Comparison of re-identification methods based on AlignedReID++

Óscar Jesús Díaz de la Fe
SIANI
ULPGC
oscar.diaz108@alu.ulpgc.es

Abstract

This article aims to conduct various experiments using an algorithm designed for person identification, which will be fed through a complex dataset, thus determining the current state of the art. The study considers the numerous external factors that influence the difficulty of these processes. For this project, a re-identification framework known as AlignedReID++ will be utilized. This framework proposes its own method called Dynamic Local Information Matching (DLIM), which combines global and local functions based on it, allowing for higher accuracy in the inference phase. The TGC20ReId dataset will be employed, acquired in a sports context, consisting of 606 participants and their images obtained from five different registration points, thus containing significant variations among each point. The obtained results will determine the effectiveness of ZSL, tl, and TS methods, opening the possibility of utilizing them as tactics to prevent cheating in sports events. In summary, this project involves the exploration and experimentation of advanced person re-identification techniques in the sports domain, with the aim of improving the accuracy and reliability of these processes in the present times.
.

1. Introduction

In recent years, sporting events have experienced exponential growth, leading to an increase in the number of participants, as well as the complexity of activities and rewards offered by them. This surge in the popularity of sporting events has been evident in many places around the world, and Spain is no exception

In Spain, a wide variety of sporting events can be found, ranging from bicycle races to marathons and cross-country races. One of the most popular events in Spain is the Vuelta Ciclista a España, which has been held annually since 1935 [1]. Since then, the race has grown in popularity and complexity, reaching its current status where the 77th edition featured the participation of 168 cyclists divided into 21 teams, with 21 stages [2] [5] and a total of €1,112,640 in prize money distributed throughout the race [7].

Federations strive to ensure that participants do not engage in any form of cheating, whether through the use of doping substances, tampered equipment, substitution of participants during the race, or even gaining aerodynamic advantages from external factors. One measure taken to address this is to ensure that participants pass through various checkpoints during the races.

The purpose of this study is to investigate the design of a system capable of detecting cheating practices related to bib theft or circumventing parts of the course by going off-track. For this, the chosen architecture will be AlignedID++ to review the following objectives:

- To test these methods, a complex dataset with a wide range of variables will be employed.
- Analyzing the obtained results will help determine the robustness and effectiveness of the method in uncontrolled environments, shedding light on how far we may be from effectively implementing these techniques.

2. Related work

Person re-identification involves identifying and tracking an individual across multiple non necessarily overlapping cameras. It is a highly interesting field of research due to its wide range of applications, ranging from security measures to surveillance systems. However, the process of manually processing a large amount of video data (or images) is highly inefficient. As a result, various proposals have been designed over the years to automate this task.

While several methods have shown promising results with existing datasets, there is still much work to be done to address the complexity of the challenge. The significant variations that can occur in the data, such as changes in pose, lighting, image quality, occlusion, and more, pose a significant obstacle. Therefore, these designs are primarily developed using datasets in controlled situations, and their effectiveness may vary when compared to experiments

conducted in real-world environments. This discrepancy is commonly referred to as the difference between closed-set and open-set experiments[1].

In general terms, a series of steps can be defined when developing a re-identification system, which are as follows:

- **Data collection:** This involves obtaining data from one or multiple cameras. The raw data must be preprocessed before it can be used effectively.
- **Bounding box generation:** This step is crucial as it involves creating bounding boxes that contain the images of individuals from the data. This can be done manually for each image or through detection algorithms, which is more practical when working with a large number of images.
- **Data annotation:** This step involves annotating labels for the images and cameras, which is essential for training a re-identification model.
- **Model training:** This step entails training a robust re-identification model using the annotated images or videos of individuals. Multiple models and techniques have been developed based on feature representation learning, distance metric learning, or a combination of both.

Additionally, deep learning is employed, which is a branch of machine learning that focuses on training artificial neural networks to learn and perform complex tasks such as pattern recognition, image classification, and natural language processing. These artificial neural networks are designed to mimic the behavior of neurons in the human brain, enabling them to autonomously learn from large datasets and improve their performance as more information is provided.

Deep learning has had a significant impact in areas such as computer vision, speech recognition, machine translation, and natural language processing. This is because deep neural networks can effectively learn and detect patterns in input data, surpassing the capabilities of traditional machine learning algorithms, leading to higher accuracy in predictions and classifications.

