SeSAME: Re-identification-based ambient intelligence system for museum environment

SeSAME: re-identification-based Ambient Intelligence system for museum environment

Marina Paolanti[a,b,**], Roberto Pierdicca[c], Rocco Pietrini[b,d], Massimo Martini[b], Emanuele Frontoni[a,b]

[a] *University of Macerata, Department of Political Sciences, Communication and International Relations, 62100 Macerata, Italy*
[b] *Universitá Politecnica delle Marche, VRAI Vision Robotics and Artificial Intelligence Lab, Dipartimento di Ingegneria dell'Informazione, 60131 Ancona, Italy*
[c] *Universitá Politecnica delle Marche, Dipartimento di Ingegneria Civile, Edile e dell'Architettura, 60131 Ancona, Italy*
[d] *Grottini Lab S.R.L., Via Santa Maria in Potenza, 62017 Porto Recanati, Italy*

ABSTRACT

Nowadays, understanding and analysing visitors activities and behaviours is becoming imperative for personalising and improving the user experience in a museum environment. Users' behaviour can provide important statistics, insights and objective information about their interactions, such as attraction, attention and action. These data represent a precious value for the museum curators, and they are one of the parameters that need to be assessed. These information are collected through manual approaches based on questionnaires or visual observations. This procedure is time consuming and can be affected by the subjective interpretation of the evaluator. From such premises, SeSAME (Senseable Self Adapting Museum Environment) a novel system for collecting and analysing the behaviours of visitors inside a museum environment is presented in this paper. SeSAME is based on a multi-modal deep neural network architecture able to extract anthropometric and appearance features from RGB-D videos acquired in crowded environments. Our approach has been tested on four different temporal modelling methods to aggregate a sequence of image-level features into clip-level features. This paper uses as a benchmark TVPR2, a public dataset of acquired videos with an RGB-D camera in a top-view configuration, in the presence of persistent and temporarily heavy occlusion. Moreover, a dataset specifically collected for this work has been acquired in a real museum environment, which is Palazzo Buonaccorsi, an important historical building in Macerata, in Marche Region in the center of Italy. During the experimental phase, the evaluation metrics show the effectiveness and the suitability of the proposed method.

1. Introduction

In Cultural Heritage (CH) domain, as well as in museum environment, understanding and analysing users' activities and behaviours is becoming imperative. Users' behaviour can provide important statistics and insights on what happens inside this space, which are the successful exhibitions, and which are the interactions with the artworks (Quattrini et al., 2020). Nowadays, it is possible for museum curators and personnel obtaining feedback on museums thanks to online purchases, social networks and other communication channels (Nisiotis et al., 2020). However, there is a limited knowledge about the circumstances that occur during the visit, and the layout, the arrangement of works, the management of flows can be designed according to the real needs of users, after collecting their information. Museum exhibits are usually arranged considering the target of users. This condition emerges from the obstacle in understanding a priori visitors' interests (Karaman et al., 2016). The concept of *smart environment* identifies a place able to acquire and apply knowledge about the environment and its inhabitants, in order to improve both their experience (i.e. by automatically reacting to some events in a even more attractive and challenging way) and the knowledge of the space itself (i.e. by providing managers with useful information for security or arrangement reasons). By means of innovative technological applications, it is possible to leverage novel human space interaction paradigms over the existing proxemic interac-

**Corresponding author: Tel.: +39-0733-258-2723
e-mail: marina.paolanti@unimc.it (Marina Paolanti),
r.pierdicca@univpm.it (Roberto Pierdicca),
r.pietrini@pm.univpm.it (Rocco Pietrini),
emanuele.frontoni@unimc.it (Emanuele Frontoni)

tion space model-based user interfaces, nowadays determined by the purely aesthetic and essentially passive fruition of cultural objects (Alletto et al., 2015).

Currently, there is a lack of reliable solutions which can fulfil such important tasks. In fact, these data represent a precious value for the museum curators, and they are one of the parameters that need to be assessed (Lanir et al., 2017). For this reason, a data-driven approach for collection and analysis provides an objective and reliable source of information. One of the current trend is to configure location-aware services, i.e., applications driven by location information, in particular, by users movements in the environment (Chianese and Picciali, 2014), (Del Fiore et al., 2016). Recently, RGB-D cameras have demonstrated their suitability for solving this task. In fact, this kind of solution provides affordable, additional rough depth information coupled with visual images, offering enough resolution and accuracy for indoor applications. Furthermore, RGB-D camera in a top-view configuration reduces the problem of occlusion, it allows precise people counting, and it has the advantage of preserving privacy by not recording faces, and it is easier to set up on ceiling installation (Paolanti et al., 2020).

In this work, it is presented SeSAME (Senseable Self Adapting Museum Environment) a novel system for collecting and analysing the behaviours of visitors inside a museum environment. SeSAME uses re-identification (re-ID) techniques to perform visual profiling of visitor interest. Person re-ID is a challenging task in various application fields: security, surveillance, people monitoring and human activity recognition Wu et al. (2019), Wu et al. (2020). Person re-ID faces the problem of recognising and identifying people inside different images or video sequences considering their characterising features (Paolanti et al., 2018).

