Person Re-Identification Using Convolutional Neural Network and Autoencoder Embedded on Frameworks Based on Siamese and Triplet Networks

Fábia Isabella Pires Enembreck (✉ enembreck@alunos.utfpr.edu.br)

Universidade Tecnológica Federal do Parana <https://orcid.org/0000-0001-7364-0912>

Erikson Freitas de Moraes

Universidade Tecnológica Federal do Parana

Marcella Scoczynski Ribeiro Martins

Universidade Tecnológica Federal do Parana

Research

Keywords: person re-identification, deep learning, surveillance, siamese neural network, triplet neural network

Posted Date: December 8th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-117671/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

Person re-identification using Convolutional Neural Network and Autoencoder embedded on frameworks based on Siamese and Triplet networks

Fábia Isabella Pires Enembreck^{1*}, Erikson Freitas de Moraes¹ and Marcella Scoczynski Ribeiro Martins²

*Correspondence:

enembreck@alunos.utfpr.edu.br

¹Department of Informatics,
Federal University of Technology
of Parana, Ponta Grossa, BR
Full list of author information is
available at the end of the article

Abstract

The person re-identification problem addresses the task of identify if a person being watched by security cameras in surveillance environments has ever been in the scene. This problem is considered challenging, since the images obtained by cameras are subject to many variations, such as lighting, perspective and occlusions. This work aims to develop two robust approaches based on deep learning techniques for person re-identification, considering these variations. The first approach uses a Siamese neural network composed by two identical subnets. This model receives two input images that may or may not be from the same person. The second approach consists of a triplet neural network, with three identical subnets, which receives a reference image from a certain person, a second image from the same person and another image from a different person. Both approaches have identical subnets, composed by a convolutional neural network which extracts general characteristics from each image and an autoencoder model, responsible for addressing high variations that input images may undergo. To compare the developed networks, three datasets were used, and the accuracy and the CMC curve metrics were applied for the analysis. The experiments showed an improvement in the results with the use of the autoencoder in the subnets. Besides, Triplet Neural Network presented promising results in comparison with Siamese Neural Network and state-of-the-art methods.

Keywords: person re-identification; deep learning; surveillance; siamese neural network; triplet neural network

1 Introduction

An intelligent system for security that uses video cameras has not only the function of environment surveillance but also the detection of events, people and the scene interpretation. One advantage by the use of these system is related to easy and automatic way for handling a large amount of data and information, instead an human operator monitoring the video, avoiding wrong registering and operator fatigue. There are many problems in intelligent security systems such as anomalous event, people or object detection, tracking and person re-identification, among others [1].

This work covers the person re-identification problem, which consists of identifying a person that have already been presented in images from the same camera or in images obtained from different cameras. This problem should consider several

characteristics and appearance changes from an analyzed person that may occur as he/she move among the cameras, caused by environmental and geometric variations and/or by partial occlusions [2].

In general, the person re-identification problem can be defined as a process in which is intended to establish a correspondence between different images of the same person. Therefore a person who has been previously seen may be identified again [3].

There are several techniques for person re-identification in the literature. However, this problem does not have a definitive solution. The analyzed images can present numerous variations of lighting, points of view, the person's position, partial occlusions, low resolutions, among other problems. These problems make the person re-identification a complex problem, since the same person can look different from one image to another, as well as different people can look very similar [4].

The main purpose in this work is to develop two approaches for person re-identification in digital images using deep learning techniques. The first approach is a Siamese Neural Network and the second a Triplet Neural Network, both built using a subnet composed by a convolutional neural network (CNN) and an autoencoder (AE) model, which performs the re-identification of people who were previously detected. The Siamese Neural Network and the Triplet Network are composed by two and three identical subnets, respectively.

The proposed methods considers two input images and compares them to verify whether they are from the same person. This is possible because the subnets generate a feature vector of images, extracted by CNN. This feature vector is reconstructed by AE, which minimizes the noise that can compromise the comparison with the other image, meanwhile maintains more relevant features for re-identification. Details of how the training of each implemented network is performed are described in section 3. The methods are validated using the accuracy and CMC curve. Experiments were also carried out to compare our neural networks against state-of-the-art methods available in the literature.

2 Related Work

In this section it will be presented other works that deal with the problem of person re-identification in digital images, using features descriptors and deep learning.

Gray and Tao developed a method called ELF for the person re-identification problem. The technique is based on the combination of histograms of RGB, YCbCr, and HSV1 color channels with image texture. A model of the most significant features of each image is learned by the AdaBoost algorithm during training. It was observed that the most significant features were the hue and saturation, due to the variations of luminosity between the images of two cameras. This method does not solve the problem of automatic person re-identification, however, by means of the similarity function it assists a human operator in the recognition of pedestrians in images of different cameras, reducing about 82% [5].

Another work that also considers the most representative histogram and texture features of each image was developed by Prosser [6]. In this method, a vector containing the features of each image is obtained and during training and associated with a set of relevant features and a vector with irrelevant features, through positive

and negative correspondences. Each image was divided into 6 rectangles, to represent body parts. For the classification was used the ranking method RankSVM, which assigns the positive matches scores between the vectors. Better results were obtained than that proposed by Gray and Tao [5]. This method can become very costly computationally since it can generate a large number of negative samples.

