Article

Directional Statistics-Based Deep Metric Learning for Pedestrian Tracking and Re-Identification

Abdelhamid Bouzid 1,*, Daniel Sierra-Sosa 2 and Adel Elmaghraby 1

1 Department of Computer Science and Engineering, University of Louisville, Louisville, KY 40208, USA
2 Department of Computer Science and Information Technology, Hood College, Frederick, MD 21701, USA
* Correspondence: abdelhamid.bouzid@louisville.edu

Abstract: Multiple Object Tracking (MOT) is the problem that involves following the trajectory of multiple objects in a sequence, generally a video. Pedestrians are among the most interesting subjects to track and recognize for many purposes such as surveillance, and safety. In recent years, Unmanned Aerial Vehicles (UAV’s) have been viewed as a viable option for monitoring public areas, as they provide a low-cost method of data collection while covering large and difficult-to-reach areas. In this paper, we present an online pedestrian tracking and re-identification framework based on learning a compact directional statistic distribution (von-Mises-Fisher distribution) for each person ID using a deep convolutional neural network. The distribution characteristics are trained to be invariant to clothes appearances and to transformations including rotation, translation, and background changes. Learning a vMF for each ID helps simultaneously in measuring the similarity between object instances and re-identifying the pedestrian’s ID. We experimentally validated our framework on standard publicly available dataset, which we used as a case study.

Keywords: pedestrians detection; tracking; re-identification; Unmanned Aerial Vehicles; drones; surveillance; von-Mises Fisher Distributions

1. Introduction

In artificial intelligence, Multiple Object Tracking (MOT) refers to the task of locating objects in a scene and maintaining their trajectories throughout the entire video. It is an essential task for a broad range of computer vision applications such as surveillance, and autonomous driving. Pedestrians are among the most interesting subjects to track in public area for many purposes such as safety and security. Therefore, lately, there has been a lot of research attention to pedestrian tracking [1,2].

In general, MOT follows the tracking-by-detection paradigm. The state of the art of this paradigm is divided into two main categories; online tracking [3–7], and batch tracking [8–11]. Online tracking is about associating object instances between a video past frame and current frame, whereas batch tracking uses the entire video frames to form a unique trajectory per ID.

The tracking association task requires a similarity measure between the object instances. This similarity is based on two types of features; motion properties and visual features. The visual features are extracted from the visual appearance of the object instance in the bounding box using a feature extractor. Thus, the feature extractor plays an important role in the model tracking performance. Recently, MOT proposed approaches [4,7,9] heavily relying on training a neural network base extractor. These approaches make use of advances in Deep Metric Learning (DML) [12–15] area in learning powerful visual representation. These base neural network extractors embed an input image into a sphere with a unit norm. The choice of the directional space is motivated by the assumption that the type of data is sometimes referred to as directional data, since the direction of the data provides more information than the magnitude of the data. However, the learning method for patch
representational learning is largely based on contrastive supervised learning in most MOT algorithms. These approaches rely on time-consuming mini-batch formats like triplets or N-pairs, which are incompatible with most MOT learning algorithm.

Machine learning-based person re-identification applications aim to successfully retrieve pedestrians by their identity for many security and safety tasks. Most of these applications aim to learn a mapping function that embeds images into compact euclidean space (unit sphere mostly) [16,17]. These embeddings should have the characteristic in which two images of the same person map to adjacent feature points, while images of two distinct persons map to feature points that are far apart. These mapping approaches should, ideally, be resistant to real-world situational changes like position, orientation, and occlusion in the same scene. It also should not rely on one’s clothes, since pedestrians wear different clothes at different lapses of time.

Most machine learning tracking and re-identification applications are heavily affected by image acquisition systems such as static cameras and the cost of collecting data. Unmanned aerial vehicles (UAV’s), which enable a low-cost way of data collection while covering broad and difficult-to-reach regions, have recently been recognized as a feasible alternative for monitoring public spaces [18–21]. The advancements in UAV’s have benefited MOT, particularly pedestrian tracking and re-identification, since it gives a viable solution to solve numerous challenges such as occlusion, moving cameras, and difficult-to-reach locations. Compared to static cameras, UAV’s are flexible enough to adapt their emplacement location and direction in the 3D space.

