확장된 RNN을 활용한 사람재인식 시스템에 관한 연구

A Study on Person Re-Identification System using Enhanced RNN

최석규*, 허문걸**

Abstract The person Re-identification is the most challenging part of computer vision due to the significant changes in human pose and background clutter with occlusions. The picture from non-overlapping cameras enhance the difficulty to distinguish some person from the other. To reach a better performance match, most methods use feature selection and distance metrics separately to get discriminative representations and proper distance to describe the similarity between person and kind of ignoring some significant features. This situation has encouraged us to consider a novel method to deal with this problem. In this paper, we proposed an enhanced recurrent neural network with three-tier hierarchical network for person re-identification. Specifically, the proposed recurrent neural network (RNN) model contain an iterative expectation maximum (EM) algorithm and three-tier Hierarchical network to jointly learn both the discriminative features and metrics distance. The iterative EM algorithm can fully use of the feature extraction ability of convolutional neural network (CNN) which is in series before the RNN. By unsupervised learning, the EM framework can change the labels of the patches and train larger datasets. Through the three-tier hierarchical network, the convolutional neural network, recurrent network and pooling layer can jointly be a feature extractor to better train the network. The experimental result shows that comparing with other researchers’ approaches in this field, this method also can get a competitive accuracy. The influence of different component of this method will be analyzed and evaluated in the future research.

Key Words : CNN, RNN, Unsupervised Learning, Person-Reidentification
Ⅰ. Introduction

Human re-identification is a basic task in automated surveillance and has been a field of extensive research over the past years. Its aim is to distinguish the same person from the gallery according to a query picture. The challenge is that there are different cameras that monitor the same person, probing the same person from different source has some difficulty. Sometimes a camera may take hundreds of people in one day and some of them have a similar appearance. The same person under the different camera have significant challenges such as viewpoint variation, illumination, deformation, occlusion, background clutter, which make the person re-identification a challenge problem. Figure 1 shows some challenging examples in a most used dataset PRID_450S [15]. Each red box contains the matched person from the dataset PRID_450S.

There are traditional ways to predict the right patches form the gallery. Which use color model and weighted covariance estimation [17] to get the most similar part of the same person. Some traditional methods use algorithm to better get the salience to represent the patches. Our work is mainly manifested in two aspects. Given a query patch, the way to find correct person form a set of patches captured by the other camera must contain two important aspects. First, it must need a good image feature to represent the similar discriminative part of query and gallery patches. Good feature is necessary to the person re-identification due to it is the elemental unit of the picture. Second, it must need a fine-tuned distance metric to determine whether the gallery contains the query patch. This is critical because it determined the accuracy and similarity of the same person from the different person. Most methods use feature selection and distance metrics separately to get discriminative representations and proper distance to describe the similarity between people but some kind of ignoring some significant features or increasing the distance between the same people. State-of-art of the research has more focus on the first part and to get reliable and suitable feature to represent the person’s visible appearance. Once get the good features from the dataset and next step is to get the suitable distance metrics. Some papers use a common distance such as L1 (Manhattan) or L2 (Euclidean) distance matrices to get the distance between patches.

This situation has encouraged us to consider a novel method to deal with this problem. Due to the good feature and distance metrics are the both critical part of the method, we use an iterative expectation maximum (EM) algorithm [16] which is contained in recurrent neural network (RNN). This novel EM based method automatically locates discriminative patches through utilizing the spatial relationships. In particular, we set that there is a variable patch form the dataset picture that state whether the feature is discriminative (the true good feature is descript by discriminative part of
the picture). First of all, we set all the values into discriminative and then we train a RNN model that output the similarity of the same person according to the input data. We utilize the spatial smoothing to the final result similarity map and choose the patches that contain the most higher similarity and make it to be the real discriminative patches. We repeat this procedure by the true discriminative patches in the novel EM algorithm. Through the EM algorithm, the non-discriminative part is eliminated by the iterative process.

In our work we use enhanced recurrent neural network with three-tier hierarchical network for person re-identification. This ERNN-based architecture creates an inside circle to express the dynamic variation in the inputs, it not only can deal with the stationary patches but also the dynamic patches, so it is applicable on a broader field.

The main contributions of our proposed method is that we use the CNN architecture to better select the discriminative feature and use RNN to learning the distance metrics and join it together to be a group. To present a person’s patch, the architecture can produce an output of 19950 dimension feature. Though the CNN and the last max pooling, the dimension can reduce. The proposed three-tier hierarchical network can learn the distance matrix from a group of three pictures and this network make the same person’s distance closer and far away the different person.