Deep learning [4] relies on models of deep neural networks, which consist of multiple interconnected layers of artificial neurons. Each layer processes the input information differently, allowing for the detection of patterns at different levels of abstraction. As information flows through the network, each layer refines and processes the input to generate a final output that represents a prediction or classification.

3. Methods and datasets

3.1. AlignedReID++

The state of the art in re-identification is constantly evolving, with numerous approaches and algorithms developed to improve the accuracy and effectiveness of the systems. Among the various available options, AlignedReID++ [9] has been chosen as the foundation for conducting tests and experiments in this project, as it has shown promising results compared to other existing methods. Furthermore, AlignedReID++ has been specifically designed to address the challenges of person re-identification, such as pose variation, lighting, and background.

It is an evolution of the AlignedReID [15] architecture. It consists of a method called 'Dynamically Matching Local Information' (DMLI), which aligns the images in horizontal stripes without the need for external supervision, thereby addressing the problem of pose variation for target individuals, bounding boxes, etc. The method is applicable to most convolutional networks based on re-identification.

The architecture of AlignedReID++ is based on the design of the ResNet50 architecture for feature map extraction (C x H x W, where 'C' is the number of channels and 'H x W' is the spatial size), with the output of the last convolutional layer. Additionally, global mean pooling is applied to convert the feature map into a global feature vector (C x 1), assuming that f_A and f_B are the global features of images A and B, allowing the global distance to be described as:

$$d_g(A, B) = ||f_A - f_B||_2 \tag{1}$$

This is because global distance can measure the similarity between two images more easily, albeit at the cost of ignoring local information. To remedy this, DMLI is used, where the local branch employs max (or average) pooling to convert the feature map (C x H x W) into a local feature map (C x H x 1), which is then resized to H x C. This allows defining l_A and l_B as the local features of the respective images. Furthermore, a distance normalization between 0 and 1 is performed using element-wise transformation, where $d_{i,j}$ represents the distance between the i-th vertical part of the first image and the j-th vertical part of the second image.

$$d_{i,j} = \frac{e^{||_A l^i - l_B^j||_2} - 1}{e^{||l_A^i - l_B^j||_2} + 1} \quad i, j \in 1, 2, 3, ..., H \tag{2}$$

With these distances $d_{i,j}$, the distance matrix D can be generated. In other words, the local distance between the images is defined as the total distance of the shortest path from (1,1) to (H,H) in the matrix D. These distances can be calculated using the equation:

$$S_{i,j} = \begin{cases} d_{i,j} & i = 1, j = 1 \\ S_{i-1,j} + d_{i,j} & i \neq 1, j = 1 \\ S_{i,j_1} + d_{i,j} & i = 1, j \neq 1 \\ \min(S_{i-1,j}, S_{i,j-1}) + d_{i,j} & i \neq 1, j \neq 1 \end{cases} \quad (3)$$

Where $S_{i,j}$ is the total shortest distance from (1,1) to (i,j) in the matrix D, and $S_{H,H}$ is the total shortest distance at the end, in this case, the local distance between the two images: $d_l(A, B) = S_{H,H}$. In the inference stage, the total distance between two images is:

$$d(A, B) = d_g(A, B) + \lambda d_l(A, B) \quad (4)$$

Where λ is the weight to balance the global and local distance.

However, local features may capture some differences. Therefore, in the AlignedReID++ network, global features are learned together with the local features based on DMLI.

The backbone network and the global branch are the same as in most previous re-identification works, trained with softmax loss (L_{ID}), combined with TriHard loss $(L_{ID} + L_T^g)$, where the latter adds a batch normalization layer and ReLU activation before horizontal grouping during training. Additionally, a 1x1 convolutional layer is used to reduce the number of channels in the feature map to facilitate training.

AlignedReID++ utilizes global distances to extract samples due to their faster computation, and optimizing global and local branches with different triplets does not provide any help to the network. Let's consider L_T^l as the triplet loss [13] of the local branch, where the triplet loss is based on the context of nearest neighbor classification. The goal is to ensure that an anchor image x_i^a of a specific individual is closer to all positive images x_i^p of the same person than to any other negative image x_i^n of any other person.

This leads to the following equations:

$$||f(x_i^a) - f(x_i^p)||_2^2 + \alpha < ||f(x_i^a) - f(x_i^n)||_2^2, \quad (5)$$

$$\forall (f(x_i^a), f(x_i^p), f(x_i^n)) \in T, \quad (6)$$

where α is a margin applied between positive and negative pairs. T represents the set of all possible triplets in the training set with a cardinality of N.