In the method proposed by Zhao [7], are considered features extracted from salient regions, described as regions that make a person different from the other and that must be present in the images even with variations in the point of view. This information is extracted by LAB color histogram and SIFT descriptor of each 10x10 pixels in the image. The correspondence between these pixels is determined on the basis of the Euclidean Distance, using the k-nearest neighbor algorithm (KNN). The method employed may aid in the person re-identification with variations in point of view, however, in some cases the overhang regions can be treated as outliers by the algorithm.

Wu [8] developed a method for extracting features of the images, combining a convolutional neural network with a modification of the method proposed by Gray and Tao [5]. The method consists of a convolutional neural network formed by two parts, the first one containing convolution and pooling layers to extract features of an image. The second step uses a modification of the ELF [5] method that divides the input image into 16 bands for extraction of the features to be performed in each band, thus producing a 16-dimensional histogram for each color channel, which will be concatenated at the end to form a single vector.

Li [9] proposed a method called DeepReID: Deep Filter Pairing Neural Network, which consists of a network that automatically learns features for identifying people. The network takes into account possible transformations that may occur in images and background interference. A set of images with 13,164 images of 1,360 different pedestrians obtained by 6 cameras, until then the largest set of data for re-identification of people in the literature, was also created for the training of the method.

Ahmed [2] used a Siamese neural network to learn similarity metrics between two images. The network is composed of two subnets containing two convolution layers used to obtain the image features maps. To verify the matching between feature maps, the network calculates the neighborhood differences between them. This operation will result in 25 neighborhood difference maps pass through a CNN to learn spatial relationships between maps.

A Siamese neural network was used by Yi [10] to find similarity in two images. The proposed network has two convolutional neural networks, formed by convolution layers, followed by a normalization layer, max-pooling, and a fully connected layer. For the calculation of the similarity of the images was used the cosine function. Yi's method obtained better results than the methods developed in [5] and [7]. However, the subnets do not share the same parameters, making a comparison between the features vectors imprecise.

A Siamese neural network composed of a convolutional neural network and a recurrent neural network to person re-identification based on video was projected in [11]. The convolutional neural network produces a features map of each frame and sends it as input to the recurrent neural network. The information from each

frame is maintained and processed by the recurring neural network as new features are obtained by the network. A temporal pooling layer is used to combine all these characteristics. The similarity function used in this method is the Euclidean Distance.

Wang [12] proposed a method for re-identifying people using unsupervised learning. This was possible due to the sharing of the knowledge of the information of an individual of origin through attributes learned from the labeled source data, without then transferring that knowledge to destination data not labeled by joint learning of identity transfer between domains. For this, each figure of the individual goes through two extractors of characteristics, the first of which aims to extract sensitive information for re-identification and the second of which aims to extract the semantic knowledge of the labels of the attributes. After that, a knowledge fusion channel is used to integrate the information obtained by the two channels. In this work, Wang used an autoencoder in the fusion channel, with the justification that he has a great capacity to capture the most important information from input and also because it is a more concise representation of features, facilitating the transfer of information between tasks.

3 Method

This section presents the proposed approaches to person re-identification in digital images obtained by different points of view and cameras. As cited in Section 1, this paper introduces two methods for person re-identification using Siamese Neural Network and Triplet Network. The first implemented method is formed by two identical subnets that are joined by a Contrastive Loss Function. The second method is presented consisting of three identical subnets joined by a Triplet Loss Function.

In both methods, the subnets are the same, formed by a convolutional neural network and an autoencoder. The Figure 1 shows the subnet model proposed in this work. Each subnet receives an input image, which goes through a process of extracting features and subsequently verifying similarity between them, based on the features found. Each of these subnet steps are described in Section 3.1 and Section 3.2.

3.1 Convolutional Neural Network - CNN

The first stage of the subnet consists of a CNN with 4 layers of convolution, responsible for producing a feature map of the input image.

In general, a Convolutional Neural Network (CNN) uses convolution operations on at least one of its layers to learn patterns from a given set of data. For images as input, a convolution layer has formed a set of filters and produces as output a features map of the image. A convolution operation can be denoted in Equation 1

$$s(x) = (f * w) = \sum_{a=-\infty}^{\infty} f(a)w(t-a) \quad (1)$$

where the first argument, represented by x , is the input and w is the kernel represents a filter to be applied in the image [13, 14].

In addition, the CNN proposed uses a layer of max-pooling and *dropout*. The pooling layer is used to reduce the size of the feature map produced by the convolution layer, making the number of parameters in the network smaller and decreasing the computational cost, since the width and height in the characteristics map also decreases. In the proposed subnet the max-pooling operation is used for a region: 2x2 of the features map replaces the highest value of this region in the output [13, 14].

Dropout is a technique developed to avoid overfitting during the training of a neural network. The method consists of randomly discarding data, preventing the network from adapting too much, breaking patterns that are not significant for the learning process. For example, if a layer of a neural network returns, for a given sample of training input, the vector (0.2; 0.5; 1.3; 0.8; 1, 1), applying Dropout the new vector could be (0; 0.5; 1.3; 0; 1.1) [15, 16].

3.2 Autoencoder - AE

A neural network autoencoder (AE) uses unsupervised learning to learn patterns through a set of data. The architecture of an AE consists of at least 3 layers: input, output and hidden layer, and the input and output are the same size.

