SIMILARITY LEARNING WITH LISTWISE RANKING FOR PERSON RE-IDENTIFICATION

Yiqiang Chen, Stefan Duffner, Andrei Stoian, Jean-Yves Dufour, Atilla Baskurt

SIMILARITY LEARNING WITH LISTWISE RANKING FOR PERSON RE-IDENTIFICATION

Yiqiang Chen* *Stefan Duffner** *Andrei Stoian* † *Jean-Yves Dufour*† *Atilla Baskurt**

*Université de Lyon, CNRS
*INSA-Lyon, LIRIS, UMR5205, France
†Thales Services, ThereSIS, Palaiseau, France

ABSTRACT

Person re-identification is an important task in video surveillance systems. It consists in matching an image of a probe person among a gallery image set of people detected from a network of surveillance cameras with non-overlapping fields of view. The main challenge of person re-identification is to find image representations that are discriminating the persons' identities and that are robust to the viewpoint, body pose, illumination changes and partial occlusions. In this paper, we proposed a metric learning approach based on a deep neural network using a novel loss function which we call the Rank-Triplet loss. This proposed loss function is based on the predicted and ground truth ranking of a list of instances instead of pairs or triplets and takes into account the improvement of evaluation measures during training. Through our experiments on two person re-identification datasets, we show that the new loss outperforms other common loss functions and that our approach achieves state-of-the-art results on these two datasets.

***Index Terms*—** Video surveillance, Person re-identification, Deep learning, Similarity learning

1. INTRODUCTION

Person re-identification is the problem of identifying people across images that have been captured by different surveillance cameras without overlapping fields of view. The task is receiving increasing attention because of its important applications in video surveillance such as cross-camera tracking, multi-camera behavior analysis and forensic search. However, this problem is challenging due to the large variations of lighting, pose, viewpoint and background. The images from the same individual can have very different appearance, and, different individuals may look similar in appearance.

Existing person re-identification approaches generally build a robust feature representation or learn a distance metric. The features used for re-identification are mainly variants of color histograms, Local Binary Patterns (LBP) or Gabor features. Some approaches use features that are specifically designed to be robust to common appearance variations, for example ELF [1], SADALF [2], LOMO features [3]. The main metric learning methods include Mahalanobis metrics like KISSME [4], LFDA [5] and XQDA [3].

With the recent success of deep learning for computer vision, many deep convolution neural network(CNN) architectures have been proposed for person re-identification. These deep learning models incorporate feature representation and distance metric into an integrated framework. To learn the features and the metric, different loss functions have been proposed such as contrastive loss, triplet loss or quadruplet loss. Unlike these existing losses, in this work, we propose a novel listwise loss function based on the predicted and ground truth ranking of a list of instances w.r.t. a query image.

Furthermore, existing deep learning methods are solely based on the minimization of a loss defined on a certain similarity metric between different examples. However, the final evaluation measures are computed on the overall ranking accuracy. Inspired by the learning-to-rank method LambdaRank, our optimisation approach directly incorporates these evaluation measures in the loss function. During training, each image in the training batch is used as probe image in turn and the rest as gallery. For each query, the mean average precision and rank 1 score are calculated. And triplets are formed by the probe image and a pair of misranked true and false correspondence. The loss of one triplet is weighted by the improvement of these evaluation measures by swapping the rank positions of the true and false correspondences.

To summarize, the main contributions of this paper are the following:

- We propose a novel listwise loss function based on list ranking for person re-identification. This loss considers the re-identification ranking problem in a conceptually more natural way than previous work by directly taking into account the ranking evaluation scores.
- We experimentally show that this loss outperforms other common loss functions and achieves state-of-the-art results.

2. RELATED WORK

Learning-to-rank is a class of techniques that learns a model for optimal ordering of a list of items. It is widely applied in information retrieval and natural language processing. Many learning-to-rank methods have been proposed in the literature, like pairwise approaches RankSVM [6], RankNet [7] and listwise approaches ListMLE [8] and LambdaRank [9]. Since person re-identification could be considered as a retrieval problem based on ranking, some person re-identification approaches applied these techniques like Prosser et al. [10] who reformulated the person re- identification problem as a ranking problem and learn a set of weak RankSVMs, each computed on a small set of data then combine them to build a stronger ranker using ensemble learning. Wang et al. [11] applied the ListMLE method to the person re-identification problem: they map a list of similarity scores to a probability distribution, then utilize the negative log likelihood of ground truth permutations as the loss function.