Considering museum environments, an automatic re-ID system can provide important information to improve the user experience. The re-ID of users that move within the museum space enables understanding which artworks are most attractive, the displacement inside the spaces, and possible stops, as well as classifying different users groups and targets. Several previous researches have adopted the top-view configuration because it facilitates the extraction of trajectory features and ensures greater robustness. Furthermore, reliable depth maps can provide valuable additional information that can significantly improve detection and tracking results (Liciotti et al., 2017). In a crowded environment (more than three people per square metre), an RGB-D system with top-view configuration provides high accuracy (Liciotti et al., 2018).

SeSAME is based on a temporal multimodal deep learning approach[1] to extract the anthropometric and the appearance features from RGB-D videos for RGB-D person re-ID. RGB-D images contain more information than RGB images, but there are two main issues to solve: how to combine these two modalities and how to extract efficient discriminative features from the depth channel. Our approach is based on (Eitel et al., 2015) to convert depth images to RGB, and then in order to merge these information, our approach is based on (Hazirbas et al., 2016). Moreover, we follow the work of (Gao and Nevatia, 2018) for video-based person re-ID, who used deep neural networks. Given the depth and RGB images, we consider a convolutional neural network (CNN) with two branches to extract their appearance features. The two branches have been then merged using a specific module to get overall image-level features as output. This approach is tested on four different temporal modelling methods to aggregate a sequence of image-level features into clip-level features. In addition, the deep learning model is robust to possible changes in people's clothes, between floors, since it learns people's features not only from the colors, but also from the depth channel and the temporal frames. In fact, it has been already demonstrated that the depth channel is more effective than the color descriptor (Liciotti et al., 2016) and (Paolanti et al., 2018).

Following the procedure outlined in (Paolanti et al., 2020), a new dataset has been collected in a real museum environment, which is Palazzo Buonaccorsi, an historical building in Macerata, in Marche Region in the center of Italy. It is composed of videos that contained RGB and depth channels. Each video recorded people on forward paths (left to right) for half the time and recorded the same people on return paths (right to left) for the other half of the time, though not necessarily in that order. The results of person re-ID are used for evaluating important indicators and statistics as the time spent in each floor of Palazzo Buonaccorsi, the most visited floor of this building as well as the attention and the interaction with the artworks.

Figure 1 depicts the architecture of SeSAME, based on a re-ID system adapted to a multi-camera museum environment.

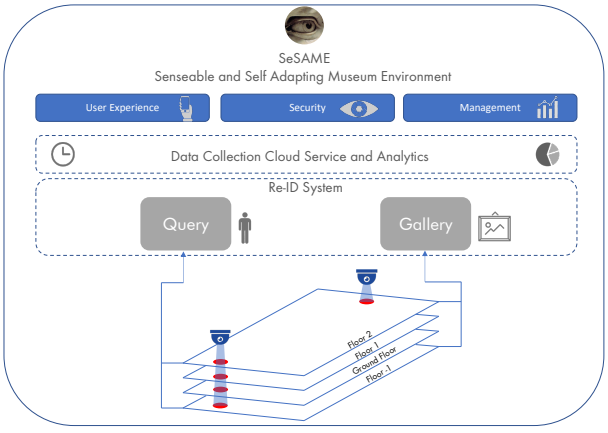


Fig. 1: Workflow of the museum multi-camera system. SeSAME consists of 6 cameras on 4 different floors. Each camera detects the passage of a person using a threshold in the depth channel frames. The network pretrained on TVPR2 was used as a feature extractor for people entering the building. These features are used to build the person gallery. The frames extracted from the RGB-D cameras are called query frames: the network extracts features from each of these frames, then compare them with all the gallery frames.

All in all, we can thus summarize the main contributions of this paper as follows:

- development of a real-world generalized smart environment system able to perform re-ID task for unseen data and in crowded environment.

[1] https://github.com/vrai-univpm/temporal_reid

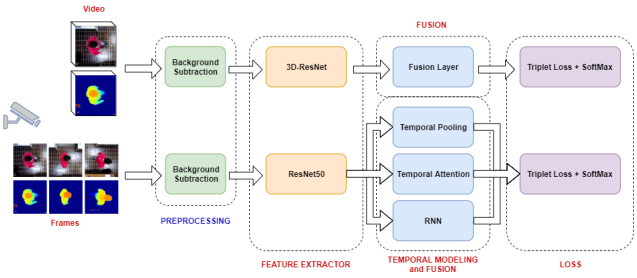


Fig. 2: Our temporal multimodal framework comprised three main components: a feature extractor, a temporal modelling module and a loss function. We tested four different methods based on this framework: a 3D-CNN, which does not need a temporal modelling method, and a 2D-CNN combined with three different temporal modelling modules. The last component was always a loss function designed to improve the network training. The dataset was initially processed in a preprocessing step to remove the backgrounds from the frames.

- development of a multimodal approach, tested in real environment which achieves results comparable with SoA approaches, which includes the temporal features which opens to the following contribution.
- development of a multi-camera re-ID system, which will allow a personalized and tailored user experience.