The purpose of an AE is to reproduce the input data at its output. This is done by means of an encoder function, which extracts characteristics from the input and produces a compressed features vector. The decompression of this vector is performed by a decoder function that translates the data as close as possible to the input, minimizing the reconstruction error. In this way, the network ends up learning to prioritize the most important properties of the input data [13].

As used by Wang [12], the autoencoder can learn the most important information for a given input representation, generating a characteristic vector with more concise data. This work takes into account that each image has several types of variations, such as lighting and perspective since they were not obtained by the same camera and pedestrian movement.

After being reconstructed, the vector goes through a normalization layer, which aims to make the samples more similar, helping the network in learning and better generalization for new data. This can be accomplished through changes in the data values for a common scale, without distorting differences in the value ranges and without losing information.

Each feature vector produced by one subnet is compared with the vector produced by the other subnet of the neural network. To make an association between them, for instance, to check if these vectors correspond to the image of the same person, two architectures are used in this work: Neural Siamese Network and Neural Triplet Network, described in the next sections.

3.3 Siamese Neural Network

A Siamese network consists of two identical subnets that are joined together in their outputs. Each subnet receives a different input that is mapped to a feature descriptor. Two descriptors are obtained and compared to estimate the similarity between them, resulting in the output of the network. Subnets need to share the same parameters and weights to make the output of each of them comparable [17].

In Figure 2 the proposed Siamese Neural Network model is presented. The pair of input images must be associated with a binary label Y that identifies whether the images are from the same person or not. For example, if the images are from the same person, the network receives $Y = 0$ as the label, otherwise, the label should be $Y = 1$. The network also receives for training a variable, where if 0, so the images correspond to the same person or 1 if they are from different people. The network entry contains two images and binary value. The network will adjust its weights during training to reach a certain value, according to the pairs of input images, using a Contrastive Loss which seeks to separate with greater distance entries belonging to different people using a predefined margin value.

During training, the network modifies its parameters to find a smaller difference between the inputs, according to the label given to them. For this, a function to calculate this difference is used (E_w), given by Equation 2. This equation passed to a Constraint Loss function L [17].

$$E_w = |G_w(X_1) - G_w(X_2)| \quad (2)$$

where X_1 and X_2 are a input pair of images, and G_w is a set of network functions.

The Contrastive Loss Function (Equation 3) is used to measure the network ability to find similarities between images. The function learns the parameters therefore the most similar examples get closer and the different ones are separated,

$$L(w, Y, X_1, X_2) = (1 - Y) \frac{1}{2} (E_w)^2 + Y \frac{1}{2} \{ \max(0, m - E_w) \}^2 \quad (3)$$

where m is the margin, if $m > 0$, such that (X_1, X_2) are a positive input pair of images, and (X_1, X'_2) are a negative input pair of images, so $E_w(X_1, X_2) + m < E_w(X_1, X'_2)$ [18].

3.4 Triplet Network

A Triplet Network is inspired by a Siamese Network, but this network is composed of three subnets with shared parameters where each subnet is fed with a sample. Thereby, a triplet samples feed the network input: anchor x_i , positive x_i^+ and negative x_i^- . The sample called anchor comprises the reference sample, so we have that the sample denoted positive must be from the same class as the anchor and the negative from a different class [19]. In the context of person re-identification, each different person represents a class, therefore the images representing the anchor sample and the positive sample are different images of the same person and the negative sample is another person, as can be observed in the Figure 3.

The purpose of a triplet network is to ensure that an anchor sample from one person is closer to all other positive samples from the same person than any negative sample from anyone else, reducing the distance to the anchor and positive sample as much as possible (Figure 4). For this, triplet loss is used, which is based on the context of the classification of the nearest neighbor. The grid generates two intermediate values of distance from the input samples, this encodes the pair of distances between each positive and negative sample relative to the anchor [19, 20].

Being the embedded representation of the network represented by $f(x)$, we have to

$$\|f(x_i) - f(x_i^+)\|_2^2 + \alpha < \|f(x_i) - f(x_i^-)\|_2^2, \quad (4)$$

$$\forall (f(x_i), f(x_i^+), f(x_i^-)) \in \tau \quad (5)$$

Where τ is the set of all possible triplets in the training set and α is the margin that is enforced between positive and negative pairs [20].

4 Experiments and Results

This section presents the results obtained with the training and tests of the proposed Siamese and Triplet Neural Networks. The networks were implemented in Python, through the library for neural networks Keras^[1], operating on Tensorflow^[2]. Three public datasets were used for network training and testing: VIPeR [21], i-LIDSVID [22] and CUHK03 [9]. The evaluation metrics used to analyze the performance of the implemented models were Accuracy and the Cumulative Correspondence Characteristic (CMC) curve.

Accuracy is one of the most used metrics in machine learning to measure the performance of a classifier. Being obtained by the fraction of correct predictions p of a given model about the total of forecasts T_p , it can be calculated by $A = \frac{p}{T}$ [23].

The CMC curve is considered a classification metric based on the concept of scores. Performance is measured for each sample, determining a set of correspondence candidates, the idea being that these candidates belong to samples with the same identity as the sample evaluated [24].

Bolle [25] described two subsets of samples to measure a CMC curve for a primary sample set B_i . The first being a set of biometric identifiers $G = \{B_1, B_2, \dots, B_m\}$ for m different identities. The second is a set of n samples $Q = \{B'_l, l = 1, \dots, n\}$ that can be from any individual that in set G , being that it can contain more than one sample for this identity. Thus, m is the total number of identities and n is the total number of samples per identity.

