I-MALL An Effective Framework for Personalized Visits. Improving the Customer Experience in Stores

Federico Becattini
federico.becattini@unifi.it
Università degli Studi di Firenze
Italy

Giuseppe Becchi
giuseppe.becchi@unifi.it
Università degli Studi di Firenze
Italy

Andrea Ferracani
andrea.ferracani@unifi.it
Università degli Studi di Firenze
Italy

Alberto Del Bimbo
alberto.delbimbo@unifi.it
Università degli Studi di Firenze
Italy

Liliana Lo Presti
liliana.lopresti@unipa.it
Università degli Studi di Palermo
Italy

Giuseppe Mazzola
giuseppe.mazzola@unipa.it
Università degli Studi di Palermo
Italy

Marco La Cascia
marco.lacascia@unipa.it
Università degli Studi di Palermo
Italy

Federico Cunico
federico.cunico@univr.it
Università degli Studi di Verona
Italy

Andrea Toaiari
andrea.toaiari@univr.it
Università degli Studi di Verona
Italy

Marco Cristani
marco.cristani@univr.it
Università degli Studi di Verona
Italy

Antonio Greco
agreco@unisa.it
Università degli Studi di Salerno
Italy

Alessia Saggese
asaggese@unisa.it
Università degli Studi di Salerno
Italy

Mario Vento
mvento@unisa.it
Università degli Studi di Salerno
Italy

ABSTRACT

In this paper we present I-MALL, an ICT hardware and software infrastructure that enables the management of services related to places such as shopping malls, showrooms, and conferences held in dedicated facilities. I-MALL offers a network of services that perform customer behavior analysis through computer vision and provide personalized recommendations made available on digital signage terminals. The user can also interact with a social robot. Recommendations are inferred on the basis of the profile of interests computed by the system analysing the history of the customer visit and his/her behavior including information from his/her appearance, the route taken inside the facility, as well as his/her mood and gaze.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

computer vision, recommendation system, fashion recommendation, tracking, recognition

ACM Reference Format:
Federico Becattini, Giuseppe Becchi, Andrea Ferracani, Alberto Del Bimbo, Liliana Lo Presti, Giuseppe Mazzola, Marco La Cascia, Federico Cunico, Andrea Toaiari, Marco Cristani, Antonio Greco, Alessia Saggese, and Mario Vento. 2022. I-MALL An Effective Framework for Personalized Visits. Improving the Customer Experience in Stores. In *Proceedings of the 1st Workshop on Multimedia Computing towards Fashion Recommendation (MCFR '22), October 14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3552468.3555365

MCFR '22, October 14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9498-7/22/10…$15.00
https://doi.org/10.1145/3552468.3555365

1 INTRODUCTION

In this project we aim to verify the extent to which Computer Vision and deep learning can support individual profiling. Although we have a real application in mind as a reference for our research - the analysis of customer's behavior inside a Shopping Mall to provide personalized suggestions at digital signage terminals - our primary goal is to investigate key scientific computer vision issues that are central to such and similar applications and develop innovative solutions with respect to the state of the art. We will focus on the following research topics: 1) Re-identification of individuals: creating anonymized identities in open-world settings to allow continuous re-identification of people across different locations. 2)

Tracking and behavior analysis: understanding interests and detecting actions of individuals either alone or in a group. 3) Extraction of personal stable traits: learn hidden information such as social class and personality from clothing and behaviors. 4) Extraction of personal temporary feelings: understand emotional status and attention from face analysis. To foster the usage in real contexts, we will design privacy-respectful solutions.

2 TRACKING AND BEHAVIOR ANALYSIS

We have investigated and developed methodologies for detection and tracking of individuals either alone or in group to support effective behavior analysis and identification of intentions.

2.1 Open world re-identification

Traditional face recognition assumes a closed-set scenario with probe images containing identities that were enrolled in the gallery. A more realistic scenario is open-set where probes may contain subjects not enrolled in the gallery and the recognition system must detect and reject such probes. A more challenging scenario is the open-world where identities are learned incrementally and unsupervisedly in the gallery as soon as they are observed. In this scenario, the approaches for closed and open-set are not suited. Parametric learning methods like deep networks are natively designed to perform closed-set recognition and have been adapted in a few cases to the open-set. Recent research on Neural Turing Machines [21] showed that deep networks may be enhanced by an external memory for quick integration of information about new items, ensuring that salient but statistically infrequent data are stabilized in the class representation. Such model does not anyway support learning from video. I-MALL uses deep networks to perform detections and extract representations of face observations, exploiting an external memory module to incrementally collect them and a smart filtering mechanism that assigns observations to identities and decides their relevance to be learned. We use the memory to break up the temporal correlation between consecutive instances disrupting the non-id nature of video data. In this way, identity clusters are incrementally built putting together frequent and rare observations with no reference to their temporal occurrence. Since the individual outfit is used to learn personality, we exploit outfit features to improve the purity of the clusters. Continuity of appearance of face and outfit in consecutive frames is used as a form of self-supervision to decide the instantiation of a new identity. Scalability of the method in large settings such as a mall with large number of individuals is a key problem to address. Open-world recognition has been addressed so far by very few researchers, none of them providing satisfactory or well validated results. None of them addressed open-world re-identification from video. Our approach grounds on preliminary research results [33] that provided good confidence on the feasibility of the method.