The paper is structured as described below. Section 2 illustrates the proposed approach. In Section 3, an exhaustive comparison of our approach with respect to state-of-the-art techniques is offered. Finally, in Section 4, conclusions and future trends for this field of research are explored and described.

2. Materials and Methods

In this section, the components of SeSAME are described; SeSAME is a vision-based person re-ID system for developing a tailored user experience in a museum environment. SeSAME is based on multimodal information and it is able to work in a multi-camera environment. As Figure 2 shows, there are three main modules of the implemented system: a frame-level feature extractor, a temporal modelling and fusion module to aggregate previous features, and a loss function (Gao and Nevatia, 2018).

The re-ID framework was comprehensively evaluated using TVPR2 (Martini et al., 2020), a publicly available dataset containing 235 recorded videos with a top-view configuration. Four different temporal approaches was tested on this dataset. Finally, the best approach was chosen for SeSAME in order to perform users' re-ID for personalising the visit experience. SeSAME has been applied on a newly collected dataset acquired in a real museum environment. ~~which is Palazzo Buonaccorsi, in the center of Italy.~~ A detailed description of the data collection and ground-truth labelling is presented in the subsection 2.3, including a preprocessing phase for the dataset.

2.1. Temporal Multimodal Person Re-Identification framework

A feature extractor typically employs a CNN. In this work, we tested two types of CNNs: a 3D-CNN and a 2D-CNN with a temporal aggregation method. The first type takes a video as an input and gives a feature vector f_v of the entire video as an output, while a 2D-CNN takes a sequence of frames as an input, produces a sequence of frame-level features $\{f_v^t\}$, then aggregates the entire sequence in a feature vector f_v using a temporal aggregation method.

First, the 3D ResNet model proposed in (Hara et al., 2018) was tested. It was a 3D-CNN designed for the task of action classification formed with ResNet architecture (He et al., 2016) and 3D convolutional kernels. ResNet is a well-known architecture in the field of image classification, a residual network which implements skip connections, allowing deeper architecture while maintaining high performance.
The 3D ResNet model was pre-trained using a Kinetics dataset (Kay et al., 2017), a kinetics human-action video dataset. The final classification layer was replaced with an adapted layer to classify people based on our dataset. The last layer before the classification was used as a feature-extractor, so its output was the representation of the recognised person.
For the 2D-CNN approach, a ResNet-50 model was tested. This network takes a sequence of frames as an input and gives a sequence of frame-level features $\{f_v^t\}$ as an output, which was fed into a temporal modelling module and produced the same output as the 3D-CNN.
For both CNNs, a second branch of the network was developed to extract depth features: the main idea is to duplicate the architecture of the RGB network and use it for other modalities, like the depth channel. The next step was to merge the feature maps extracted from each architecture, similar to the approach proposed in (Eitel et al., 2015). The merging layer is based on an element-wise summation of the features maps.
Another problem to solve was how to feed the network with depth frames. The network was designed to receive RGB images, that is, 3-channel images. Its depth duplicate must also receive images in that format. To convert depth images to RGB, our approach was based on (Eitel et al., 2015). As a first step, all depth values was normalised to the range [0, 255] by choosing a threshold for the maximum value to compare with 255. Then, a jet colormap was applied to this value matrix, which transformed it from a single to a 3-channel matrix (colorising the depth). This method essentially mapped every distance value to a pixel RGB value, ranging from blue (small distance) to red (large distance). According to (Eitel et al., 2015), the jet colormap is the best of all those used to convert depth information, even better than the HHA (Gupta et al., 2014) method, based on height above ground, horizontal disparity, and pixelwise angle between a normal surface and the direction of gravity.

2.1.1. Temporal Modelling and Fusion Module

Three different temporal modelling methods were tested:

- Temporal pooling
- Temporal attention (Liu et al., 2017)(Zhou et al., 2017)
- Recurrent neural network (RNN) (McLaughlin et al., 2016)(Yan et al., 2016).

Each method was tested by using the best parameter combination of (Gao and Nevatia, 2018). The temporal pooling module performs the average pooling of N frames:

$$f_v = \frac{1}{N}\sum_{t=1}^{N} f_v^t \tag{1}$$

In the temporal attention model, temporal attention scores was computed for each frame using a temporal generation network formed by a spatial and a temporal convolutional layer. Then, after computing the final attention score a_v^t by a softmax function, an attention-weighted average was applied:

$$f_v = \frac{1}{N}\sum_{t=1}^{N} a_v^t f_v^t \tag{2}$$

Finally, the tested RNN module was formed by long short-term memory cells only. The RNN outputs $\{o^t\}$ were averaged to produce the final feature vector:

$$f_v = \frac{1}{N}\sum_{t=1}^{N} o_v^t \tag{3}$$

This module also included the merging phase between the RGB and depth features based on the approach of (Hazirbas et al., 2016): it contained two branches to extract temporal features from RGB and depth data, in addition depth feature maps were constantly fused into the RGB branch. In our framework, the fusion layer was implemented as an element-wise summation. This was the implementation of the temporal pooling approach. The fusion of the features was performed at four different points of the temporal approach: the first point was the output of the Resnet-50 (with the classification part removed), while the following points were those between the average pooling layers. The same method was used both for the temporal attention approach and for those based on the RNN. In the approach in which 3D-ResNet was used, there was no phase for temporal modelling, as it was already part of the 3D network. The fusion of the features was thus implemented in five main points within the network itself.

