Person Re-Identification with RGB-D camera in Top-view configuration

Sara Abbonizio[1], Davide Manzoni[2], Massimo Martini [2], Marina Paolanti[2], Emanuele Frontoni[2].

***Abstract*—Video-based reidentification of people is an important task, which has received much attention in recent years due to increased demand in the area of surveillance and environmental monitoring. RGB-D cameras are used to analyze customer behavior and interactions. This solution provides additional approximate depth information, providing sufficient accuracy and resolution for indoor applications. For this project a data set for the re-id of people (named TVPR2, Top View Person Re-Identification 2) using an RGB-D camera in a top view configuration was used, where the camera was installed on the ceiling above the area to be analyzed. The main objective of the document is to carry out the phases of extraction and classification of the functionalities for the re-id activity in a configuration scenario with a top-down view that uses a series of functions extracted from the colour and depth images.**

The document is organised as follows: in Section 1 a brief introduction is given to the topic dealt with in the rest of the document; Section 2 provides a description of the approaches used in the context of the re-id; Section 3 provides details about the dataset used, (Section 3.1) and (Sections 3.2, 3.3) the methodology proposed for the extraction phase of the characteristics and the model of automatic learning implemented; Section 4 provides the results obtained from the various tests carried out. The conclusions and future work in this direction are proposed in Section 5.

I. Introduction

Person re-Identification (re-ID) tackles the problem of retrieving a specific person (i.e. query) in different images or videos, possibly taken from different cameras in different environments. Recently, the re-id of people has emerged as a very interesting and widely used tool for detecting and monitoring people under occlusion or partial coverage of the camera. This was made possible by the spread of cameras that are used in most public places such as shopping malls, public offices, airports, stations and museums.

Within stores, re-id can provide useful information to improve customer service and commercial space management. In fact, in this area, the re-id person becomes a useful tool for correctly recognizing customers in a store, for studying returning consumers and for classifying different groups and targets of buyers. RGB-D cameras are used to analyze customer behavior and interactions. This solution provides additional approximate depth information that can be combined with visual images, providing sufficient accuracy and resolution for indoor applications. The choice of the RGB-D camera in a top view configuration is preferred because of its greater suitability over a front view configuration, as it reduces the problem of occlusions and has the advantage of preserving privacy because a person's face is not recorded by the camera. The camera's viewpoint in the top view configuration is also the only one that allows you to simultaneously measure the anthropometric characteristics of people passing by and the interactions between buyers and the surrounding environment. In this regard, we have created a new data set for the re-id of people (TVPR, Top View Person Re-identification [1]) that uses an RGB-D camera in a top view configuration, in which the camera has been installed on the ceiling above the area to be analyzed. This data set includes data from 1000 people, acquired at different intervals of days and at different times.

Specifically, we focus on video-based person re-ID, that is, given a query video of one person, the system tries to identify this person in a set of gallery videos. Most of the recent existing video-based person reID methods are based on deep neural networks. Typically, three important parts have large impacts on a video-based person reID system: an image-level feature extractor (typically a Convolutional Neural Network, CNN), a temporal modeling module to aggregate image-level features and a loss function to train the network.

After the work done only with RGB images, we wanted to integrate our project with a Depth function through the use of Depth map, that is, images of the same size as the sample image, in which for each pixel the distance of the relative point of the scene is indicated with respect to the sensor of the acquisition device. In the following we will analyze in more detail these methods.

II. Related work

In recent years, with the introduction of cameras for monitoring and surveillance of environments, various approaches have been developed for the re-identification of people through images or video. These approaches can be of different types:

Biometric approaches, where the different instances of person are combined together and assigned to the same identity through the use of biometric features. The examples adopted in the real situation concern gait, faces, fingerprints, iris scans and so on [2]. They are reliable and effective solutions, but they require collaborative behaviour of people and appropriate sensors. Therefore, in the case of low resolution, poor visualizations and a non-cooperative audience, as in the case of common settings for surveillance cameras, these techniques are often not applicable.