To minimize the loss, we arrive at the equation:

$$L = \sum_i^N [||f(x_i^a) - f(x_i^p)||_2^2 - ||f(x_i^a) - f(x_i^n)||_2^2 + \alpha]_+. \quad (7)$$

This allows for many possible triplets that satisfy the above equations. Although these do not contribute to training or speed up convergence, they still play a role in the network.

Considering the above, the total loss of AlignedReID++ is determined as:

$$L = L_{ID} + L_T^g + L_T^l, \quad (8)$$

The backpropagation of AlignedReID++ is easily calculated through automatic differentiation in the deep learning environment. By using equations 1 and 2 and denoting $x = ||l_A^i - l_B^j||2$, where l_A^i and l_B^j are corresponding local features, we consider that d_{ij} satisfies two conditions. The first condition, $d_{ij}\alpha x$, does not change the monotonicity of x in the range between 0 and 1, allowing for small changes to align the local features and make the model more stable. The second condition, $\frac{\partial \mathrm{d}_{ij}}{\partial x}\frac{1}{\alpha}x$, is calculated using the following formula:

$$\frac{\partial \mathrm{d}_{ij}}{\partial x} = \frac{2}{e^x + \frac{1}{e^x} + 2}, \quad e^x \in [1, e^{max(x)}]. \quad (9)$$

Here, $\frac{\partial \mathrm{d}_{ij}}{\partial x}\frac{1}{\alpha}x$ holds since e^x is not less than 1. Non-corresponding alignment has a large L2 distance, and its gradient is close to 0 in the above equation. Hence, the contribution of such alignments in the shortest path is small. Thus, the local distance between images is primarily determined by the corresponding alignments.

3.2. Datasets

Re-identification datasets are collections of images and associated metadata used to train and evaluate person re-identification algorithms. These datasets are crucial in re-identification research and development as they provide a foundation for model training and evaluation.

Re-identification datasets come in varying sizes and complexities, and AlignedReID++ has been evaluated to work with some of the well-known datasets, including MSMT17 [14], CUHK03 [8], DukeMTMCReID [12], and Market1501 [16].

- **Market1501:**

  Market1501 [16] is a dataset that contains images captured by six cameras positioned in front of a supermarket. These cameras include five high-resolution HD cameras (1280×1080) and one standard-resolution SD camera (720×576), with overlapping coverage. The dataset includes 32,668 bounding boxes corresponding to 1,501 identities. Due to the open environment, each identity in the dataset is captured by a maximum of six cameras to enable cross-camera search, ensuring that each annotated identity is captured by at least two cameras. This dataset exhibits the following characteristics:

- Market-1501 differs from other datasets in that it utilizes the Deformable Parts Model (DPM) [6] for bounding box detection instead of manually cropped bounding boxes.
- Within this dataset, a classification is applied to the detected bounding boxes based on their area ratio. If the area ratio exceeds 50%, the detected bounding box is considered "good". On the other hand, if the area ratio is less than 20%, it is labeled as "distractor". Any bounding box that does not meet these conditions is classified as "junk", indicating that the image has no significant influence on the re-identification accuracy. Additionally, certain bounding boxes that are evident false alarms in the dataset are identified and marked as "distractors".
- Each identity can have multiple images in each camera, allowing for multiple queries and analyses for each identity during cross-camera search. This better reflects real-world scenarios where multiple queries can be utilized to gather more discriminative information about the person of interest.

Dataset	Market-1501
# identity	1501
# bounding boxes	32668
# distractors	2,793 + 500K
# camaras	6

Table 1. Dates Market-1501 [17]

- **TGC20ReID:**

  TGC20ReID [10] dataset contains images of participants participating in the Transgrancanaria (TGC) 2020 Classic, a 128KM race that spans from the north to the south of Gran Canaria Island. In the 2020 race, 435 out of 677 participants successfully completed the race within the time limit. A significant subset of these participants was recorded at five checkpoint locations (RP1, RP2, RP3, RP4, RP5) along the course. Two of these checkpoints (RP1 and RP2) occurred during the nighttime, while the rest were during the daytime. Due to the 30-hour duration of the race, only 109 participants were captured at all checkpoints.