2.1.2. Loss Function

Our networks were trained using a combination of a triplet loss function and a softmax cross-entropy function, as described in (Gao and Nevatia, 2018). This setup choice is due considering the literature in the field, specifically for re-ID task Hermans et al. (2017), Yuan et al. (2020). The triplet loss function implemented is the batch hard function presented in (Hermans et al., 2017), which allowed performing end-to-end learning between the input and the desired embedding space. This particular function can achieve state-of-the-art performance both with a pre-trained CNN and a model trained from scratch. The key idea is the following: P person IDs and K frames of each individual are randomly sampled, so that batches of PK frames are created for the training. The entire loss function is evaluated by defining triplet like in (Gao and Nevatia, 2018): the hardest positive and the hardest negative samples are selected within the batch, for each sample a of the batch itself.

$$L_{BH} = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{all\ anchors} + [m + \overbrace{\max_{p=1..K} D(f_a^i, f_p^i)}^{hardest\ positive} - \overbrace{\min_{\substack{j=1..P \\ n=1..K \\ j\neq 1}} D(f_a^i, f_n^j)}^{hardest\ negative}] \tag{4}$$

where the frames of the same person in different poses are related to the hard positive samples, while similar people are related to hard negative samples.

Finally, the L_{BH} function was combined with a classic softmax loss function, which helped the network correctly recognise the person in the PK input frames.

$$L = L_{BH} + L_{Softmax} \tag{5}$$

2.2. Museum multi-camera system

The best person re-ID approach has been chosen for a multi-camera system of a museum. The museum system consists of 6 cameras on 4 different floors. Each camera detects the passage of a person using a threshold in the depth channel frames. The saved frames also contain the timestamp of the detection, which is useful for obtaining important visitors statistics.
The network pretrained on TVPR2 was used as a feature extractor for people entering the building. These features are used to build the person gallery. The frames extracted from the RGB-D cameras are called query frames: the network extracts features from each of these frames, then compare them with all the gallery frames. The comparison is done by Euclidean distance. The result of the classification is the class of the gallery frame that has the smallest distance respect the query frame.

2.3. Top-View Visitors' Museum Dataset

A dataset containing museum's visitors has been specifically collected for this work, with 6 RGB-D cameras placed in 4 different floors of a real museum environment, which is Palazzo Buonaccorsi in Macerata, a city in Marche region in the center of Italy. The multi-camera system is as follows: one camera at the entrance of floor 0, one camera for the entrance and exit of floor -1, one camera for the entrance and one for the exit of floor 1, one camera for the entrance and one for the exit of floor 2. The acquisition has been made by using Orbbec Astra cameras, which recorded a total of 240.000 frames with $640x480$ pixel resolution (both RGB and depth). The main camera has collected 66.808 frames.
The proposed dataset consists of 55 days of video recordings from 10 July 2020 to 13 November 2020. People detected during these period has been 6200. Then, three days of registrations have been used for fine-tuning the proposed approach.

Figure 3 shows some frame examples for every camera of the museum surveillance system.

Fig. 3: Top-View Visitors' Museum Dataset: frame examples for every camera of the museum surveillance system.

2.3.1. Preprocessing Phase

Before using the dataset, the individual frames were subjected to a preprocessing phase, as seen in Figure 4: a people-detection algorithm was written in Python, making a crop of the person through a 150×150 pixel-bounding box. This was possible by using the depth channel and a threshold of the minimum height of people. This phase allows to remove noise produced by the frame background and with a crop size experimentally chosen. Depth information is very useful because it can be also used to remove the background inside the crop, as showed in Figure 4: using the same previous mask, is it possible to find the largest area's contour, then remove everything outside it. This preprocessing step is used in both phases of the proposed approach, namely the choice of the re-ID system and the multi-camera museum system. This phase also allows to handle in the best way the presence of more people inside the same frame. In fact, in this way, the network will receive in input only the portion containing the single person and not the whole frame, removing the background that would introduce only noise.

3. Results and Discussion

In this section, experiments carried out on the proposed framework will be described. The first experiment concerns the choice of the best person re-ID system, exploiting all the information available: the RGB colour channels, the depth channel, the temporal information given by the acquisition of videos with respect to the single images. As stated in the previous Sections, the best approach is used for the multi-camera system in museum environments. In particular, the proposed approach has to recognise people detected by 6 different cameras, and then evaluate statistics regarding the behaviour of the visitors inside the museum.

3.1. Deep Learning models for Re-Identification results

The first experiments concern 3D-CNN and 2D-CNN (with related temporal modelling methods) tested on TVPR2 dataset. The performance of these networks are evaluated in terms of mAP and CMC. The mean average precision (mAP) metric is defined as the mean of the maximum precision as a function of recall values. mAP can be used to measure accuracy in classification problems of ordered sequences. The cumulative matching characteristic (CMC) curves indicate the probability of finding the right match in the first n most-expected matches. The CMC curve metric is a common metric in the evaluation of re-ID methods.