II. RELATED WORK

In the work of person re-identification, person re-ID based on still images and person re-ID based on video frames are two kinds of person re-identification.

Generally, there are two main categories on re-identification for still images: the method for extracting features from the input images usually called invariant feature representation, and the measure for comparing these features between images usually called distance metric learning. The first of methods uses a method based on invariant features that attempt to extract features that are distinguishable and invariant for changes in the environment and viewpoints.\[23\][34] Employs an ensemble of discriminant localized features and classifiers, which selected by contributing to improve invariance of viewpoint, combining spatial and color information. Kviatkovsky et al.\[24\] show that the color as a single cue under extremely variable imaging conditions has relatively good performance in identifying persons, and present a novel illumination invariant feature representation which is based on the log chromaticity (log) color space. The existing features include: local binary patterns\[25\], salient color names \[26\], variations on color histograms\[25\], Gabor features \[14\], and local patches \[28\].

Secondly, after extracting features, metric learning has been widely used for person re-identification to learn the metric emphasizing inter-personal distance and de-emphasizing intra-person distance. The basic idea of metric learning methods is to find the mapping from feature space to the new space in which eigenvectors from different image pairs are further than eigenvectors from same image pairs. The metric learning is applied to make a final decision as to if the person has been on the re-identified or not. \[7\] Presents Large margin nearest neighbor metric (LMNN) to improve the traditional KNN classification performance. Scale Invariant Local Ternary Pattern (SILTP) histograms \[13\] and a high dimensional color representation are proposed. While preserving local discrimination, it constructs the histogram of pixel features and then acquires its maximum values within horizontal band to handle the viewpoint variations.

Although it is often assumed in many ways for re-identification that a single image represents each person, in many real-world scenarios the use of video means that multiple images can be applied to enhance performance. Recently, performing person re-id in video sequences has been considered. Existing methods
for video–based person re-id include Dynamic Time Warping (DTW)\textsuperscript{[31]}, conditional random field (CRF)\textsuperscript{[30]}, and top-push distance learning (TDL) Model \textsuperscript{[32][33]}.

The approaches of deep learning have also been exploited to video based person re-id to simultaneously solve the problem of feature representation and metric learning. In general, DNNs are applied to learn ranking functions based on image pairs\textsuperscript{[3]} or triplets\textsuperscript{[33]}. Such approaches typically depend on a deep network such as Siamese network\textsuperscript{[22]}, learning the direct mapping from original images to a feature space in which images from the identical person are approach while images from disparate person are widely separated.\textsuperscript{[27]} Propose a “Siamese” convolutional network for metric learning. It includes three independent convolutional networks acting on three overlapping parts of the two input images. Each particular part of the network is composed of two convolutional layers with max pooling. Then the fully connected layer following, generates an output vector for each input image, and through a cosine function compare the two output vectors. The final similarity score are obtained by the cosine outputs for each of the three parts.\textsuperscript{[10]} Propose two network: a recurrent neural network (RNN) and a Siamese network, one (RNN) is used to learn the interaction between multiple sequences in a video, the other is used to learn the distinguishing video–level features for person re-id. Long–Short Term Memory (LSTM) network\textsuperscript{[8]} is proposed to assemble frame–wise person features in a recurrent way.

In this paper, the proposed enhanced recurrent neural network with three–tier hierarchical network for person use the CNN architecture to select the discriminative feature and use RNN to learning the distance metrics. The iterative EM algorithm can fully use the feature extraction ability of convolutional neural network (CNN) which is before the RNN. By unsupervised learning, the EM framework can change the labels of the patches and train larger datasets.

### III. PROPOSED METHOD

In this section, we descrip the proposed methods in details. At the first, we will explain the whole neural network architecture about our person re-ID methods. Then we will specifically explain the proposed EM algorithm that joint with the RNN. Lastly, how can the proposed three–tier hierarchical network learn the distance metric from a group of three picture will be in detail.