Considering the total number of samples given by $N = n * m$, each sample is compared with the remaining samples $(N - 1)$ resulting in a set of S scores of two types: 1) genuine scores: both samples belonging to the same individual; 2) impostor scores: samples belonging to different individuals [25],[24].

When comparing a set Q with the set of identities G , a set of scores $S'_l = \{s(B'_l, B_1), s(B'_l, B_2), \dots, s(B'_l, B_m)\}$ is generated. Thus, for each example B'_l , the scores s are ordered from highest to lowest, generating a list C . The CMC curve must represent the probability that a genuine score is in the first k classifications of C [25],[24].

^[1] Available in <https://keras.io>

^[2] Available in <https://www.tensorflow.org>

4.1 VIPeR dataset

The ViewPoint Invariant Pedestrian Recognition (VIPeR) dataset was developed to simulate real-world safety environments, where pedestrians appear in large open environments and can be viewed from any angle. It contains a total of 1264 images of 632 pedestrians, and for each pedestrian there are two images captured by different cameras, containing variation between points of view and also changes in the lighting and poses of individuals [21]. In the Figure 5 there are samples of 10 different pedestrian image pairs.

To increase the number of training images and reduce overfitting, the data augmentation technique was performed on the images of this dataset. There were 11 transformations in the indicated original image obtaining a total of 15144 images.

The networks were trained with 100, 200, 400 and 600 epochs. Table 1 shows the accuracy results obtained for each season in the set of images intended for testing and Figure 6 presents the graph.

Table 1 Accuracy (%) of each network model in relation to the number of training epochs using the VIPeR dataset

Epoch	Siamese	Triplet	Siamese	Triplet
			Without AE	Without AE
100	51.33%	64.27%	35.48%	44.25%
200	54.38%	75.15%	39.87%	53.76%
400	69.5%	89.86%	46.54%	58.35%
600	81.62%	94.35%	58.49%	66.65%
800	89.54%	96.29%	62.32%	71.04%

To validate the use of AE in the subnets, experiments were carried out with and without AE. As can be seen in the Table 1 and Figure 6, the tests without AE had the worst accuracy rates, compared to tests with AE, in all cases.

The Triplet Network with AE started the experiment with better accuracy than the Siamese Network, with 51.33% and 64.27% respectively. As the number of epochs increased, the accuracy of the Siamese Neural Network increased, but remained ever smaller, ending the experiment with 89.54% accuracy. The Triplet Network achieved its best result in this experiment, raising 96.29% with 800 seasons.

Figure 6 Experiments on VIPeR: Accuracy (%) of each network model in relation to the number of training epochs.

Figure 7 shows the graph of the CMC curve for the four networks implemented and tested with the VIPeR dataset. It can be seen that the Triplet Network with AE obtained the best results in all cases. Table 2 shows the hit rates for each rank.

Table 2 CMC curves on VIPeR dataset

Epoch	Siamese	Triplet	Siamese Without AE	Triplet Without AE
1	28.98%	47.36%	20.52%	23.61%
2	51.52%	58.91%	24.14%	25.45%
3	52.34%	61.85%	27.52%	33.32%
4	52.89%	66.87%	29.41%	36.43%
5	67.58%	72.72%	31.45%	41.32%
6	75.84%	81.32%	36.65%	48.21%
7	80.62%	86.58%	39.54%	50.95%
8	83.80%	90.90%	44.32%	55.95%
9	86.16%	92.65%	49.51%	56.83%
10	87.91%	92.65%	53.65%	59.98%
11	89.34%	93.43%	56.43%	61.43%
12	90.50%	95.69%	59.25%	64.12%
13	91.41%	97.12%	61.04%	67.54%
14	92.17%	97.12%	62.64%	68.90%
15	92.79%	99.04%	64.23%	71.78%

4.2 i-LIDSVID dataset

The i-LIDS Video Re-Identification (i-LIDSVID) dataset contains images of 300 different pedestrians obtained through two non-overlapping cameras. There are two sets of images for each individual: one of still images and the other of sequential images. In the case of sequential images, the images were obtained by pedestrian tracking, containing from 23 to 192 frames for each person other than dataset. Images have many variations between different cameras, such as lighting, point of view variations, and occlusions [22].

Figure 7 Experiments on ViPER: CMC curves on ViPER dataset

Figure 8 iLIDS-VID: Images of the same pedestrian, obtained through two non-overlapping cameras [22].

Figure 8 presents an example of sequential images of the same pedestrian, however, captured by different cameras. There will not always be frame-by-frame parity, as one person may not appear in both camera views at the same time.

The accuracy of the test set for different training epochs of the implemented networks can be seen in Table 3 and in the graph in Figure 9. Networks that do not have AE in the subnets obtained lower accuracy rates in all cases, compared to networks with AE.

Analyzing networks with AE, the Triplet Network got the highest hit percentages with 97.48% for 800 epochs. The networks chart has remained more consistent and kept growing with an increasing number of epochs. In this case, the best result to Siamese Neural Network was 92.48% for 800 epochs. The CMC curve of the experiments performed with this dataset can be seen in Figure 10, with its data in Table 4.