  Unlike other sports datasets, TGC20ReID captures the same runner in highly diverse conditions due to the race's duration and the characteristics of the checkpoints. Another notable difference compared to existing datasets is the variation in lighting conditions from night to day, including changes in lighting within the checkpoints due to shadows caused by buildings, mountains, etc. Additionally, the appearance of the participants is not consistent throughout the race due to changes in clothing. All these factors contribute to making the dataset a highly challenging collection of data for testing different re-identification algorithms.

Dataset	TGC20ReId
#Sports	Trail race
Annotation	Identies
Media	Imagenes
#Sequences	4373
Conditions	Day & Nightlight
#Identities	416 2 5 locations
Quality	1920x1080

Table 2. Dates TGC20ReId [10]

4. Experiments

Three experiments were conducted to evaluate different methods: zero-shot learning [11] (ZSL), training from scratch (TS), and transfer learning [3] (TL). For each experiment, optimal weights were obtained for each training method. The next step involved computing the embeddings of the images of the participants that appeared at all checkpoints. This process involves representing these images with numerical vectors, which has significantly improved knowledge discovery and content recommendation tasks. With these vectors, we can determine the proximity matrix of the images and calculate a Cumulative Matching Curve (CMC), which is a commonly used evaluation metric in the field of Computer Vision. The CMC curve represents the probability that a model correctly matches a target person within a ranked set of retrieved and classified images, based on the rank. It provides a comprehensive evaluation of the model's performance. Specifically, we will examine ranks 1, 5, 10, and 20.

4.1. Zero-shot learning

ZSL [11] is a machine learning technique that allows a model to classify objects from unseen classes, without specific training on those particular classes. This is achieved through the model's ability to understand and utilize descriptive information about the classes, such as attributes or associated features.

The goal of zero-shot learning is to expand the classification capability of models, enabling them to recognize and classify new and unknown objects without the need for specific training on each class.

The first method involved applying this technique to the TGC20ReID dataset, using pre-trained weights from Market1501.

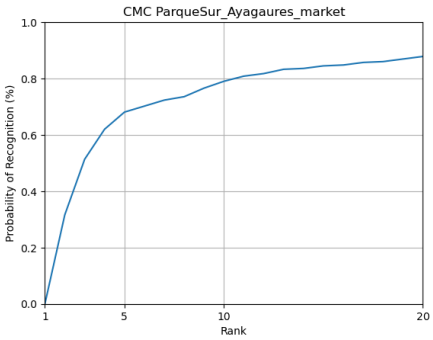

Figure 1. CMC zero-shot learning

4.2. training from scratch

The second method involves training AlignedReID++ solely on the TGC20ReID dataset. Since the dataset only contains a set of participants that appear in all image capture points, it is important to ensure that the network can correctly identify any new runner that may be added in the future. To achieve this, all the participants appearing in all image capture points are placed in the test folder, while the remaining images are used for training the network. It is worth mentioning that these training images appear in multiple capture points but not in all of them. The results will be evaluated using the AlignedReID++ model trained from scratch. This will allow us to assess the performance when working exclusively with our dataset.

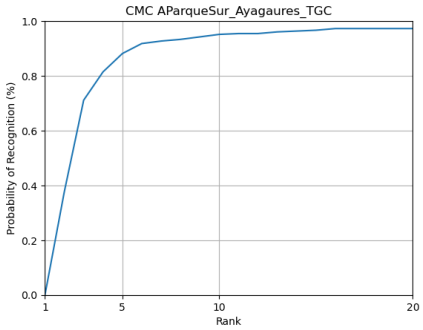

Figure 2. CMC training from scratch

4.3. Transfer learning

TL [3] is a technique that involves using a pre-trained model on a specific task and adapting it for another related task instead of training a model from scratch. By utilizing a model that has already been trained on a large and complex dataset, we can then fine-tune and retrain the model on a smaller and more specific dataset.

In this experiment, we leverage the pre-trained weights from the Market1501 dataset. By employing this strategy, we aim to achieve a significant improvement in results by utilizing the previously learned features from Market1501 as a starting point in training the model on the TGCReID dataset. By doing so, we expect the model to capture more complex and representative patterns, enhancing its ability to recognize and classify individuals more accurately and efficiently.