The re-identification has been applied in MOT for a long time. However, the contribution of this paper could be highlighted by discussing the limitations of the state-of-the-art re-identification MOT solutions. For example, although the existing methods (e.g., [22]) provide powerful re-identification MOT solutions, they are offline and cannot match the online processing in the drone application. In contrast, the proposed method is an online solution that is well-suited for drones.

Motivated by these facts, we adapted in this paper the approach proposed by [15] for image classification and retrieval to the pedestrian tracking and re-identification from aerial devices problem. This approach is a simple and efficient method for learning a von-Mises-Fisher (vMF) distribution for each ID in the directional space. For spherical data, this distribution can be considered a gaussian distribution. Learning a vMF for each ID helps simultaneously in measuring the similarity between object instances and identifying the person ID.

The following are the primary contributions of this paper:

- Introducing directional statistical distribution to pedestrian tracking and re-identification.
- A new object similarity measure based on pedestrian re-identification.
- An end-to-end multiple object tracking and re-identification of pedestrian from aerial devices.

The remainder of this paper is laid out as follows: We start by giving a background and an overview of related previous work in Section 2. In the latter a basic understanding of directional statistics is provided alongside a review of learning algorithm of vMF proposed by [15]. Our proposed online pedestrian tracking and re-identification is presented in Section 3. In Section 4, we present the dataset case study. Penultimately, we present our experimental results. Finally, we summarize our conclusions and outline potential future work.

2. Related Work

2.1. Pedestrian Tracking

In machine learning, pedestrian tracking refers to the task of detecting and tracking humans from video clips. It plays an important role in many safety systems such as in video surveillance and autonomous vehicle driving [1,2]. However, pedestrian tracking from a fixed camera suffers from several issues such as occlusion, access to area, and covering large areas. To overcome the aforementioned issues, unmanned aerial vehicles (UAV’s)
such as drones provide a cheap and effective way for collecting data. The UAV’s can record the scene from different angles and follow the flow of objects. In recent years, tracking from UAV’s has gained a lot of attention from researchers [18–21].

2.2. Tracking

The task of finding objects and retaining their identities over all video frames is known as Multiple Object Tracking (MOT). It is a vital problem for a broad range of computer vision applications such as surveillance, and autonomous driving. In recent years, MOT has been dominated by detection followed by the tracking paradigm, where first detections are obtained, then, instances of the same object are linked together. An object instance from frame \( t-1 \) is at maximum associated with a single object instance from frame \( t \). The state of the art of this paradigm can be divided into two main sub-categories:

2.2.1. Online Tracking Mode

Online tracking mode [3–7] is based on associating the detections between the past frame and the current frame. Usually, these methods are applied in real-time tracking scenarios tasks such as surveillance. Generally, this association problem is formulated as a bipartite graph problem [23]. The disjoint sets of vertices are the objects instances in frame \( t-1 \) and frame \( t \). The edges between the nodes represent the cost (similarity) between their corresponding instances.

2.2.2. Batch Tracking Mode

Methods that are based on batch tracking mode [8–11] use the entire video frame detections to build a unique track (trajectory) per ID. This is accomplished by associating past, present, and future directions. The association is generally formulated as a graph optimization problem such as maximum flow [24], or minimum cliques [25].

2.3. Deep Metric Learning

Many machine learning tasks require learning a measure of similarity between data objects such as MOT. The goal of metric learning is to learn a mapping function that quantifies the similarity between data points. The metric learning objective is to minimize the similarity between data points from the same category and maximize the distance between data points from different categories.

In recent years, deep learning has gained enormous success in a variety of machine learning tasks such as image classification, image embedding, and multiple object tracking. Deep learning brought revolutionary advances due to its representational power in extracting high abstract non-linear features. This fact led to a new research area known as Deep Metric Learning (DML) [12–15].

MOT has benefitted from the success of DML, by training a neural network feature extractor to learn a similarity measure between object instance patches.

2.4. Pedestrian Re-Identification

Pedestrian re-identification research has a broad range of aspects spanning from feature-based [26] to metric-based [27] and from hand-crafted features to deeply learned features [28,29]. In this report, we review three of the recent and relevant sub-research areas related to pedestrian re-identification problems.