Other approaches can be geometric, and they verify when more than one camera or sensor simultaneously collects information from the same area and geometric relationships

can be adopted between the fields of view to match the data [3]. Geometric relationships, when available, guarantee strong matches or, at least, a strict selection of candidates.

Another type of approach is based on aspect [4]. In aspect-based approaches, re-id can only be performed correctly if the aspect is retained between views. It consists of exploiting the colors and textures of clothing, perceived heights and other similar signals and can be considered a biometric approach. Occlusions, lighting changes, different sensor qualities and different points of view are some of the most difficult problems that make appearance-based re-id difficult. The approach we followed to realize this project was the one used in Jiyang Gao's article [5], that is an approach for video-based person reID based on deep neural networks. We also made use of Eitel's article[6] for the conversion of our depth images to RGB and finally Hazirbas's paper[7] for their approach into integrating depth and RGB data.

Neural Networks are parallel computational models that try to reproduce the functioning of biological neural networks. The neural networks are formed by various layers of processing units, called neurons, strongly interconnected between them. The network receives external signals on a layer of input nodes, each of which is connected with numerous internal nodes, organized into several levels. Each node processes the received signals and transmits the result to successive nodes up to the output layer. Through more or less numerous cycles of input-elaboration-outputs, in which the inputs present different variables, these neural networks become able to generalize and supply correct outputs associated to inputs which are not part of the training set. In the system we used there are three important parts that have a strong impact on video-based re-identification and that we will then analyze in the following parts of the article: an image-level feature extractor (typically a Convolutional Neural Network, CNN), a temporal modeling module to aggregate image-level features and a loss function to train the network.

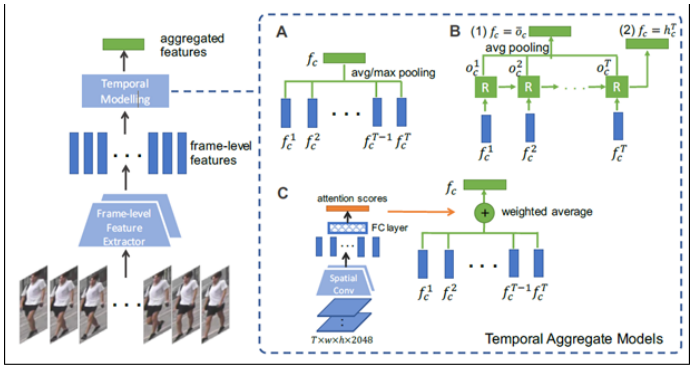

Fig. 1. Temporal Attention Methods

III. Methods

A. Dataset

Unlike other articles, a different dataset was used for the training and testing phase of the network, which was carried out previously. The dataset in question is called TVPR (Top View Person Re -identification) [1] which contains videos of 100 people recorded over several days by an RGB-D camera installed in a top view configuration. The recordings were made in an indoor scenario, where people passed under the camera installed on the ceiling of a laboratory. The reason why an RGB-D camera is used in a top-down view configuration is because of the improved applicability of the approach proposed in crowded public environments. The top view configuration reduces the problem of occlusion and has the advantage of preserving privacy because a person's face is not recorded by the camera. However, this difficult configuration does not allow you to retrieve front view features, which can be highly discriminating in the identification of the subject. Possible applications of this approach can be: security and protection in crowded environments, analysis of the flow of people, access control and counting of people.