The implemented networks are the following:

- ResNet-50 standard, pre-trained on ImageNet (2D-CNN);
- 3D ResNet-50, pre-trained on Kinectics (3D-CNN).

The networks are trained by using an Adam optimiser and a batch size of 32. The test phase needed gallery and query sets for each person. The gallery was created using the videos which contain first passage of the person under the camera (left to right), while the query set was based on videos in which it is show the return to the initial position (right to left). Each video was divided into many clips of fixed lengths (T frames) before the shorter-length clips were increased by duplicating the frames.

First, performance was evaluated by comparing two different ways of fusing the features extracted from the two networks – concatenating and summing them – after changing the encoding of the depth frames into false RGB colours by using JET colormap.

Table 1 reports the results using the JET colormap for the depth channel and concatenating the features extracted from both networks. Table 2 summarises the results using the JET colormap for the depth channel and summing the features extracted from both networks.

Table 1: Results using the JET colormap for the depth channel and concatenating the features extracted from both networks.

Model	*mAP*	CMC_1	CMC_5	CMC_{10}	CMC_{20}
3D-CNN	72.6	82.6	88.3	91.7	94.4
Temporal Pooling	74.2	82.7	90.3	95.2	96.6
Temporal Attention	**77.9**	**84.2**	**94.6**	**97.1**	**98.8**
RNN	73.5	80.4	91.5	93.9	95.1

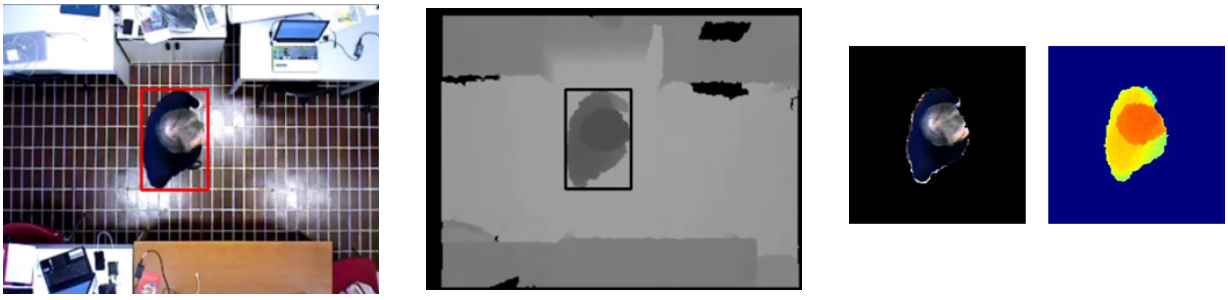
Table 2: Results using the JET colormap for the depth channel and summing the features extracted from both networks.

Model	*mAP*	CMC_1	CMC_5	CMC_{10}	CMC_{20}
3D-CNN	70.8	80.4	86.7	90.1	93.3
Temporal Pooling	72.1	80.6	88.1	90.2	94.3
Temporal Attention	**75.9**	**83.1**	**90.8**	**93.9**	**95.0**
RNN	70.5	81.3	85.5	89.3	93.8

This first part of experiments demonstrated the effectiveness of the feature concatenation respect the summation one, both in terms of mAP and CMC values. In fact, concatenation increased the mAP of all approaches by at least 1.8%, compared to the summation. The latter, sometimes, introduces some noise inside the learned features, since it is not guaranteed that the meaning of some features of the RGB branch is the same as those of the depth branch, given the same position. Instead, concatenation allows to use the information learned in both branches without adding any errors.

The second part regarded tests on the four temporal approaches to understand the benefits of the multimodal approach, where both RGB and depth information are merged.

The experiments performed were based on the following idea. Initially, the temporal approaches were tested using only the RGB stream without the depth information. Then, the results obtained were compared using both streams, RGB and



(a) RGB Frame (b) Depth frame (c) Person Detection

Fig. 4: This figure shows an example of TVPR2 Dataset and the relative output of the preprocessing phase. This phase finds the person inside the original RGB(Fig.5a) and depth (Fig.5b) frames. Finally the person is cropped, removing the entire background (Fig.5c). In addition, the depth channel is converted into the JET colormap.

depth, through the feature-fusion method. These two tests were carried out by training the networks with videos of 100 people and then validating them on another 100 people never seen by the network. This way allowed to verify whether the addition of the depth information improved the performances of the networks, which use only the depth ones. Finally, these trained networks have been tested on another 800 people never seen by the network. This test prove the generalisation of the network on larger datasets.

Table 3 shows the results a dataset of 100 people for training and another 100 people for the test with only the RGB features. We can infer that the Temporal Pooling approach achieves good results both in terms of mAP and CMC curves.