Open-world person re-identification is defined as one-to-one set matches. Given two sets of pedestrian, the first called probe and the second called gallery, every person appears in both sets, and the task is to match between them [30]. In this problem, the pedestrian set must be known. However, in some open-scenarios it is not guaranteed that the probe will always have a corresponding match in the gallery which need to be catered to.

Generalized-view re-identification approached is mainly based on discriminative learning from different views acquired by two different fixed cameras [31,32]. However, in
practice this problem is costly since it requires data collected, annotated and matched from two different cameras.

Recently, pedestrian re-identification from drones gained a lot of attention and new benchmark datasets are published \[20,33\]. Drones provide a new tool for data acquisition, especially for video surveillance and analysis. With this new tool, problems such as pedestrian detection, tracking, and re-identification can be taken to new challenges as it helps overcome some of the static camera issues.

2.5. Directional Statistics in Machine Learning

Directional data is defined as points in a Euclidean space with norm \[\|x\|_2 = 1\], where \[\|\cdot\|_2\] is the Euclidean norm. In other words corresponding to points on the surface of the unit sphere. Thus, statistics that deal with directional data are called directional statistics.

The topic of directional statistics has gained a lot of attention due to demands from fields such as machine learning or the availability of big data sets that necessitate adaptive statistical methodologies, as well as technical improvements. Directional statistics method has led to a tremendous success in many computer vision tasks, such as image classification and retrieval \[15\], pose estimation \[34\] and Face Verification \[35\]. It has also been introduced to other machine learning fields such as text mining \[36\].

2.5.1. Von Mises-Fisher Distribution

Von Mises-Fisher Distribution (vMF) is a probability distribution function for directional data. It can be seen as a Gaussian distribution since they have very similar properties. In a directional data space \(S^{p-1}\), the probability distribution is density function defined as:

\[
f_p(x; \mu, \kappa) = Z_p(\kappa) \exp(\kappa \mu^T x),
\]

where, \(\mu\) is the mean direction of the distribution, \(\kappa \geq 0\) is a concentration parameter which can be seen as the standard deviation for Gaussian distribution, \(p\) is the space dimension, \(Z_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2}I_{p/2-1}(\kappa)}\) is a normalization term and \(I_v\) is the modified Bessel function of the first kind with order \(v\).

Given \(N\) samples from a vMF distribution, we can estimate its parameters as follows:

\[
\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{\| \sum_{i=1}^{N} x_i \|_2},
\]

and

\[
\hat{\kappa} = \frac{R(p - R^2)}{1 - R^2}.
\]

In Equation (3), \(R = \frac{\| \sum_{i=1}^{N} x_i \|}{N}\).

2.5.2. Learning Von-Mises Fisher Distribution

The learning problem is defined as follows. Given \(C\) identities, the goal is to learn a vMF distribution for every ID parameterized by \(\{\kappa_i, \mu_i\}\), where \(i = 1 \ldots C\).

Given a point \(x\) in the mapping space, the normalized probability of \(x\) belonging to a chosen class \(c\) is defined as:

\[
P(c|x, \{\kappa_i, \mu_i\}_{i=1}^{C}) = \frac{Z_p(\kappa_c) \exp(\kappa_c \mu_c^T x)}{\sum_{i=1}^{C} Z_p(\kappa_i) \exp(\kappa_i^T \mu_i x)}
\]
Equation (4) can be used to increase the likelihood that the sample belongs to the correct class while decreasing the likelihood that it belongs to other classes. Given a mini-batch with \( N \) samples and for a \( C \) identity, we can maximize the following objective function:

\[
P(Y|X, \Theta, \cup, \kappa) = \prod_{n=1}^{N} P(c|\{\kappa_i, \mu_i\}_{i=1}^{C})
\]

\[
= \prod_{n=1}^{N} \frac{Z_p(\kappa_c) \exp(\kappa_c \mu_c^T x)}{\sum_{i=1}^{C} Z_p(\kappa_i) \exp(\kappa_i \mu_i^T x)},
\]

where \( X \) and \( Y \) represent the data points in the mini-batch and their ID labels, \( \Theta \) contains the deep model parameters, and \( \cup = \{\mu_i\}_{i=1}^{C}, \kappa = \{\kappa_i\}_{i=1}^{C} \). For a simplification purpose, we assumed \( \kappa \) to be a constant for all IDs, and by applying the negative likelihood, Equation (6) can be simplified to:

\[
\arg \min_{\Theta, \cup} L = -\sum_{n=1}^{N} \log \left( \frac{\exp(\kappa \mu_c^T x)}{\sum_{i=1}^{C} \exp(\kappa \mu_i^T x)} \right)
\]