4.3 CUHK03 dataset

The CUHK03 dataset was developed taking into account that better training of deep neural networks requires the use of a large number of images. Thus, the CUHK03 dataset was built with 13164 images of 1360 different people, obtained through six

Table 3 Accuracy (%) of each network model in relation to the number of training epochs using the i-LIDSVID dataset

Epoch	Siamese	Triplet	Siamese Without AE	Triplet Without AE
100	42.06%	67.04%	24.55%	36.32%
200	46.42%	81.58%	31.47%	42.36%
400	67.88%	92.36%	39.54%	49.84%
600	86.13%	94.47%	43.89%	52.47%
800	92.85%	97.48%	54.28%	59.82%

Figure 9 Experiments on i-LIDSVID: Accuracy (%) of each network model in relation to the number of training epochs.**Table 4** CMC curves on i-LIDSVID dataset

Epoch	Siamese	Triplet	Siamese Without AE	Triplet Without AE
1	31.45%	55.55%	15.34%	24.76%
2	33.69%	64.32%	19.12%	27.43%
3	38.21%	73.65%	24.54%	30.65%
4	39.03%	76.76%	29.32%	31.96%
5	42.90%	79.32%	31.76%	34.21%
6	46.32%	81.21%	32.88%	35.85%
7	51.87%	87.87%	34.56%	38.52%
8	55.62%	87.87%	37.24%	42.05%
9	67.21%	91.94%	41.87%	45.58%
10	75.45%	94.95%	44.09%	47.65%
11	88.10%	96.86%	46.33%	50.95%
12	94.33%	97.41%	48.47%	56.49%
13	96.25%	97.97%	50.03%	59.33%
14	97.77%	98.32%	51.78%	61.48%
15	98.59%	98.99%	54.43%	62.92%

surveillance cameras. Examples of images from this dataset can be seen in Figure 11. Images of manually cropped people are also available and also detected using a pedestrian detector in addition to the original frames [9].

This dataset presents several problems such as misalignment, occlusions, missing body parts, among others. Also, cameras monitor an open area, subjecting to variations in lighting caused by various weather factors in a single camera view. Other transformations may also occur due to the various directions pedestrians walk [9].

Figure 10 Experiments on i-LIDSVID: CMC curves on i-LIDSVID dataset

Figure 11 CUHK03: Samples of pedestrian images observed in different camera views, two adjacent images have the same identity. [9]

Table 5 presents the results obtained with the experiments using the four implement network models and the Figure 12 illustrates a graph of the accuracy results, in relation to the number of training epochs.

Table 5 Accuracy (%) of each network model in relation to the number of training epochs using the CUHK03 dataset

Epoch	Siamese	Triplet	Siamese Without AE	Triplet Without AE
100	44.76%	51.91%	21.84%	39.87
200	49.74%	74.85%	32.95%	46.28
400	66.88%	88.64%	43.28%	51.21
600	90.51%	91.26%	54.95%	55.64
800	94.09%	97.62%	65.52%	68.21

Figure 12 Experiments on CUHK03: Accuracy (%) of each network model in relation to the number of training epochs.

As can be seen from the graph in the Figure 12, In these experiments the networks that were implemented and the AE also obtained better results. In this case, Triplet Network obtained the best accuracy of this dataset, with 97.62% to 800 epochs. In both networks the accuracy increased as the number of epochs increased. The best result to Siamese Neural Network was 94.09% to 800 epochs.

Figure 13 and Table 6 show the results for the CMC curve. As in the experiments carried out with the other datasets, the Triplet Network with AE obtained the best hit rates in all cases, compared with the Siamese Neural Network with AE for ranks from 1 to 15 and with the neural networks without AE.

Table 6 CMC curves on CUHK03 dataset

Epoch	Siamese	Triplet	Siamese Without AE	Triplet Without AE
1	24.06%	47.50%	15.43%	19.35%
2	28.94%	63.98%	18.67%	22.09%
3	31.89%	71.35%	21.65%	25.63%
4	35.98%	75.63%	23.65%	28.52%
5	41.22%	81.32%	28.52%	31.12%
6	44.54%	84.85%	30.75%	32.94%
7	51.58%	86.00%	32.56%	34.45%
8	58.36%	90.57%	34.32%	35.34%
9	66.97%	92.50%	36.98%	40.56%
10	72.41%	93.76%	39.34%	43.21%
11	77.65%	95.58%	41.37%	47.65%
12	83.47%	98.43%	44.22%	52.34%
13	88.21%	99.25%	48.34%	55.32%
14	92.48%	100.00%	51.27%	63.76%
15	94.35%	100.00%	52.92%	66.83%

4.4 Discussion of Results

Given the considered experiments, the use of AE provided better results for both implemented approaches. For the results with the dataset i-LIDSVID, the networks

Figure 13 Experiments on CUHK03: CMC curves on CUHK03 dataset

obtained the highest rates, increasing accuracy in all cases, with 71.05% and 62.95% for Siamese Neural Network and Neural Triplet Network, respectively.

The experiments carried out with the networks that use AE, it is concluded that Neural Network Triplet obtained better results in all cases, compared to the Neural Siamese Network, using the same datasets and subnets. The accuracy of each case can be seen in Figure 14.

Figure 14 Accuracy (%) of each network model in relation to the number of training epochs.

As the Neural Network Triplet receives a positive and a negative image pair per entry, the network may have had an easier time, separating images with different

identities better and also bringing together images with identical identities, compared to the neural network that receives only one positive or one negative image pair per entry.