Deep metric learning based person re-identification in which the similarity of pedestrian is well measured. Several loss functions are proposed or applied in person re-identification. Yi et al [12] first proposed to apply a Siamese network to person re-identification. Ding et al. [13] applied the triplet loss to train a CNN for person re-identification. Chen et al. [14] applied a quadruplet loss which minimizes the difference between a positive pair from one identity and a negative pair from two different identities. Some methods exploit hard examples mining to enhance the learning procedure. Ahmed et al. [15], for example, used the difference of feature maps to measure the similarity and performing hard negative example mining. Shi et al. [16] proposed to perform moderate positive and negative example mining to ensure a stable training process and avoid perturbing the manifold learning by using hard examples. On the contrary, Hermans et al. [17] proposed to use the hardest positive and negative examples in each training batch to perform an effective triplet learning.

3. PROPOSED METHOD

In the following, we will first describe the learning-to-rank method LambdaRank and the person re-identification evaluation measures. Then we will explain how to perform our proposed Rank-Triplet loss learning in terms of the evaluation measures. An overview of our approach is shown in Fig. 1.

3.1. LambdaRank

LambdaRank is an improved learning-to-rank method based on RankNet. RankNet uses a neural network with a pair-based cross entropy cost. It is optimizing for the number of pairwise errors, which does not consider with some other information retrieval measures. However, the evaluation measures are not differentiable. Thus, they cannot directly be incorporated in the optimization. To tackle this problem, Burges et al. [9] proposed LambdaRank which simply scales the gradient of the loss function by the difference of the evaluation measure incurred by swapping the rank positions of two items, and they show an improvement of the overall ranking performance. In triplet learning for person re-identification, we face a similar problem. The classical triplet loss is defined on the partial order relations among identities, however, the ranking measures are calculated on the global order. That means that the triplet loss iteratively enforces pair-wise order relationships w.r.t. reference examples, but it is difficult to generalize this approach for optimizing the global order. In this regard, a listwise ranking is a better approximation of this global order relation, and adapt it to the person re-identification problem, as explained in Section 3.3.

3.2. Person re-identification evaluation measure

Cumulated Matching Characteristics (CMC) and mean average precision (mAP) are widely used performance measures for person re-identification. CMC evaluates the top n nearest images in the gallery set w.r.t. one probe image. If a correct match of a query image is at the k^{th} position (k⩽n), then this query is considered as success of rank n. In most cases, we look at the success of rank 1 (R1). The CMC curve shows the probability that a query identity appears in different-sized candidate lists. As for mAP, for each query, we calculate the area under the Precision-Recall curve, which is known as average precision (AP):

$$AP = \int_0^1 p(r)\,\mathrm{d}r \tag{1}$$

where p is the precision function of recall. Then, the mean value of APs of all queries, *i.e.* mAP, is calculated, which considers both precision and recall of an algorithm, thus providing a more suitable evaluation for a multi-shot re-identification setting.

According to the evaluation code provided by [18], the area under the precision-recall curve is approximated as:

$$AP = \sum_{k=1}^{N} \frac{p(k) + p(k-1)}{2} [r(k) - r(k-1)], \tag{2}$$

where k is the rank in the sequence of retrieved items. p and r are respectively the precision and recall at the rank k position. We define also p(0)=1 and r(0)=0. N is the number of images in the gallery set.