Algorithm 1 vMF learning algorithm

1. Initialize CNN parameters \( \Theta \).
2. \textbf{while} (Converge not achieved) \textbf{do}:
3. \hspace{0.5cm} Estimate mean directions using (8) and all the training data.
4. \hspace{0.5cm} Train CNN for several iterations and update \( \Theta \).

During the inference phase, we can predict the ID of a given object by measuring cosine similarity with the learned mean directions. The object will be assigned with the ID label of its nearest mean vector.

3. Dataset Description

Over the recent years, the pedestrian tracking and re-identification topic has gained a lot of research interest due to its importance and wide applicability, such as surveillance systems and traffic control. However, the acquisition system is limited by the stationary camera, which represents a main issue for the tracking and especially the pedestrian re-identification. Lately, Unmanned Aerial Vehicles (UAV’s) have been viewed as a viable option for monitoring public areas, as they provide a low-cost method of data collection while covering large and difficult-to-reach areas.

The P-DESTRE [33] dataset is publicly available and fully annotated dataset for detecting, tracking, re-identifying, and searching pedestrians from aerial devices. A group of researchers from the University of Beira Interior (Portugal) and the JSS Science and Technology University gathered this data (India). They recorded packed sights on both institutions using drones called “DJI Phantom 4”. These drones, as shown in Figure 1, are
piloted by humans to fly and collect data from a volunteer audience walking at altitudes ranging from 5.5 to 6.7 meters. In total, 75 videos with a Frame Per Second (fps) equal to 30 are collected. In these videos there are 318,745 annotated instances of 269 different IDs. These statistics are summarized in Table 1.

Table 1. P-DESTRE Dataset Statistics Summary.

|                          |             |
|--------------------------|-------------|
| Total number of videos   | 75          |
| Frames Per Second (fps)  | 30          |
| Total number of identities | 269        |
| Total number of annotated instances | 318,745 |
| Camera range distance    | [5.5–6.7] m |

The pedestrian search challenge, in which data is collected over long periods of time (e.g., days/weeks), with constant ID labels across observations, is the primary distinguishing characteristic between the P-DESTRE and comparable datasets. The re-identification techniques in this problem cannot rely on clothing appearance-based features, which is a key property that distinguishes search from the (less difficult) re-identification problem, in which the consecutive observations of each ID are assumed to have been taken in short intervals of time and clothing appearance features can be reliably used.

We believe that this dataset is an excellent case study for training and evaluating our frameworks for pedestrian tracking and re-identification from aerial devices. This dataset was used for all of the experiments and we plan to look at more aerial datasets in the future.

4. Methodology

4.1. Multiple Object Tracking and Re-Identification Framework

Our suggested approach is based on the generic MOT task paradigm of detection followed by tracking. We utilized YOLO V4 [37] as a detector. The proposed tracking method is then applied to track and recognize pedestrians. Figure 2 illustrates the structure diagram of this framework in the inference phase. It is comprised of 4 major steps:

1. We used YOLO V4 for real-time object detection. The YOLO V4 detector receives input videos frame by frame. As an output, the detection bounding boxes in each frame are obtained.
2. Patches are cropped from each frame using the bounding boxes that have been detected.
3. These patches are resized to the unified shape \((H, W, C)\), where \(H\) denotes height, \(W\) denotes width, and \(C\) denotes channel count. A feature vector is then generated for each detected object using the trained features extractor.

4. Using features vectors from frames \(t\) and \(t-1\), the object association and re-identification algorithm matches and recognizes the objects.

Figure 2. Multiple object tracking online framework. First, patches from frame \(t\) are extracted and resized to the same shape. Second, these patches are mapped through the trained feature extractor network to obtain the visual descriptors. Third, a pairwise similarity measure is conducted between the object descriptors of frame \(t\) and others of frame \(t-1\). Finally, an association algorithm is performed to match object instances based on their similarity.