The generated CMC curves for networks with AE were also compared and the Neural Triplet Network also obtained better results against the results of the Siamese Neural Network. From the analysis of Figure 15, we can highlight the CMC Curve of the Siamese Neural Network using the i-LIDSVID dataset obtained identification rates very close to the rates obtained by the other network. This may be due to the fact that this dataset is a multi-shot type, presenting more images with the same identity.

The CMC curves generated by the experiments with the CUHK03 dataset showed that the results started with lower identification rates than the other datasets for the same rank. One explanation for this is that this dataset has a much larger number of images and identities than the other two. Throughout the experiments, the CMC Curve of the Neural Triplet Network starting from rank-12, obtained the highest identification rates, and in the last two ranks, the network reached 100% correct with the CUHK03. Total candidates, in this case, had about 30 images for each identity.

4.5 Comparison with State of the Art

This section presents a comparison of the CMC Curves obtained for each dataset and proposed method with the AE with some methods available in the state of the art.

The methods used to compare the performance of the networks using the VIPeR dataset were: ELF [5], KISSME [26], SDALF [27], DGD [28], NullSpace [29] and the method developed by Ahmed [2]. As can be seen in Table 7 and in Figure 16, our Neural Triplet Network, obtained the best performance in rank-2 and from rank-6

it was the better between methods. In other cases, our network was among the top three methods.

Table 7 Comparison of CMC Curves with state of the art methods, using the dataset VIPeR

Rank	Siamese	Triplet	ELF	KISSME	SDALF	DGD	NullSpace	Ahmed
1	28.98%	47.36%	18.62%	21.25%	20.26%	33.87%	51.33%	49.22%
2	51.52%	58.91%	22.44%	27.85%	24.52%	37.74%	56.42%	51.26%
3	52.34%	61.85%	25.84%	32.91%	29.78%	45.45%	63.49%	54.46%
4	52.89%	66.87%	28.43%	38.61%	37.59%	58.64%	66.84%	63.72%
5	67.58%	72.72%	29.61%	43.67%	41.23%	60.41%	74.55%	74.73%
6	75.84%	81.32%	31.29%	47.47%	44.41%	61.58%	78.69%	77.54%
7	80.62%	86.58%	33.49%	50.95%	46.27%	62.08%	81.65%	80.65%
8	83.80%	90.90%	35.73%	54.11%	49.47%	65.21%	84.71%	82.45%
9	86.16%	92.65%	39.12%	56.33%	50.23%	66.79%	86.28%	84.78%
10	87.91%	92.65%	40.67%	57.91%	50.96%	68.46%	89.67%	85.48%
11	89.34%	93.43%	42.46%	59.18%	55.43%	69.07%	91.43%	86.04%
12	90.50%	95.69%	46.28%	62.03%	56.19%	69.45%	92.98%	86.48%
13	91.41%	97.12%	47.66%	62.97%	58.49%	70.96%	94.52%	87.58%
14	92.17%	97.12%	48.27%	65.19%	60.05%	72.64%	95.67%	89.21%
15	92.79%	99.04%	50.03%	66.14%	62.17%	74.56%	97.52%	90.15%

Figure 16 Comparison of CMC Curves with state of the art methods, using the dataset VIPeR

The i-LIDSVID dataset was used in experiments to compare the following methods available in the state of the art: KISSME [26], SDALF [27], DGD [28] and Ahmed [2]. In experiments with i-LIDSVID, the Triplet Neural Network had the best results from rank-2. In rank-1 this network was in second place and the DGD method obtained the best result. The Siamese Neural Network was in second place in all cases starting from rank-11. The results are shown in Table 8 e Figure 17.

For the CUHK03 dataset, the approaches developed in this work were compared with the following methods: KISSME [26], SDALF [27], DGD [28], NullSpace [29], NormXcorr [30] and Ahmed [2]. The results obtained can be seen in Table 9 and Figure 18. The DGD method obtained a better identification rate in the first ten ranks. Starting from rank-11, our Neural Network Triplet obtained the best results,

Table 8 Comparison of CMC Curves with state of the art methods, using the dataset i-LIDSVID

Rank	Siamese	Triplet	KISSME	SDALF	DGD	Ahmed
1	31.45%	55.55%	13.23%	10.39%	58.43%	38.76%
2	33.69%	64.32%	17.44%	12.87%	59.61%	47.13%
3	38.21%	73.65%	22.92%	15.48%	60.66%	56.98%
4	39.03%	76.76%	28.28%	18.76%	61.98%	62.24%
5	42.90%	79.32%	30.42%	20.14%	63.58%	66.32%
6	46.32%	81.21%	33.74%	22.35%	64.52%	68.45%
7	51.87%	87.87%	35.76%	24.56%	68.81%	71.22%
8	55.62%	87.87%	39.43%	25.87%	72.44%	75.21%
9	67.21%	91.94%	40.15%	26.93%	75.27%	77.42%
10	75.45%	94.95%	42.32%	28.52%	81.03%	78.23%
11	88.10%	96.86%	44.63%	29.05%	86.32%	79.46%
12	94.33%	97.41%	45.91%	31.96%	89.35%	81.09%
13	96.25%	97.97%	48.33%	32.80%	91.86%	84.03%
14	97.77%	98.32%	50.90%	33.53%	94.33%	87.66%
15	98.59%	98.99%	53.21%	35.47%	97.69%	89.23%

reaching a rate of 100 % in rank-14 and rank-15. It is worth mentioning from rank-6 that our network was among the two best methods from rank-6.

