEE/CAA J. Autom. Sin. **10**(5), 1192–1208 (2023)

[48] Liang, J., Zhou, T., Liu, D., Wang, W.: CLUSTSEG: Clustering for universal segmentation (2023) arXiv:2305.02187 [cs.CV]

[49] Yan, L., Wang, Q., Ma, S., Wang, J., Yu, C.: Solve the puzzle of instance segmentation in videos: A weakly supervised framework with spatio-temporal collaboration. IEEE Transactions on Circuits and Systems for Video Technology **33**(1), 393–406 (2022)

[50] Qin, Z., Lu, X., Nie, X., Yin, Y., Shen, J.: Exposing the Self-Supervised Space-Time correspondence learning via graph kernels. AAAI **37**(2), 2110–2118 (2023)

[51] Qin, Z., Lu, X., Liu, D., Nie, X., Yin, Y., Shen, J., Loui, A.C.: Reformulating graph kernels for self-supervised Space-Time correspondence learning. IEEE Trans. Image Process. **PP** (2023)

[52] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)

[53] Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022)