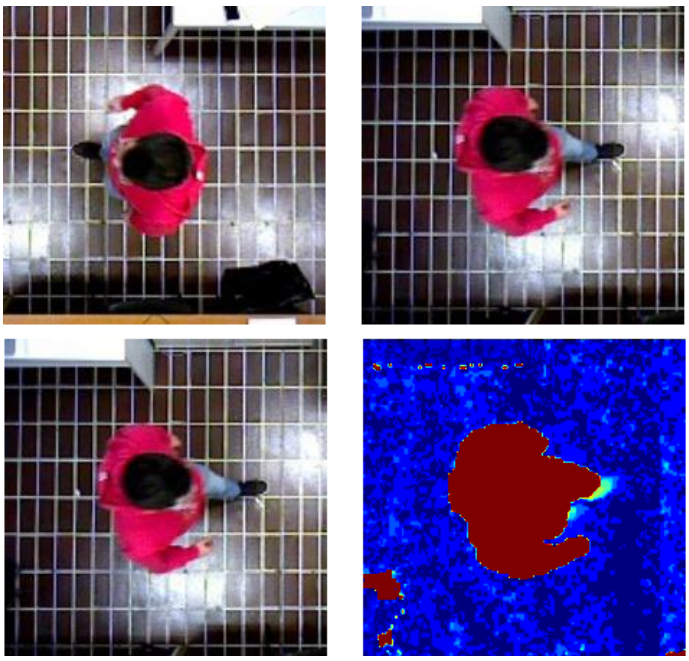

Fig. 2. An example of training and test set in our dataset and an exaple of a RGB image on the relative depth map

B. Input preprocessing

Before using our dataset in the training and testing of our network, some additional operations were necessary. To integrate depth images, we first had to modify the data loader so that each sequence of frames that forms a clip also includes the corresponding depth map for each frame. The depth images are then color encoded into RGB images to make them compatible as inputs for the CNN, using the jet colormap to transform the input from a single to a three channel image. This approach has been used becauce contrary to other encoding methods like HHA it is computationally inexpensive.

C. Methodology

In this section we will detail the general functioning of our network. To begin with, a network is generally formed of two parts, a feature extractor for images and a loss function to train the network. Our network is formed of two separate CNN that work in parallel, one on the dataset of RGB video frames and the other on the respective Depth Map of the video. The two CNNs form the two branches of our network, the RGB branch and the Depth branch. The feature maps from the depth branch are fused into the RGB branch at four different steps, allowing the RGB branch to learn how to combine rgb and depth features.

1) Video Clip Encoder: Since we were dealing with images (as the videos were divided into several frames), to build a video clip encoder, we have only considered the 2D CNN with temporal aggregation method, because a 3D CNN directly takes a video clip c which contains n frames as input and output a feature vector fc, while 2D CNN first extracts a sequence of image-level features and then aggregates them into a single vector fc by a temporal modeling method.

For 2D CNN, we adopt a standard ResNet-50 model as an image level feature extractor. We assume that we are given a training set consisting of M four-channel RGB-D images, having the same size H x W, along with the ground-truth labelling, while in output, we have a sequence of image level features $f_c^t, t[1,T]$, which is a $T \times D$ matrix (D is image level feature dimension).

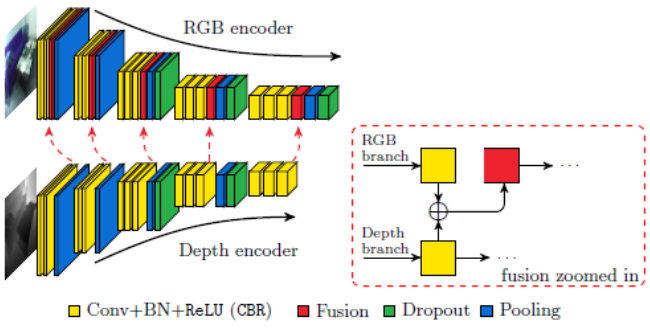

Fig. 3. The architecture of the proposed neural network. Colors indicate the layer type. The network contains two branches to extract features from RGB and depth images, and the feature maps from depth is constantly fused into the RGB branch, denoted with the red arrows. In our architecture, the fusion layer is implemented as an element-wise summation, demonstrated in the dashed box