1. **The overall structure**

As showed in figure 2, in order to train the network from the input images to get the most discriminative patches, we use triplet pictures as the input. Descript by the function of $\mathcal{T}_m = \rho(M,F) = \prod_{n=1}^{N} \rho(M_{im}, (F)_{im})$, there are three input images formed the $m$–th label, where the $I_i$ and $I_j$ are from the same person but $I_k$ are from the different person. As the other kind of CNN architecture\textsuperscript{[8]}, all the network share the same parameters $\mathbf{W}$. we extracted the discriminative features from the raw images into a learned feature space. This is like $\mathbf{A}(\mathbf{C}) \rightarrow \mathbf{A}(\mathbf{C}) \rightarrow \mathbf{A}(\mathbf{C}) \rightarrow \mathbf{A}(\mathbf{C})$. As a complete network. Each row of the architecture from the figure is able to extract features from the training data. After the proposed network model is trained through the three–tier hierarchical network, we define that between $T^2_{ik}$ and $T^2_{ij}$is more than the distance between $T^3_{ik}$ and $T^3_{ij}$, at the same time the $\mathbf{D}(I_i, I_k) < \mathbf{U} (\mathbf{U}$ is the defined number that limited the distance$.)$. the three–tier hierarchical network can learning the distance from the training data make the same person distance near and the different person far away.
2. The proposed recurrent neural network (RNN) model contain an iterative expectation maximum (EM) algorithm.

An overview of the EM are descript next. The dataset that we used is lower resolution images as patches that is extracted from the separate cameras. We have calculated the whole ground truth label that is contains the whole images not only some of the dataset. We decided the situation is just discriminative or non-discriminative as a binary tree.

We set that \( M = \{ M_1, M_2, M_3 \} \) as the group that contains \( N \) bags. For every bag \( M_j = \{ M_{j,1}, M_{j,2}, \ldots, M_{j,N} \} \) have \( N_j \) situations. At the j-th bag, the j-th situation and relative label is in the \( M_{j,i} < m_{j,i}, y_i > \).

Given that the bags are standalone and identically allocated, the \( M \) and the variables \( V \) can be calculated by the following model:

\[
P(M, V) = \prod_{i=1}^{n} \prod_{j=1}^{m_j} P(M_{j,i} | y_i) P(V_{m_j})
\]  

We maximize the patch similarity \( P(M) \) using the Expectation maximization.

First, at the beginning step, we should initial the \( E_i \) we set \( \eta_i = 1 \) for all the \( i, j \). From this set, we can consider all the situation is discriminative. In this step, we consider all the patch in a positive way to set the dataset.

At the maximization step, we make the patches similarity to become closer and make the parameter \( E \) to get the higher possibility.

\[
E_i = \arg \max P(M | V; \epsilon)
= \arg \max \prod_d \prod_{i=1}^{n} P(x_{i,d} | y_{i})
\]

Where the \( D \) represent the discriminative part of the patches. If we make a generative function for all the discriminative and latent instances, the optimization of the above function can be this:

\[
\arg \max \prod_d \prod_{i=1}^{n} P(x_{i,d} | y_{i}) P(x_{i,d} | \epsilon)
\]

Finally, we valued the hidden variables \( V \), specially, \( \eta_i = 1 \) is a certain threshold and it is initial at the first. As for the person re-ID, \( \eta(M_i, \epsilon) \) is got by applying Gaussian smoothing at \( P(x_{i,d} | y_{i}) \), though this way, the smoothing step can utilizes the spatial Mutual position of \( P(x_{i,d} | y_{i}) \). At last, we repeat back to the Maximization step until it is become convergence.

IV. EXPERIMENT RESULTS

In this section, we evaluate our method on the most popular dataset, and perform a comprehensive evaluation of the proposed method. And then we compare different variations of the functions to compare the results. First, we describe these datasets and evaluation principle. Second, through the experiment, we predict some reliable results compare with some traditional methods and state-of -art method for some specific situation. Finally, we illustrate some disadvantages and advantages of this method and state the future work that we will do next.

1. Datasets PRID–450S.

To perform our experiments, we consider the most challenging datasets. The PRID 450S Dataset consists of 450 images pairs of pedestrians captured by two
non-overlapping cameras. The main challenges are related to changes in viewpoint, pose as well as significant differences in background and illumination.

2. Evaluation principle

For the evaluation, we use the most useful cumulative match curve (CMC) metric for feature match. Towards the PRID_450S dataset, we choose half of it as the gallery and then choose the other half as the query for testing the network. We randomly choose the gallery and query to avoid some errors. For this dataset that we used, we use the camera A as the gallery and the picture from camera B are chosen to be the query. As for the multi-camera groups, we plan to choose two same person, one as the gallery and the other as the query. We make sure that each individual is included in the gallery dataset. When start to search the query picture from the gallery, the previous network should be executed first. The whole gallery picture is calculated through the network and get the discriminative feature from the result. Then, we compare the distance between the query picture and the whole gallery dataset using the L2 distance metric. Finally, we get the top-n nearest images. If the match picture is in the n nearest scale, the rank of the correct is rank n. In order to eliminate the deviation, we repeat this process ten times to get the mean number.