Two neural network architectures, YOLO V4 and features extractor, must be trained for this framework. Using P-DESTRE dataset, we trained these two architectures independently, using the 10-fold learning/validation/test splits provided in the dataset web page. The data is randomly divided into 60% for learning, 20% for validation, and 20% for testing.

- **YOLO V4**: For each fold, we fine-tuned YOLO V4 pre-trained architecture to detect only pedestrian using the train set. The detector performance is monitored using the error on the validation set.

- **Feature extractor**: the feature extractor architecture is composed of a base model and a header. We used Wide ReseNet-50 (WRN) [38] as a base model, and header composed of two Fully connected Pooling Layers (FPL) of sizes \([4096, 128]\), respectively. From each fold a train, validation, and test patches dataset is created from train, validation, and test sets correspondingly. The model is trained using the vMF learning algorithm (Algorithm 1). Figure 3 illustrates the training mechanism of the features extractor.

During training, the detection and vMF are trained separately using the detection ground truth which are the bounding boxes. During inference, the detections are used to get the bounding boxes and then the earned vMF mean direction is used to track and re-identify the person.
4.2. Object Association and Re-identification

Learning a vMF distribution for each ID class has the ultimate objective of predicting the ID of an object instance during inference. These learned distributions, on the other hand, can be used to compare the similarity of any two object instances. As a result, learned distributions can be used to predict ID and assess similarity at the same time. We integrate the feature extractor into the online multiple object tracking and re-identification system once it has been trained in an offline way using the suggested framework. Current detections at frame $t$ are cropped and scaled to fit the feature extractor’s input form in this framework. Then, these patches are mapped to the embedding space. Since the system is online, it only retains the previous frame objects and their IDs at $t-1$. The pairwise similarity measure between the current object instances at frame $t$ and the objects instances at frame $t-1$ is the next stage in the system. The object detection association is the final phase.

4.2.1. Similarity Measure

Since the detection patches are mapped to a euclidean space with unit norm $\|x\|_2$, the similarity can be measured using cosine similarity or Euclidean distance. We opted to utilize one of them since the two metrics are inversely connected and provide the exact same results.

4.2.2. Object Detection Association

The final step in the online tracking and re-identification system is the object association and re-identification of the IDs. For object association, the current object detections in frame $t$ needed to be matched with the others in frame $t-1$. However, a pair of matched objects needs to belong to the same identity. In other words, we cannot match two object instances, while they are assigned to two different identities. Therefore, a consistency between the data association and the identity prediction needs to be established.

Prior to the start of the data association step, data preparation has to be completed. Let $D_t$ denotes the set of object detection in frame $t$, where $d_i^t$ denote the $i$th detected object. Let $\mu_c$ denotes the learned mean direction for ID $c$. The data association can be done in two different ways. First, we only rely on ID prediction by assigning the class, of which the mean direction is the nearest to the object representation:

$$\arg \max_c = \cos (d_i^t, \mu_c), c = 1 \ldots C,$$

where $\cos$ is the cosine similarity function. This technique is simple and straightforward, where no data association algorithm is needed. It can be seen as a recognition problem. Moreover, it is pertinent to mention that all subjects are in-distribution that is, they are present in the training dataset.

A second way to correlate targets from the previous frame $D_{t-1}$ with detections produced by the current frame $D_t$, we utilize the product result of appearance and position
criteria to indicate the amount of similarity of the target-detection pair of $i$th target and $j$th detection. This product is defined as follows:

$$\text{arg max}_j = \cos \left( d_i, d_{i-1} \right) + \frac{1}{2} \times \frac{\text{Max}_p - \| p_i - p_{i-1} \|_2}{\text{Max}_p - \text{Min}_p}$$

(10)

where $\| p_i - p_{i-1} \|_2$ is the Euclidean distance between the two object centroids in the pixel(image) space. $\text{Max}_j$ and $\text{Min}_j$ are the maximum and the minimum distances, respectively, between the two sets $D_i$ and $D_{i-1}$. Each of these criteria is normalized to $[0..1]$ range, so that we have a normalized score. This similarity is motivated by the fact that, the similarity between two object instances is not enough, it has to be reinforced by a position confidence score. We suppose that two detections belonging to the same ID have a similarity measure that is higher than a predetermined threshold $T$. The user defines this threshold, which is a hyper-parameter. We assign IDs to the unmatched subset of objects in frame $t$, which is denoted as $D^u_t$, using Equation (9).