Model	*mAP*	CMC_1	CMC_5	CMC_{10}	CMC_{20}
3D-CNN	27.4	24.8	37.0	48.4	62.0
Temporal Pooling	**58.4**	**55.7**	**68.1**	**73.7**	**86.8**
Temporal Attention	11.9	8.9	20.3	31.1	42.0
RNN	16.5	10.9	27.1	34.9	46.1

Table 3: Experimental results using 100 people for training and another 100 people for the test. The approach only took the RGB stream as input without using the depth information.

Table 4 shows the results of using both the RGB and Depth features with a dataset of 100 people for training and another 100 people for the validation. As we can notice, depth information significantly improve the performances of the temporal approaches. Figure 5 shows the validation accuracy of all the temporal modelling approaches on 1000 epochs of training. All the methods tended to converge to a specific value after 400 epochs. The results on the validation set show that the Temporal Attention method obtains good performance in terms of mAP. However, it does not performs well on accuracy. The Temporal Pooling achieves the best performance in terms of CMC curves yet.

The trained networks are tested on the other 800 person recorded in the TVPR2 dataset, which have never been seen from the networks. The 100 validation people were chosen to be very heterogeneous among them, to have an exhaustive sample of the entire population, and thus discriminating different people carefully. Table 4 shows that the Temporal Pooling approach reaches performances in terms of mAP and CMC curves, hence demonstrating that this network generalised well on a larger dataset.

Finally, in Table 5 we compared our best approach, which is Temporal Pooling, with state-of-the-art methods proposed in the literature. To obtain a more realistic validation of our approach, we used the same parameters in all the tested approach, which are state-of-the-art methods concerning person re-ID from a top-view perspective. In particular, we have chosen: RGB-D-CNN Lejbolle et al. (2017), MAT Lejbolle et al. (2018), which has been improved by the same authors by a multimodal attention network called adding an attention module to extract local and discriminative features that were fused with globally extracted features and SLATT Lejbølle et al. (2019). Furthermore, in order to have a fair comparison, we chose the TVPR2 Paolanti et al. (2020) which is the bigger available dataset that comprises RGB and Depth streams in Top-View configuration labeled for people re-id. The results achieved show an increase both in terms of mAP and CMC curves, demonstrating the effectiveness and the suitability of our approach.

3.2. Experiments on Top-View Visitors' Museum Dataset results

The Temporal Pooling approach proved to be the best method for re-ID in a top-view configuration, with a multi-modal approach of RGB-D and Temporal data. This network, pre-trained on the TVPR2 dataset, was adapted for SeSAME.

The network allows to extract the features of the people detected at the entrance of the museum, generating a gallery of the incoming people. The camera system, thanks to the same network, was able to detect the same people at the various entrances and exits of the building floors. All this information was constantly sent to the final module of the framework, which generated very useful statistics about the visitors.

Each day, the system generates information for each user, concerning: day of acquisition; unique person id; all visited floors; entry timestamp; floor 1 entry timestamp; minutes spent on floor 1 ; floor 2 entry timestamp; minutes spent on floor 2; floor entry timestamp -1; minutes spent on floor -1 ; total minutes spent inside the museum.

From these information, SeSAME can generate the following daily statistics: overall people counting; people counting for floor 1 ; people counting for floor 2; people counting for floor

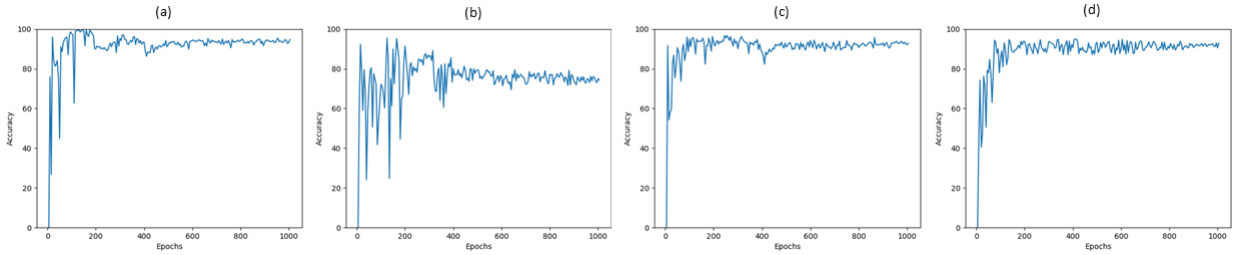


Fig. 5: Validation accuracy of all the Temporal Modeling Approaches on the training phase. (a) Temporal Pooling. (b) Temporal Attention. (c) RNN. (d) 3D-CNN

	Validation (100 People)					**Test (800 People)**				
Model	*mAP*	CMC_1	CMC_5	CMC_{10}	CMC_{20}	*mAP*	CMC_1	CMC_5	CMC_{10}	CMC_{20}
3D-CNN	67.0	91.3	94.1	96.0	97.9	77.2	93.0	96.5	98.1	99.0
Temporal Pooling	68.0	**94.7**	96.7	97.2	**99.2**	**81.6**	**98.8**	**99.4**	**99.7**	**99.7**
Temporal Attention	**68.3**	74.7	82.5	85.6	89.6	78.6	93.1	96.4	97.8	98.6
RNN	66.7	90.9	**97.5**	**98.5**	98.7	75.6	98.7	99.3	99.4	99.6

Table 4: Experimental results using 100 people for training and another 100 people for the Validation. The approach took both the RGB and depth streams as inputs. Finally these trained networks were tested on another 800 people.