Based on this assumption, we may define a CNN model to perform multinomial logistic regression. The network extracts features from the input layer and through filtering provides classification score for each label as an output at each pixel. We propose an encoder type network architecture as shown in Figure 2. This type of neural network has been already introduced in a previous work of Hazirbas et al. [7], where the neural network is divided in two branches: these two branches extract features from RGB and depth images. We note that the depth image is normalized to have the same value range as color images, i.e. into the interval of [0,255]. In order to combine information from both input modules, we use a fusion block. The fusion layer is implemented as element-wise summation in four different parts so that the RGB branch can learn features from both rgb and depth maps. By making use of fusion the discontinuities of the features maps computed on the depth image are added into the RGB branch in order to enhance the RGB feature maps. Moreover, with fusion, we do not only increase the activation values of neurons, but also preserve informations used for the recognition of features in the following layers of the neural network, so the proposed fusion strategy, preserves well all the useful information from both branches.

Then we apply a temporal aggregation method to aggregate the features into a single clip level feature fc, which is a D-dimensional vector. Specifically, we test two different temporal modeling methods: (1) temporal pooling, (2) temporal attention. The architectures of these methods are shown in Figure 2.

2) Temporal Pooling: In temporal pooling model, we consider average pooling:

$$fc = \frac{1}{T}\sum_{t=1}^{n} f^t c \tag{1}$$

3) Temporal Attention: In the temporal attention model, we use an attention weighted average on the sequence of image features:

$$fc = \frac{1}{T}\sum_{t=1}^{n} a^t c f^t c \tag{2}$$

where $a^t c$ is the attention for clip c.

4) Loss function: To train the network we use a triplet loss function[8] and a Softmax cross-entropy loss. To form a batch, we randomly sample P identities and randomly sample K clips for each identity with a total of PK clips in a batch. For each sample in the batch, the hardest positive and hardest negative samples are selected when forming the triplet to compute the loss. A baseline input called anchor is compared to a positive(truthy) and a negative(falsy) input; the distance from the anchor to the truthy input is maximized while the distance from the anchor to the falsy input is minimized. The goal of triplet this way is to make sure that two examples with the same label have their embeddings close together while two examples with different labels have their embeddings far away.

$$\begin{aligned} Ltriplet = \sum_{i=1}^{P}\sum_{a=1}^{K}[m + \max_{p=1,\ldots,K} D(fi,a,f^i p) \\ - \min_{j=1,\ldots,P} n = 1,\ldots,KD(fi,a,fj,n] \end{aligned} \tag{3}$$

The softmax function instead encourages the network to classify the PK clips to the correct identities.

$$Lsoftmax = -\frac{1}{PK}\sum_{i=1}^{P}\sum_{a=1}^{K} pi,a \log_q i,a \tag{4}$$

where pi,a and qi,a are the correct and predicted identities respectively of the sample i,a. Finally, the total loss is simply the sum of these losses.

$$Ltotal = Ltriplet\,total + Lsoftmax \tag{5}$$

IV. Experimental Protocol

A. Data

Our dataset consists of 1000 clips of 1000 different persons recorded from a top-down perspective by walking first from one direction which forms our train set and then from the opposite direction which will form our test set. Each clip consists of a collection of individual frames of the video, saved in a jpeg image. In addiction, each frame has a respective Depth map.

B. Parameters settings

We tested our network with different settings to examine its performance. We first set our sequence length to S=4 and train batch to B=6 and observed how our network performed with different splits of the dataset in train and test sets. The dataset was first split with a 300/700 ratio, with 300 of the 1000 clips being used for training and 700 for testing, and used to evaluate the performance of the newtwork with both temporal pooling and temporal attention. The dataset was then split with a 100/900 ratio, with 100 of the 1000 clips being used for training and 900 for testing, and still used both with temporal pooling and temporal attention. Finally, we kept the same 100/900 ratio and increased our sequence lenght to S=8 and train batch to B=16 and examined its performance. The learning rate for all tests is set to 0.0003

C. Evaluation metrics

Metrics are measures calculated in order to be able to give an opinion on the quality of the results of the research carried out. In our case,we used the mean average precision score (mAP) and the cumulative matching curve (CMC) at rank-1, rank-5, rank-10 and rank-20.