Since in our method the AP is calculated online during training, we propose to simplify this computation. In ranking problems, recall is the fraction of the items that are relevant to the query that are successfully retrieved, the variation r(k)-r(k-1) is different from zero only when a relevant item is retrieved through the sequence of retrieved items. We only

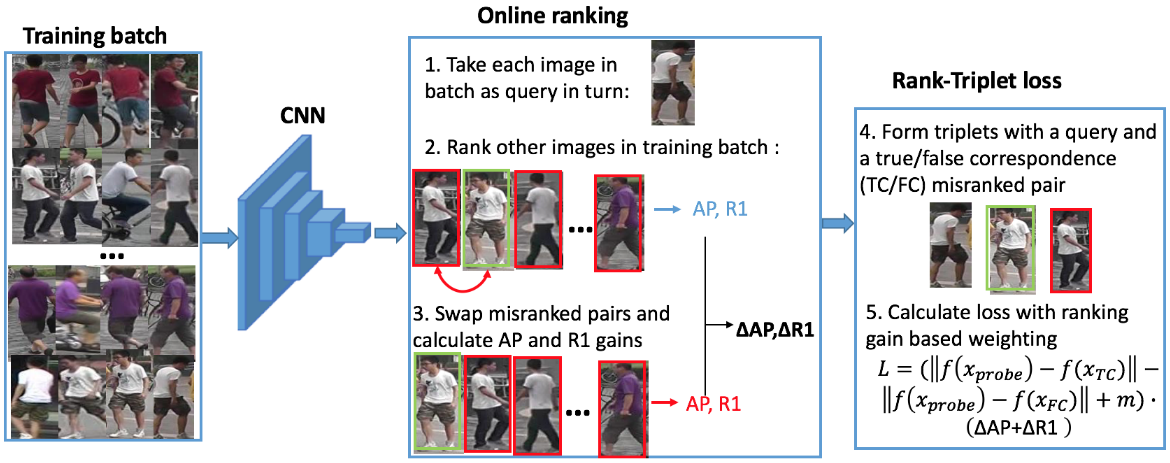
Training batch
CNN
Online ranking
1. Take each image in batch as query in turn:
2. Rank other images in training batch :
AP, R1
3. Swap misranked pairs and calculate AP and R1 gains
ΔAP,ΔR1
AP, R1
Rank-Triplet loss
4. Form triplets with a query and a true/false correspondence (TC/FC) misranked pair
5. Calculate loss with ranking gain based weighting
$L = (\|f(x_{probe}) - f(x_{TC})\| - \|f(x_{probe}) - f(x_{FC})\| + m) \cdot (\Delta AP + \Delta R1)$

Fig. 1. Overview of the training procedure of the proposed Rank-Triplet approach

need to calculate at the true correspondence ranking position and the variation of recall equals always $\frac{1}{M}$, where M is the number of the true correspondences of a query. thus the AP can be calculated as:

$$AP = \frac{1}{2M}[1 + p(\pi_1) + \sum_{i=2}^{M} p(\pi_i) + p(\pi_{i-1}))], \quad (3)$$

where π_i is the rank index of the i^{th} true correspondence. Precision is defined as the proportion of non-relevant items that are retrieved, out of all non-relevant items available. Thus the precision at ranking position π_i : $p(\pi_i) = \frac{i}{\pi_i}$. We can further simplify the equation:

$$AP = \frac{1}{M}\sum_{i=1}^{M}[\frac{i}{\pi_i}] + \frac{1}{2\pi_M} + \frac{1}{2M}. \quad (4)$$

3.3. RankTriplet loss

The triplet loss uses triplets of examples to train the network with an anchor image a, a positive image p from the same person as a and a negative image n from a different person. The weights of the network for the three input images are shared, and to train the network, the following triplet loss function is minimized:

$$E_{triplet} = -\frac{1}{N}\sum_{i=1}^{N}[max(\|f(a_i) - f(p_i)\|_2^2 - \|f(a_i) - f(n_i)\|_2^2 + m, 0)], \quad (5)$$

where N is the number of triplets, f is the projection of the network, and m is a margin. With the triplet loss function, the network learns a semantic distance metric by ”pushing” the negative image pairs apart and ”pulling” the positive images closer in the feature space.

A major drawback of the triplet loss is that the trivial triplets become inactive at a later learning stage. Hard triplet mining is an effective way to tackle this problem, but some too hard triplets may distort the manifold [16]. We propose to take into account all possible triplets to stabilize the training procedure and weight the triplet in function of their contribution to make the learning more effective.