5. Experiments

In this section, we report the performance results of the proposed framework. In order to provide a comparable results, we follow the same experimental protocol provided by the P-DESTRE web page http://p-destre.di.ubi.pt/experiments.html (accessed on 20 August 2022). The experiments were divided into three categories: (1) pedestrian; (2) detection; tracking (3) re-identification. The results of the state-of-the-art are directly reported from the original paper.

5.1. Pedestrian Detection

For the YOLO V4 detector, we set the Non-Max-Suppression (NMS) hyper-parameter intersection over union between the predictions to 0.5 and the threshold score for pre-filtering to 0.05. The maximum number of positive detections (predicted pedestrian per image) is set at 100. During the inference phase, the score threshold is set to 0.3. To train the detector, 300 epochs with a learning rate of 0.0001 were sufficient. In order to align with the reported results from the original paper, we used the same development kit http://host.robots.ox.ac.uk/pascal/VOC/voc2012/#devkit (accessed on 20 August 2022) to evaluate YOLO V4 detection performance in P-DESTRE dataset. Following the 10-fold cross validation scheme provided http://p-destre.di.ubi.pt/pedestrian_detection_splits.zip (accessed on 20 August 2022), in which each fold the data was randomly divided into 60% for learning, 20% for validation, and 20% for testing.

The results of all detection methods are summarized in Table 2. The approaches were assessed using the Average Precision (AP) measure with an Intersection over Union (IoU) of 0.5 ($\text{AP}_{@\text{IoU}=0.5}$). As can be observed, the YOLO V4 outperformed the other detection methods considerably which motivated to use it in overall pipeline. It’s also built for online real-time detection, making it simple to combine with other online tracking and re-identification systems as a detector. Overall, we noticed that the YOLO V4 struggled in crowded situations with occlusions and only small portion of the area was viewable. The other approaches in the P-DESTRE paper were also challenged by this problem.

Table 2. The Average Precision (AP) results obtained by 4 detection methods in the P-DESTRE dataset [33]. RetinaNet, R-FCN, SSD taken from [33].

| Method       | Backbone   | AP          |
|--------------|------------|-------------|
| RetinaNet [39] | ResNet-50  | 63.10 % ± 1.64 % |
| R-FCN [40]   | ResNet-101 | 59.29 % ± 1.31 % |
| SSD [41]     | Inception-V2 | 55.63 % ± 2.93 % |
| YOLO V4 [37] | CSPResNext50 | 65.70 % ± 2.40 % |
5.2. Pedestrian Tracking

We trained the feature extractor, using the 10-fold cross validation scheme of the P-DESTRE set. In this scheme, each split randomly divided into 60% for learning, 20% for validation and 20% for test, i.e., 45 videos were used for learning, 15 for validation and 15 videos for test purposes. Only the subset of identities that are available in the training videos are used for learning the vMF. From each frame, the bounding boxes are cropped and scaled to patches of dimension \((48, 64, 3)\). The input patch is normalized to the \([0, 1]\) range by dividing by 255 for the prepossessing. We set the concentration parameter \(\kappa\) to 15 for the learning algorithm hyper-parameter. This number produced the best outcomes experimentally. The feature extractor was trained using 50 epochs with a batch size of 64. 128 is the embedding space dimension. The similarity threshold is set to 0.2.

As detailed in [42], we assessed the performance of our proposed tracking system using three metrics: Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), and F1 score. Table 3 summarizes the performance results. As shown below, our proposed tracking method slightly outperformed the other tracking methods in terms of MOTA, but provides comparable results in terms of MOTP to TracktorCV. In terms of F - 1 score, our method is significantly better.

From a qualitative standpoint, our approaches struggle with severe occlusions. Normally, it is difficult to reestablish the person ID in online tracking after occlusion happens, but our method is able to restore the person ID in several cases due to re-identification based on vMF mean direction (9).