Table 9 Comparison of CMC Curves with state of the art methods, using the dataset CUHK03

Rank	Siamese	Triplet	KISSME	SDALF	DGD	NullSpace	Norm Xcorr	Ahmed
1	24.06%	47.50%	12.12%	12.52%	72.41%	54.12%	69.08%	44.21%
2	28.94%	63.98%	21.45%	20.46%	75.92%	66.54%	72.45%	52.69%
3	31.89%	71.35%	33.53%	25.59%	79.25%	73.42%	75.87%	63.29%
4	35.98%	75.63%	37.65%	28.07%	85.62%	79.65%	78.09%	68.56%
5	41.22%	81.32%	39.26%	29.41%	87.23%	82.63%	79.78%	72.65%
6	44.54%	84.85%	41.54%	31.04%	88.67%	83.98%	81.73%	73.78%
7	51.58%	86.00%	44.62%	32.65%	91.42%	85.32%	84.22%	74.87%
8	58.36%	90.57%	45.87%	35.41%	93.17%	89.52%	87.94%	77.12%
9	66.97%	92.50%	47.21%	36.44%	94.89%	90.63%	88.78%	79.77%
10	72.41%	93.76%	49.67%	37.19%	95.13%	93.54%	90.06%	83.63%
11	77.65%	95.58%	51.36%	42.58%	95.51%	94.09%	95.54%	85.11%
12	83.47%	98.43%	55.43%	46.41%	96.32%	94.64%	95.58%	86.56%
13	88.21%	99.25%	59.58%	49.28%	97.14%	95.26%	96.73%	89.52%
14	92.48%	100%	62.54%	50.56%	97.84%	95.65%	97.23%	90.64%
15	94.35%	100%	65.48%	52.65%	98.41%	96.09%	97.23%	92.96%

5 Conclusion and Futures Works

This work implemented two neural network models: a Siamese Neural Network and a Triplet Network. Both models contain identical subnets composed by a Convolutional Neural Network (CNN), which is responsible for the extraction of image features, and an Autoencoder, which encodes the features produced by CNN and then decodes to maintain the most relevant information for the image identified during the network training.

Experiments were performed with three different datasets, with and without AE at the end of the subnets. The results showed that the use of AE generated a significant gain, both for the Neural Siamese Network and the Neural Triplet Network. With the AE, the accuracy rates increased in all cases, as well as the CMC Curve, and it was found that the AE provided a gain of up to 71.05 %.

From the experiments carried out using the networks with the AE, it was also found that the Neural Triplet Network obtained better performance in the people re-identification in comparison with the Neural Siamese Network, in all tested cases.

The Neural Triplet Network also presented a potential for person re-identification compared to other methods available in the state-of-the-art approaches. This network largely gained from the results in the CMC Curves, for the three considered datasets.

As future work, it is suggested to carry out experiments using other public datasets, including video datasets. Besides, an implementation of a Neural Quadriple model using the same subnet as in this work can be performed. This network could prove whether the number of subnets increases, the potential to re-identify people in images will increase as well.

Abbreviations

CNN: Convolutional Neural Network; AE: Autoencoder; CMC: Cumulative Match Curve.

Ethical Approval and Consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Public datasets were used to test our methods. Contact the corresponding author for data requests.

Competing interests

The authors declare that they have no competing interests.

Funding

The work has been funded by the Federal University of Technology of Parana.

Author's contributions

Enembreck contributed to the experimental results and the main idea of the article. Morais guided the development of the work and Martins worked on the text of the article. All authors read and approved the final paper.

Acknowledgements

The authors would like to thank UTFPR for their support in this research and Fábíá Enembreck is especially grateful for UTFPR's financial support.

Authors' information

Enembreck has a degree in Computer Engineering from the State University of Ponta Grossa, and a master's degree in Computer Science from the Federal University of Technology of Parana. Her current research interests include computer vision, pattern recognition, and deep learning.

Morais has a degree in Computer Science from Integrated Colleges of Caratinga, a Master's in Computer Science from the Federal University of Minas Gerais, and a Ph.D. in Computer Science from the State University of Campinas. Currently, he is a professor at the Department of Informatics at the Federal Technological University of Paraná, at the Ponta Grossa Campus. He works as a teacher and developer, with an emphasis on Computer Vision, Digital Image Processing, Pattern Recognition, Machine Learning, and Visual Tracking of Multiple Objects.

Martins has a degree in Computer Engineering from the University of Ponta Grossa and in Electronics Technology with an emphasis on Industrial Automation from the Federal University of Technology of Parana. She holds a master's degree in Electrical Engineering and Industrial Informatics from the Federal University of Technology of Parana and a Ph.D. in Electrical Engineering and Industrial Informatics from the Federal University of Technology of Parana. She is currently an adjunct professor at the Federal University of Technology of Parana, Campus Ponta Grossa. Works in the area of optimization and problem simulation, metaheuristics, evolutionary computing, and artificial intelligence.

Author details

¹Department of Informatics, Federal University of Technology of Parana, Ponta Grossa, BR. ²Department of Electronic, Federal University of Technology of Parana, Ponta Grossa, BR.