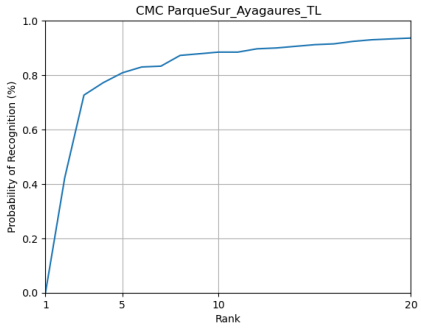

Figure 3. CMC Transfer learning

5. Results

In table displays the results obtained for the previous experiments: zero-shot learning, TS, and transfer learning, for the different tested ranks. Additionally, a comparative CMC graph is provided to visualize the variations between the experiments(figure 4).

Methods	mAP	Rank-1	Rank-5	Rank-10	Rank-20
ZSL	45.2%	41.9%	66.0%	79.0%	90.0%
TS	79.3%	77.8%	93.9%	97.3%	98.2%
TL	76.0%	76.6%	90.3%	93.6%	97.3%

Table 3. Result of experiments

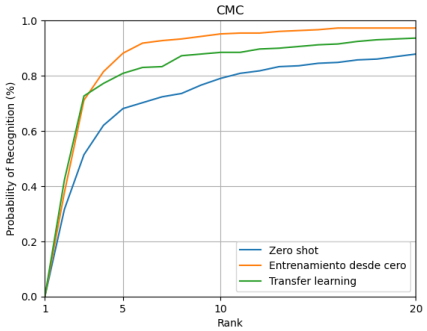

Figure 4. CMC comparison

The results obtained from training the TGCReId dataset from TS have shown significantly better performance com-

pared to the other experiments conducted using the same dataset. One possible reason for this is the increased complexity of our dataset, as observed in the ZSL experiment, where directly evaluating with the mentioned weights yielded the worst results. This may be due to the lack of necessary features required for better performance, a limitation that was not observed in the TS and TL experiments, where more discriminative features were detected, resulting in improved identification of the participants.

6. Conclution

After analyzing the results, it is evident that ZSL yielded the poorest performance with a Rank-1 accuracy of 41.9

It is important to note that while we have achieved significant improvement through the use of TL, notable improvements have also been observed with training directly on the dataset of interest. This could be attributed to the fact that the dataset used in our evaluations was not extensive enough to provide a complete representation of underlying patterns and features. Therefore, it is crucial to continue researching and expanding the dataset to further enhance our results.

In conclusion, this study has demonstrated the potential and effectiveness of TL and training directly on the dataset of interest in person re-identification. However, there is still room for improvement in these techniques, particularly through dataset expansion and exploration of additional approaches. These ongoing efforts will continue to drive significant advancements in improving the accuracy and robustness of person re-identification systems across various scenarios and applications.

References

[1] Bici Home. **https://bicihome.com/historia-de-la-vuelta-a-espana-desde-sus-inicios-hasta-hoy/**, 19 de Enero de 2023.

[2] Ciclo21. **https://www.ciclo21.com/vuelta-espana-2022/**, 19 de Enero de 2023.

[3] DataScientest. **https://datascientest.com/es/que-es-el-transfer-learning**, 19 de Enero de 2023.

[4] S. Dong, P. Wang, and K. Abbas. A survey on deep learning and its applications. *Computer Science Review*, 40:100379, 2021.

[5] eitb.eus. **https://www.eitb.eus/es/deportes/ciclismo/detalle/8934218/listado-de-ciclistas-participantes-de-vuelta-a-espana-2022-nombre-de-corredeores-y-sus-dorsales/**, 19 de Enero de 2023.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[7] Las Provincias. **https://www.lasprovincias.es/deportes/ciclismo/vuelta-espana/premios-vuelta-ciclista-espana-20220816173706-nt.html**, 19 de Enero de 2023.

[8] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[9] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, and C. Zhang. Alignedreid++: Dynamically matching local information for person re-identification. *Pattern Recognition*, 94:53–61, 2019. Cited By :123.

[10] A. Penate-Sanchez, D. Freire-Obregón, A. Lorenzo-Melián, J. Lorenzo-Navarro, and M. Castrillón-Santana. Tgc20reid: A dataset for sport event re-identification in the wild. *Pattern Recognition Letters*, 138:355–361, 2020. Cited By :4.

[11] Petru Potrimba. **https://blog.roboflow.com/zero-shot-learning-computer-vision/**, 2023.

[12] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking, 2016.

[13] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[14] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification, 2018.

[15] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun. Alignedreid: Surpassing human-level performance in person re-identification, 2017.

[16] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[17] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015.