Table 3. Comparison between 3 tracking algorithms from the state-of-the-art and our tracking algorithm in the P-DESTRE dataset [33]. TracktorCv, V-IOU, IOU taken from [33].

| Method               | MOTA    | MOTP    | F-1       |
|----------------------|---------|---------|-----------|
| TracktorCv [43]      | 56.00% ± 3.70% | 55.90% ± 2.60% | 87.40% ± 2.00% |
| V-IOU [44]           | 47.90% ± 5.10% | 51.10% ± 5.80% | 83.30% ± 8.40% |
| IOU [45]             | 38.27% ± 8.42% | 39.68% ± 4.92% | 74.29% ± 6.87% |
| Ours (based on vMF)  | 59.30% ± 3.50% | 56.10% ± 5.61% | 92.10% ± 3.20% |

We designed an experiment to display the tracking results in inference mode on a chosen video as a proof-of-concept for our tracking plus detection approach. We picked one of the 5-folds at random, we trained the algorithm, and then applied it to a particular video from the test set in inference mode. The results of the detection and tracking at three distinct timestamps are shown in Figure 4. The tracking algorithm was successful in keeping track of the majority of the pedestrians in the scenario. Our investigation revealed that the majority of switched identities occur as a result of miss-detection at particular frames, which results in two cases: (1) assign new ID to the pedestrian when it is detected again; and (2) two pedestrians switch their identities.

Figure 4. Example of applying our detection and tracking in inference mode on one of the videos (named “11-11-2019-1-2_out.mp4”). The selected video is part of the test data. We show the results at three different timestamps. The tracking algorithm was able to keep track of most of the identities in the scene.
5.3. Long-Term Pedestrian Re-Identification

Long-term pedestrian re-identification is one of the main reasons for the development of tracking and re-identification based on vMF distribution. In a long-term real world scenario pedestrian clothing appears to differ between lapses of time spanning over days and weeks. Therefore, a reliable identifier must depend on persistent features such as face or body rather than clothing.

In this experiment, the data was randomly divided into 5-folds. The ratios of these folds are 50% learning, 10% gallery, and 40% query. The detailed information about this split is provided in http://p-destre.di.ubi.pt/pedestrian_search_splits.zip (accessed on 20 August 2022).

Our approach to re-identify pedestrians can be assessed in two ways. First, we utilize Equation (9) to get the nearest mean direction of the IDs. Second, top-N recall performance is used. We report using the same metrics as the state-of-the-art results in order to be consistent. We also present the results of the re-identification using the initial measure we presented in (9). The results are summarized in Table 4. Our method significantly outperformed other methods using different metrics. This demonstrates that a feature extractor based on vMF is able to learn reliable features that help in recognizing the person, rather than the person’s clothing appearances.

Table 4. Comparison between the re-identification performance attained by the state-of-the-art methods and ours based on vMF on the P-DESTRE dataset [33]. ArcFace + COSAM taken from [33].

| Method                  | mAP   | Rank-1          | Rank-20         | Mean Direction |
|-------------------------|-------|-----------------|-----------------|----------------|
| ArcFace [46] + COSAM [47]| 34.90%±6.43% | 49.88%±8.01%    | 70.10%±11.25%   | —              |
| vMF identifier          | 37.85%±3.42% | 53.81%±4.50%    | 74.61%±8.50%    | 64.45%±3.90%   |

5.4. Analysis

As proof of concept that the proposed framework can learn a vMF distribution for each pedestrian identity, we designed an experiment to visually represent the learnt embeddings. The experiment findings are illustrated in Figure 5. We randomly selected 7 pedestrian identities, each identify is associated with a color in the figure. We applied vMF learning algorithm on their patch dataset. For the purpose of visualization, we set the embedding dimensions to 3. We also set the concentration $\kappa = 15$. The results demonstrate that the algorithm was successful in learning vMF distributions for each identity, where patches belonging to the same identity are grouped together into a single cluster and set apart from embeddings belonging to other identities. The experiments also showed that high concentration does actually help in moving data points with identical identities in the direction of their mean, but it makes the model learn vMF distribution that are close to each other. This means that with high concentration values the learning is more dominated by the intra-variations, variations between different appearance and looking of the same pedestrians in different scenes, rather than with the inter-variations, variations between different pedestrian identities.
Figure 5. The learned embedding of 7 randomly selected pedestrian identities, where each color is associated with one single identity. For the purpose of visualization we set the embedding dimensions to 3. We also set the concentration $\kappa = 15$.