In order to optimize directly the AP and R1 scores, we estimate the gain for AP and R1 of the triplets from an online ranking within a training batch. The training batch is formed by M images of N identities. For each example in the batch, we preform a ranking among the rest of images in the batch. For the sake of a robust metric, we add a margin *m* to the distance between the true correspondences and the probe before ranking. The AP and R1 scores are computed for each query ranking. Then w.r.t. one probe, we form all possible mis-ranked pairs (false correspondences ranked before the true correspondence), and we re-calculate the new AP and R1 scores by swapping positions of the pair in the ranking and thus obtain the gain ΔAP and $\Delta R1$. The loss of each triplet is weighted by the sum of the gain on AP and R1. The final Rank-triplet loss is calculated as follows:

$$E_{rank-triplet} = \frac{1}{MN}\sum_{i=1}^{MN}\frac{1}{K_i}\sum_{j \in TC_i}\sum_{\substack{k \in FC_i \\ r_k^i < r_j^i}}[\|f(x_i) - f(x_j)\|_2^2 - \|f(x_i) - f(x_k)\|_2^2 + m] \cdot (\Delta AP_{jk}^i + \Delta R1_{jk}^i), \quad (6)$$

where x_i is the i^{th} training example in a training batch. K_i is the number of misranked pairs w.r.t. the i^{th} example as query. r_j^i is the rank of the j^{th} example w.r.t. the i^{th} image as query. TC_i/FC_i is the true/false correspondence set of the i^{th} example. ΔAP_{jk}^i is the gain of AP by swapping the j^{th} and k^{th} examples w.r.t. the i^{th} example as query and analogously for R1.

Methods	R1	mAP
Classification loss	74.3	51.0
Hardbatch triplet loss [17]	81.0	63.9
Baseline	82.1	66.5
Rank-Triplet loss	83.6	67.3
Rank-Triplet+re-rank [19]	86.2	**79.8**
LOMO+XQDA [3]	43.8	22.2
LSRO [20]	78.1	56.2
SVDNet [21]	82.3	62.1
K-reciprocal re-rank [19]	77.1	63.6
JLML [22]	85.1	65.5
DPFL [23]	**88.6**	72.6

Table 1. Re-identification result on Market-1501

Methods	R1	mAP
Classification loss	62.7	40.4
Hardbatch triplet loss[17]	62.8	42.7
Baseline	72.4	52.0
Rank-Triplet loss	74.3	55.6
Rank-Triplet+re-rank [19]	78.6	**71.4**
LOMO+XQDA [3]	30.8	17.0
LSRO [20]	67.7	47.1
SVDNet [21]	76.7	56.8
DPFL [23]	**79.2**	60.6

Table 2. Re-identification result on DukeMTMC-Reid

This evaluation measure-based weighting makes better use of difficult triplets which can bring a larger rank improvement and are more effective for the learning, and at the same time, keep the learning stable by using all misranked pairs, since only using the hardest examples can in practice lead to bad local minima early in training.

4. EXPERIMENTS AND RESULTS

4.1. Datasets

The **Market-1501** dataset [18] is one of the largest publicly available datasets for human re-identification with 32,668 annotated bounding boxes of 1501 subjects. All images are resized to 128 × 48. The dataset is split into 751 identities for training and 750 identities for testing as in [18].

The **DukeMTMC-Reid** dataset [20] is collected with 8 cameras and used for cross-camera tracking. It contains 36,411 total bounding boxes from 1,404 identities. Half is used for training and the rest for testing.

4.2. Implementation Details

We take Resnet-50 [24] as the model architecture and the pre-trained weights from the ImageNet dataset are used as initialization. We replace the final layer of the Resnet-50 by a fully-connected layer with 256 output dimensions. Each input image is resized to 224 ×112 pixels. The augmentation is performed by randomly flipping the images and cropping central regions with random perturbation. The margin in the triplet loss is set to m =1. Adam optimizer is used and the initial learning rate is set to 10^{-4}. Each 80 epochs the learning rate is decreased by a factor of 0.1. The weight decay is set to 0.0005. The training is performed in 200 epochs. And the batch size is set to 128 from 32 identities with 4 images each. We implement the baseline with the loss function without evaluation gain weighting. We compared also to the classification softmax cross-entropy loss and the hard batch triplet loss in which the triplet loss is calculated as follows:

$$L_{hard-batch} = \frac{1}{MN}\sum_{i=1}^{MN} max(\max_{j \in TC_i} \|f(x_i) - f(x_j)\|_2^2 - \min_{k \in FC_i)} \|f(x_i) - f(x_k)\|_2^2 + m, 0). \quad (7)$$

The hardbatch triplet learning on DukeMTMC-Reid had difficulty to converge with an initial learning rate of 10^{-4}. We reduced the learning rate to 2×10^{-5}.