3. Experimental setting

Data augmentation: Similar to the references [1, 2], we use random partitions of the dataset and half of it as gallery and half it as the query. In order to eliminate deviations, we repeat this process ten times. In order to improve accuracy and increase training data, we use data augmentation before we apply the dataset. This is also conducive to prevent over-fitting problem. In practice, the original image is 100*220 pixels. But during the input process, we cut a region of 95*210 with a random shiver, finally we get the more reliable and abundant data to train. The result are get by the CMC curve by top n positions. If the match picture show up at the n-th position, then the result is rank n.

4. Feature Extraction

To represent the most discriminative part of the picture, we use the hand-craft label and ERNN followed by CNN at the same time. The dimension of the image is 19950, but after the CNN and ERNN, the dimension is reduced and get the relative feature.

5. Compared methods

To our best knowledge, our CNN+ERNN is the most novel method with better result compared with the other methods. We trust the better result comes from the combination of CNN and RNN with the three-tier architecture. Nevertheless, learning how to limit the matching features to the gallery from the non-overlapping camera is a very challenging task. But here we just do the experiment in the dataset of PRID_450S.

According to the results from table 1, we completed the improved outputs while compared to efficient impostor–based metric learning [18]. Large scale metric learning [22], Salient color names [19] and other approaches. For example, we reached much higher ranking rate comparing to KISSME method, this can be explained though the lower dimensional features and at the same time, same the calculating time.

<table>
<thead>
<tr>
<th>Methods</th>
<th>R=1</th>
<th>R=5</th>
<th>R=10</th>
<th>R=20</th>
<th>R=30</th>
</tr>
</thead>
<tbody>
<tr>
<td>KISSME [15]</td>
<td>33.0</td>
<td>59.8</td>
<td>71.0</td>
<td>79.0</td>
<td>84.5</td>
</tr>
<tr>
<td>EIML [4]</td>
<td>35.0</td>
<td>58.5</td>
<td>68.0</td>
<td>77.0</td>
<td>83.0</td>
</tr>
<tr>
<td>KPCA [5]</td>
<td>42.9</td>
<td>67.7</td>
<td>76.6</td>
<td>84.7</td>
<td>89.5</td>
</tr>
<tr>
<td>KPLS [6]</td>
<td>40.5</td>
<td>63.1</td>
<td>71.6</td>
<td>81.3</td>
<td>87.2</td>
</tr>
<tr>
<td>KPLS Mode A</td>
<td>50.0</td>
<td>78.6</td>
<td>87.0</td>
<td>93.7</td>
<td>96.0</td>
</tr>
<tr>
<td>CNN+ERNN</td>
<td>52.8</td>
<td>81.3</td>
<td>90.0</td>
<td>94.0</td>
<td>96.2</td>
</tr>
</tbody>
</table>

표 1. PRID–450S 데이터 세트에 대한 실험 결과

Table 1. Experimental result on PRID-450S dataset.
The figure 3 shows that CMC curve representation on the most state-of-art method of KPLS mode A and our method. This is applied on the person re-ID dataset PRID-450S. The contributions of our paper is that our proposed person re-ID architecture has three innovative part: 1) we use the CNN architecture to better select the discriminative feature and use RNN to learning the distance metrics and join it together to be a group. 2) the proposed EM algorithm that joint with the RNN can product patch-level outputs to see the discriminative part as a binary number, through the iteration, the most discriminative feature for person has been extracted. 3) the proposed three-tier hierarchical network can learn the distance matric from a group of three pictures and this network make the same person’s distance closer and far away the different person.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed an enhanced recurrent neural network with three-tier hierarchical network for person re-identification. Specifically, the proposed recurrent neural network (RNN) model contain an iterative expectation maximum (EM) algorithm and three-tier Hierarchical network to jointly learn both features and distance. The iterative EM algorithm can fully use of the feature extraction ability of convolutional neural network (CNN) which is in series before the RNN. Through the three-tier hierarchical network, the convolutional neural network, recurrent network and pooling layer can jointly be a feature extractor to better train the network. Comparing with the other result in this field, this method also can get a competitive accuracy. We will fine-tune the program to increase the accuracy and increase the compatibility to the other kind of dataset. Because RNN is an internal state of network that could express the dynamic temporal gesture, this makes it suitable for person re-ID in the video. This is a big step in the person tracking area also. We will explore this later and use this method run smoothly on the other kind of dataset.

References

Kernel Hierarchical PCA for Person Re-Identification” in ICLR 2016.


[26] Y. Yang, J. Yang, J. Yan, S. Liu, D. Yi, and S. Li. Salient color names for person re-identification. In