6. Conclusions

An online (real-time) pedestrian tracking and re-identification approach based on images acquired from aerial devices (drones) is proposed and evaluated. The foundation of the proposed framework is a neural network encoder that was trained to learn the vMF distribution for each ID in directional space. The learned distribution aided in tracking and recognizing pedestrians at the same time. The P-DESTRE dataset case study was utilized to assess our proposed system’s tracking and re-identification performance using standard metrics. We demonstrated that we can outperform other approaches in re-identifying pedestrians with more efficient feature retention. We plan to expand this approach and apply it to additional types of datasets in the future. Another aspect that could be looked into is online learning. Our proposed approach has been tested for other applications and can be adopted for different object tracking such as vehicles, military equipment, and parcels.

It is worth noting that these approaches are computationally intensive for training and comparison purposes. However, with computational advances deploying these models for tracking and re-identification is becoming feasible.

Author Contributions: All authors contributed equally to the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used is publicly available at http://p-destre.di.ubi.pt/ (accessed on 20 August 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Suhr, J.K.; Jung, H.G. Rearview camera-based backover warning system exploiting a combination of pose-specific pedestrian recognitions. IEEE Trans. Intell. Transp. Syst. 2017, 19, 1122–1129. [CrossRef]
2. Camara, F.; Bellotto, N.; Cosar, S.; Nathanael, D.; Althoff, M.; Wu, J.; Ruenz, J.; Dietrich, A.; Fox, C. Pedestrian models for autonomous driving part I: low-level models, from sensing to tracking. IEEE Trans. Intell. Transp. Syst. 2020, 22, 6131–6151. [CrossRef]
3. Han, S.; Huang, P.; Wang, H.; Yu, E.; Liu, D.; Pan, X.; Zhao, J. Mat: Motion-aware multi-object tracking. arXiv 2020, arXiv:2009.04794.
32. Koestinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P.M.; Bischof, H. Large scale metric learning from equivalence constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2288–2295.

33. Kumar, S.A.; Yaghoubi, E.; Das, A.; Harish, B.; Proença, H. The p-destre: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. IEEE Trans. Inf. Forensics Secur. 2020, 16, 1696–1708. [CrossRef]

34. Prokudin, S.; Geihsler, P.; Nowozin, S. Deep directional statistics: Pose estimation with uncertainty quantification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 534–551.

35. Hasnat, M.; Bohné, J.; Milgram, J.; Gentric, S.; Chen, L. Von mises-fisher mixture model-based deep learning: Application to face verification. arXiv 2017, arXiv:1706.04264.

36. Straub, J.; Chang, J.; Freifeld, O.; Fisher, J., III. A Dirichlet process mixture model for spherical data. In Proceedings of the Artificial Intelligence and Statistics, San Diego, CA, USA, 9–12 May 2015; pp. 930–938.

37. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. arXiv 2020, arXiv:2004.10934.

38. Zagoruyko, S.; Komodakis, N. Wide residual networks. arXiv 2016, arXiv:1605.07146.

39. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.

40. Dai, K.; R-FCN, Y. Object detection via region-based fully convolutional networks. arXiv 2016, arXiv:1605.06409.

41. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Computer Vision–ECCV 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.

42. Bernardin, K.; Stiefelhagen, R. Evaluating multiple object tracking performance: The clear mot metrics. EURASIP J. Image Video Process. 2008, 2008, 246309. [CrossRef]

43. Bergmann, P.; Meinhardt, T.; Leal-Taixe, L. Tracking without bells and whistles. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 941–951.

44. Bochinski, E.; Senst, T.; Sikora, T. Extending IOU based multi-object tracking by visual information. In Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 27–30 November 2018; pp. 1–6.

45. Bochinski, E.; Eiselein, V.; Sikora, T. High-speed tracking-by-detection without using image information. In Proceedings of the 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; pp. 1–6.

46. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 4690–4699.

47. Subramaniam, A.; Namibiar, A.; Mittal, A. Co-segmentation inspired attention networks for video-based person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 562–572.