4.3. Experimental results

The results on Market-1501 and DukeMTMC-Reid are respectively shown in Tables 1 and 2. **Compared to different losses**, the Rank-Triplet loss gives a better performance. The improvement w.r.t. the baseline showed the effectiveness of the listwise evaluation measure-based weighting. The Hardbatch triplet gave an inferior result and a converge problem occurred on DukeMTMC-Reid. This could be due to some very similar negative examples and to some very different positive examples in the dataset. This demonstrates that hard example mining could make the learning more effective, but some too hard examples may severely perturb the learning procedure. **Comparison with state-of-the-art methods**. The proposed approach using Rank-Triplet loss outperforms most state-of-art methods. By combining it with the re-ranking techniques in [19], our approach achieves state-of-the-art results on both the Market 1501 and Duke-MTMC dataset.

5. CONCLUSION

In this paper, we presented a novel listwise loss function based on ranking evaluation measures. An online ranking within training batches is performed to evaluate the importance of different triplets of probe, misranked true and false correspondences and to weight the loss with the rank improvement for a given query. We experimentally showed that taking into account the evaluation measures and calculate the loss in a listwise way can improve the results. Also our proposed loss outperforms some other loss functions and achieved a state-of-the-art result on two different benchmarks.

6. REFERENCES

[1] Douglas Gray and Hai Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, 2008, pp. 262–275.

[2] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*. IEEE, 2010, pp. 2360–2367.

[3] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.

[4] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012, pp. 2288–2295.

[5] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *CVPR*, 2013, pp. 3318–3325.

[6] Ralf Herbrich, "Large margin rank boundaries for ordinal regression," *Advances in large margin classifiers*, pp. 115–132, 2000.

[7] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender, "Learning to rank using gradient descent," in *ICML*. ACM, 2005, pp. 89–96.

[8] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li, "Listwise approach to learning to rank: theory and algorithm," in *ICML*. ACM, 2008, pp. 1192–1199.

[9] Christopher J Burges, Robert Ragno, and Quoc V Le, "Learning to rank with nonsmooth cost functions," in *NIPS*, 2007, pp. 193–200.

[10] Bryan James Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary, "Person re-identification by support vector ranking.," in *BMVC*, 2010, vol. 2, p. 6.

[11] Jin Wang, Zheng Wang, Changxin Gao, Nong Sang, and Rui Huang, "Deeplist: Learning deep features with adaptive listwise constraint for person reidentification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 513–524, 2017.

[12] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li, "Deep metric learning for person re-identification," in *International Conference on Pattern Recognition*, 2014, pp. 34–39.

[13] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[14] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *CVPR*, 2017, vol. 2.

[15] Ejaz Ahmed, Michael Jones, and Tim K Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015, pp. 3908–3916.

[16] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z Li, "Embedding deep metric for person re-identification: A study against large variations," in *ECCV*. Springer, 2016, pp. 732–748.

[17] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[18] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.

[19] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li, "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017.

[20] Zhedong Zheng, Liang Zheng, and Yi Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *ICCV*, 2017.

[21] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang, "Svdnet for pedestrian retrieval," in *ICCV*, 2017.

[22] Wei Li, Xiatian Zhu, and Shaogang Gong, "Person re-identification by deep joint learning of multi-loss classification," in *International Joint Conference on Artificial Intelligence*, 2017.

[23] Yanbei Chen, Xiatian Zhu, and Shaogang Gong, "Person re-identification by deep learning multi-scale representations," in *CVPR*, 2017, pp. 2590–2600.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.