Model	*mAP*	CMC_1	CMC_5	CMC_{10}	CMC_{20}
RGB-D-CNN Lejbolle et al. (2017)	70.7	81.4	85.2	91.1	93.7
MAT Lejbolle et al. (2018)	78.8	91.1	93.7	94.1	95.3
SLATT Lejbølle et al. (2019)	75.1	90.6	92.1	94.3	95.1
Temporal Pooling (our)	**81.6**	**98.8**	**99.4**	**99.7**	**99.7**

Table 5: Testing using the TVPR2 dataset comparing TL-DCNN with SLATT. Results are based on mAP and CMC curves.

-1; people counting for floor -1 and floor 1; people counting for floor -1 and floor 2; people counting for floor 1 and floor 2; people counting for floor -1, floor 1 and floor 2; dwell time inside the museum; dwell time inside floor 1; dwell time inside floor 2; dwell time in floor - 1. This is all important information that allows the museum to study users' behaviour on the different floors, and to make decisions in order to improve visitors' satisfaction. SeSAME evaluates the overall people counting and the people counting per each floor. A specific dashboard using this insight-based methodology has been designed as an in-depth evaluation of SeSAME[2]. The results show that the floor that generated the most interest was floor 1, while the least interesting floor was floor -1. It also reports the graph of the dwell time per day of the dataset: the system evaluates the overall dwell time and the dwell time on each floor. Even in this case, the results show that the floor that generated the most interest was floor 1, while the least interesting floor was floor -1.

Finally, Table 6 shows the average values of the statistics concerning people counting and dwell times, both in the museum and on each floor.

Statistic	*Museum*	*Floor1*	*Floor2*	*Floor-1*
People	112,27	77,4	51,2	20,9
Minutes	34,53	27,10	14,58	12,98

Table 6: Mean values of the statistics processed on the entire Museum dataset.

[2]https://public.tableau.com/app/profile/grottinilab/viz/SeSame/Museum.

4. Conclusion

In this paper, a novel deep learning approach for person re-ID in a top-view configuration was described and evaluated on real scenario. The problem has been solved through a multi-modal network that allows to extract visual features from RGB frames and useful information resulting from the depth channel. Starting from videos acquired in realistic conditions, the approach also allows to extract temporal features by using 4 different temporal approaches. All these information are efficiently and effectively integrated by the method proposed in this paper. The results demonstrate that the proposed methodology is relevant and accurate. Metrics like mAP and CMC curves prove the effectiveness and the suitability of the proposed approach on crowded scenario, where accurate people counting and re-ID are necessary (e.g. intelligent museum environments). SeSAME opens to an incredible number of possibilities. First of all, the data collected offer an objective way for evaluating the installation's performances, besides favouring the rearrangement of the museums' layout. Moreover, the knowledge coming from data might help the definition of security guidelines. And more, digital applications can be designed according to the needs of the users, reacting with the prior knowledge gathered by the ambient intelligence system. About this latter point, it is important to highlight that the system is multi-camera, meaning that Re-Id on multiple focus point may act, in a future implementation, as a driver for a personalized user experience based on his/her behaviour. Future works are focused on improving the performance by inte-

grating more complex CNN architectures. Incremental learning methods should also be investigated because can allows to enhance the online performance of this re-ID approach. Further investigations could cover strategies to get some demographic information such as gender, keeping the top-view approach to minimize occlusions, preserve privacy and be non-intrusive in the installation. Age group estimation can also be performed as well with a frontal view camera with a face analysis approach. The solution could be the integration of different systems (top-view and frontal view cameras) keeping in consideration the limitations of the frontal view approach and the privacy implications.

Finally, as the museum scenario can be very complex, further investigations of CNN generalisations are needed to test the effectiveness of the approach in different categories and in cross-country human behaviours and attitudes.

Acknowledgments

This work was funded by Grottini Lab (`www.grottinilab.com`). The authors would like to thank Massimo Martini, Mauro D'Aloisio, Luigi Di Bello and Marco Contigiani for their support to the work.

References

Alletto, S., Cucchiara, R., Del Fiore, G., Mainetti, L., Mighali, V., Patrono, L., Serra, G., 2015. An indoor location-aware system for an iot-based smart museum. IEEE Internet of Things Journal 3, 244–253.

Chianese, A., Piccialli, F., 2014. Designing a smart museum: When cultural heritage joins iot, in: 2014 eighth international conference on next generation mobile apps, services and technologies, IEEE. pp. 300–306.

Del Fiore, G., Mainetti, L., Mighali, V., Patrono, L., Alletto, S., Cucchiara, R., Serra, G., 2016. A location-aware architecture for an iot-based smart museum. International Journal of Electronic Government Research (IJEGR) 12, 39–55.

Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M., Burgard, W., 2015. Multimodal deep learning for robust rgb-d object recognition, in: Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on, IEEE. pp. 681–687.