V. Results

In the following section we will report the results of the tests described in the previous section. First, we compare the results of the 300/700 split using temporal pooling and temporal attention, with S=4 and B=6

	mAP	CMC-1	CMC-5	CMC-10	CMC-20
TP	87.2	86.0	88.7	90.3	90.9
TA	89.5	88.6	90.6	90.9	90.9

Then we increased the sequence lenght to S=8 and the train batch to B=16

	mAP	CMC-1	CMC-5	CMC-10	CMC-20
TP	76.7	72.3	81.7	84.1	87.6
TA	93.4	92.9	94.1	94.4	94.6

Then we compare the results of the 100/900 split for temporal pooling and temporal attention, with S=4 anf B=6

	mAP	CMC-1	CMC-5	CMC-10	CMC-20
TP	87.1	85.7	88.7	89.2	89.3
TA	87.4	86.7	88.1	88.7	89.2

And finally, with the same 100/900 split we again increased the sequence lenght to S=8 and the train batch to B=16

	mAP	CMC-1	CMC-5	CMC-10	CMC-20
TP	89.9	88.2	92.0	93.1	93.1
TA	90.3	89.0	92.0	92.6	93.3

VI. Discussion

We can see that temporal attention on average has better performances than temporal pooling, with a difference between the two that ranges from 2.3% in the case of a training batch equal to B=6 to a difference of the mAP scores of 16.7% by increasing the training batch from 6 to 16 and sequence lenght to S=8 . The network keeps high performances even when reducing the training set to 100 clips, with temporal attention getting slightly worse while temporal pooling's performance decreases only by a small margin. Increasing the batch size and the sequence length produces a noticeable increase in performance with temporal attention, with temporal pooling performing less efficiently.

VII. Conclusions

We examined some of the state-of-the-art networks available at the moment for person re-identification and modified one of the better performing ones to adapt it to a different dataset and integrate Depth images along RGB images. Experimental results show that temporal attention generally performs better than temporal pooling, and gets slightly worse when training batch and sequence length are decreased. Future development could test the network on datasets with different perspectives from top view and implement the other two architectures, RNN and 3DCNN.

References

[1] Linciotti D., Paolanti M., Frontoni E., Mancini A., Zingaretti P., *Video Analytics for Face, Face Expression Recognition, and Audience Measurement*,Springer; Berlin, Germany: 2017. Person Re-Identification Dataset with RGB-D Camera in a Top-View Configuration.[Google Scholar]

[2] Havasi L., Szlvik Z., Szirnyi T. Eigenwalks, *Walk detection and biometrics from symmetry patterns; Proceedings of the IEEE International Conference on Image Processing*, Genoa, Italy,2005.

[3] Calderara S., Prati A., Cucchiara R. Hecol, *Homography and epipolar-based consistent labeling for outdoor park surveillance*,Comput. Vis. Image Understand,2011.

[4] Farenzena M., Bazzani L., Perina A., Murino V., Cristani M., *Person re-identification by symmetry-driven accumulation of local features; Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,San Francisco, USA,2010.

[5] Jyiang Gao, Ram Nevatia, *Revisiting Temporal Modeling fo Video-Based Person ReID. Computer Vision and Pattern Recognition.*,2018

[6] Eitel A., Springerberg J.B., Spinello L., Riedmiller M., Burgard W., *Multimodal Deep Learning for Robust RGB-D Object Recognition*,2015

[7] Hazirbas C., Ma L., Domokos C., Cremers D., *FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-based CNN Architecture*, 2017

[8] Kaiming H., Xiangyu Z., Shaoqing R., Jian S. *Deep Residual Learning for Image Recognition*,2015

[9] Hermans A., Beyer L., Leibe B., *In Defense of the Triplet Loss for Person Re-Identification*,2017