References

1. Tu, P.H., Doretto, G., Krahnstoeber, N.O., Perera, A.A., Wheeler, F.W., Liu, X., Rittscher, J., Sebastian, T.B., Yu, T., Harding, K.G.: An intelligent video framework for homeland protection. In: Unattended Ground, Sea, and Air Sensor Technologies and Applications IX, vol. 6562, p. 65620 (2007). International Society for Optics and Photonics
2. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3908–3916 (2015)
3. Bedagkar-Gala, A., Shah, S.K.: A survey of approaches and trends in person re-identification. *Image and Vision Computing* **32**(4), 270–286 (2014)
4. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1335–1344 (2016)
5. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European Conference on Computer Vision, pp. 262–275 (2008). Springer
6. Prosser, B.J., Zheng, W.-S., Gong, S., Xiang, T., Mary, Q.: Person re-identification by support vector ranking. In: BMVC, vol. 2, p. 6 (2010)
7. Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3586–3593 (2013)
8. Wu, S., Chen, Y.-C., Li, X., Wu, A.-C., You, J.-J., Zheng, W.-S.: An enhanced deep feature representation for person re-identification. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–8 (2016). IEEE
9. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159 (2014)
10. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Deep metric learning for person re-identification. In: 2014 22nd International Conference on Pattern Recognition, pp. 34–39 (2014). IEEE
11. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1325–1334 (2016)

12. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2275–2284 (2018)
13. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT press, ??? (2016)
14. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *nature* **521**(7553), 436–444 (2015)
15. Chollet, F.: Deep Learning Mit Python und Keras: Das Praxis-Handbuch Vom Entwickler der Keras-Bibliothek. MITP-Verlags GmbH & Co. KG, ??? (2018)
16. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **15**(1), 1929–1958 (2014)
17. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a " siamese" time delay neural network. In: Advances in Neural Information Processing Systems, pp. 737–744 (1994)
18. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, pp. 1735–1742 (2006). IEEE
19. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International Workshop on Similarity-Based Pattern Recognition, pp. 84–92 (2015). Springer
20. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
21. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), vol. 3, pp. 1–7 (2007). Citeseer
22. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: European Conference on Computer Vision, pp. 688–703 (2014). Springer
23. Sokolova, M., Japkowicz, N., Szpakowicz, S.: Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: Australasian Joint Conference on Artificial Intelligence, pp. 1015–1021 (2006). Springer
24. DeCann, B., Ross, A.: Relating roc and cmc curves via the biometric menagerie. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–8 (2013). IEEE
25. Bolle, R.M., Connell, J.H., Pankanti, S., Ratha, N.K., Senior, A.W.: The relation between the roc curve and the cmc. In: Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05), pp. 15–20 (2005). IEEE
26. Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2288–2295 (2012). IEEE
27. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2360–2367 (2010). IEEE
28. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1249–1258 (2016)
29. Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1239–1248 (2016)
30. Subramaniam, A., Chatterjee, M., Mittal, A.: Deep neural networks with inexact matching for person re-identification. In: Advances in Neural Information Processing Systems, pp. 2667–2675 (2016)

Figures

Figure 1

Subnet model proposed: the subnet receives an RGB image as input and is formed by a CNN with 4 convolutional layers, which produces a feature vector. This vector undergoes an AE for reconstruction. The subnet output is this vector reconstructed after going through a normalization step.

Figure 2

Network Architecture of the proposed using Contrastive Loss Function. Siamese neural network formed by two identical subnets that receive a different image to verify the similarity between them. For this, each image passes through a CNN and an DAE, that form the subnet, resulting in a features vector for each image. To verify the similarity, the Euclidean Distance between the vectors is calculated, passing through a Contrastive Loss function.

Figure 3

Network Architecture of the proposed using Triplet Loss Function. The network receives as input a triplet given by anchor, positive and negative. Since the image identified by anchor is a reference image, positive is a different image, but the same person as the anchor and negative is a different person image. In this model, each subnet treats one of the input images and finally the network uses a Triplet Loss Function that attempts to bring the anchor closer to the positive and away from the negative so that re-identification can be performed.

Figure 4

Triplet Loss minimizes the distance between an anchor and a positive, both of which have the same class, and maximizes the distance between the anchor and a negative of a different class [20].

Figure 5

ViPER: Examples of pairs of images from 10 different pedestrians taken by two cameras [21]

Figure 6

Experiments on VIPeR: Accuracy (%) of each network model in relation to the number of training epochs.

Figure 7

Experiments on ViPeR: CMC curves on ViPeR dataset

Figure 8

iLIDS-VID: Images of the same pedestrian, obtained through two non-overlapping cameras [22].

Figure 9

Experiments on i-LIDSVID: Accuracy (%) of each network model in relation to the number of training epochs.

Figure 10

Experiments on i-LIDSVID: CMC curves on i-LIDSVID dataset

Figure 11

CUHK03: Samples of pedestrian images observed in different camera views, two adjacent images have the same identity. [9]

Figure 12

Experiments on CUHK03: Accuracy (%) of each network model in relation to the number of training epochs.

Figure 13

Experiments on CUHK03: CMC curves on CUHK03 dataset

Figure 14

Accuracy (%) of each network model in relation to the number of training epochs.

Figure 15

CMC curves: Comparison of the CMC curves of the proposed networks.

Figure 16

Comparison of CMC Curves with state of the art methods, using the dataset VIPeR

Figure 17

Comparison of CMC Curves with state of the art methods, using the dataset i-LIDSVID

Figure 18

Comparison of CMC Curves with state of the art methods, using the dataset CUHK03