Gao, J., Nevatia, R., 2018. Revisiting temporal modeling for video-based person reid. arXiv preprint arXiv:1805.02104 .

Gupta, S., Girshick, R., Arbeláez, P., Malik, J., 2014. Learning rich features from rgb-d images for object detection and segmentation, in: European Conference on Computer Vision, Springer. pp. 345–360.

Hara, K., Kataoka, H., Satoh, Y., 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 18–22.

Hazirbas, C., Ma, L., Domokos, C., Cremers, D., 2016. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture, in: Asian conference on computer vision, Springer. pp. 213–228.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hermans, A., Beyer, L., Leibe, B., 2017. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 .

Karaman, S., Bagdanov, A.D., Landucci, L., D'Amico, G., Ferracani, A., Pezzatini, D., Del Bimbo, A., 2016. Personalized multimedia content delivery on an interactive table by passive observation of museum visitors. Multimedia Tools and Applications 75, 3787–3811.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al., 2017. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 .

Lanir, J., Kuflik, T., Sheidin, J., Yavin, N., Leiderman, K., Segal, M., 2017. Visualizing museum visitors' behavior: Where do they go and what do they do there? Personal and Ubiquitous Computing 21, 313–326.

Lejbolle, A.R., Krogh, B., Nasrollahi, K., Moeslund, T.B., 2018. Attention in multimodal neural networks for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 179–187.

Lejbolle, A.R., Nasrollahi, K., Krogh, B., Moeslund, T.B., 2017. Multimodal neural network for overhead person re-identification, in: 2017 International Conference of the Biometrics Special Interest Group (BIOSIG), IEEE. pp. 1–5.

Lejbølle, A.R., Nasrollahi, K., Krogh, B., Moeslund, T.B., 2019. Person re-identification using spatial and layer-wise attention. IEEE Transactions on Information Forensics and Security .

Liciotti, D., Paolanti, M., Frontoni, E., Mancini, A., Zingaretti, P., 2016. Person re-identification dataset with rgb-d camera in a top-view configuration, in: Video Analytics. Face and Facial Expression Recognition and Audience Measurement. Springer, pp. 1–11.

Liciotti, D., Paolanti, M., Frontoni, E., Zingaretti, P., 2017. People detection and tracking from an rgb-d camera in top-view configuration: Review of challenges and applications, in: International Conference on Image Analysis and Processing, Springer. pp. 207–218.

Liciotti, D., Paolanti, M., Pietrini, R., Frontoni, E., Zingaretti, P., 2018. Convolutional networks for semantic heads segmentation using top-view depth data in crowded environment, in: Pattern Recognition (ICPR), 2018 24rd International Conference on, IEEE.

Liu, Y., Yan, J., Ouyang, W., 2017. Quality aware network for set to set recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5790–5799.

Martini, M., Paolanti, M., Frontoni, E., 2020. Open-world person re-identification with rgbd camera in top-view configuration for retail applications. IEEE Access 8, 67756–67765.

McLaughlin, N., Martinez del Rincon, J., Miller, P., 2016. Recurrent convolutional network for video-based person re-identification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1325–1334.

Nisiotis, L., Alboul, L., Beer, M., 2020. A prototype that fuses virtual reality, robots, and social networks to create a new cyber–physical–social ecosociety system for cultural heritage. Sustainability 12, 645.

Paolanti, M., Pietrini, R., Mancini, A., Frontoni, E., Zingaretti, P., 2020. Deep understanding of shopper behaviours and interactions using rgb-d vision. Machine Vision and Applications 31, 1–21.

Paolanti, M., Romeo, L., Liciotti, D., Pietrini, R., Cenci, A., Frontoni, E., Zingaretti, P., 2018. Person re-identification with rgb-d camera in top-view configuration through multiple nearest neighbor classifiers and neighborhood component features selection. Sensors 18, 3471.

Quattrini, R., Pierdicca, R., Paolanti, M., Clini, P., Nespeca, R., Frontoni, E., 2020. Digital interaction with 3d archaeological artefacts: evaluating user's behaviours at different representation scales. Digital Applications in Archaeology and Cultural Heritage 18, e00148.

Wu, D., Zheng, S.J., Zhang, X.P., Yuan, C.A., Cheng, F., Zhao, Y., Lin, Y.J., Zhao, Z.Q., Jiang, Y.L., Huang, D.S., 2019. Deep learning-based methods for person re-identification: A comprehensive review. Neurocomputing 337, 354–371.

Wu, Y., Zhang, K., Wu, D., Wang, C., Yuan, C.A., Qin, X., Zhu, T., Du, Y.C., Wang, H.L., Huang, D.S., 2020. Person reidentification by multiscale feature representation learning with random batch feature mask. IEEE Transactions on Cognitive and Developmental Systems 13, 865–874.

Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., Yang, X., 2016. Person re-identification via recurrent feature aggregation, in: European Conference on Computer Vision, Springer. pp. 701–716.

Yuan, Y., Chen, W., Yang, Y., Wang, Z., 2020. In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 354–355.

Zhou, Z., Huang, Y., Wang, W., Wang, L., Tan, T., 2017. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification, in: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE. pp. 6776–6785.