2.2 Detection and tracking with 360 camera

We used a 360-degree camera to monitor the scene. These cameras consist of at least two lenses and can capture spherical images with a field of view of 360 degrees horizontally, and 180 degrees vertically. Thus, with each shot, they can fully sense the surrounding environment. The acquired spherical images are often stored through their equirectangular projection, shown in Fig. 1.(b). The pixel coordinates (x, y) of the equirectangular image represent the normalized values of the polar (ϕ) and azimuth (θ) angles of the corresponding point on the surface of the sphere.

Let us consider a 2D coordinate reference system centered on the projection of the 360-degree camera on the ground plane. As proposed in [29], under mild conditions, given the camera height h_c and the pixel coordinates (x_g, y_g) of a point P lying on the ground plane, it is possible to estimate the real world location in polar coordinates (d, θ) by the following equations:

$$\begin{aligned} d &= h_c \cdot \alpha \\ \alpha &= \frac{\frac{H}{2} - y_g}{\frac{H}{2}} \cdot \frac{\pi}{2} \\ \theta &= \frac{\frac{W}{2} - x_g}{\frac{W}{2}} \cdot \pi. \end{aligned}$$

where d represents the distance of P from the projection of the 360-degree camera on the ground palne, α is the angle between the ground plane and the line through the camera center and P (see Fig. 1.(c)), while H and W are the height and width of the equirectangular image respectively. The correspondence between pixel coordinates of the equirectangular image and ground-plane points holds only when the horizontal camera plane is parallel to the ground plane [29].

We have used this correspondence within a multi-object tracking method, preliminary presented in [28]. The method is based on the tracking-by-detection paradigm [5, 12, 43, 46]. It uses a pre-trained pedestrian detector to locate persons on the image plane. To account for the image circularity, the image is expanded at both sides and duplicated detected bounding-boxes are removed. The location on the ground in pixel coordinates is approximated by the middle point of the lower side of the bounding box, and is then transformed into real world coordinates by the above equations. At each frame, the targets' locations on the ground are predicted by Kalman filter; then, associations between the detector outputs and the predictions are found by using the Munkres algorithm considering both the distances measured on the ground and in an appearance feature space. Appearance features are extracted from a ResNet-50 [23] model. As pedestrian detector, we experimented with both Faster-RCNN [38] and YOLOv5 [7]. The latter yielded accurate detection results and it is faster. Inspired by SORT [5], we also implemented a different association strategy: first, we associate the most recently detected targets; later, the ones missing from the scene for more time. To detect pedestrians within apriori known regions of interests (RoIs), we first represent each rectangular RoI by means of two corners in the 2D coordinated reference system on the ground (hence, we use the relative position of the RoI and the camera). The tracking output is then used to locate pedestrians on the ground. Detection of persons within the RoI is done by comparing the person location with the RoI corners in the 2D coordinate reference system (see Fig. 1.(a)). We also keep track of the state associated with each person. A person can be in one of the following states: entered RoI, within RoI, exiting RoI, not in a RoI. Whenever a person enters a RoI, we initialize a counter of the number of frames the person is detected in the region. While the person stays inside the RoI, the counter is incremented. Furthermore, we attempt to detect the face by running a pre-trained face detector [35] within the pedestrian

image crop (based on the associated bounding box). Over time, we keep the face detection with the highest confidence score. When the person exits from the RoI, if the number of frames he/she has been detected inside the RoI is higher than a predefined threshold, the facial features (if available) and the information about the visited RoI are sent to a centralized database.

2.3 Gaze analysis

Capturing the attention of people towards specific elements in a scene, like advertising boards or shop windows in a mall, is an attractive yet unsolved problem in computer vision. In particular, we consider the visual selective attention (VSA) towards the scene, which is the process of directing the gaze to relevant visual stimuli while ignoring the irrelevant ones in the environment [9]. The task of detecting the target being looked at by a person in an image or video is known as attention target detection [2, 10, 18, 22, 24, 26, 37, 45]. However, in the scenario presented in this work we have a special case of this task in which the interest targets are always outside the video frame. While in most cases the problem is tackled in the 2D image space [10, 37, 45], we decided to extend our approach to work in a fully 3D manner. The scenario consists of 3 shop windows, where a camera has been placed inside each one to record the outside. When a customer approaches one of the windows, the bounding box around the head is extracted in order to identify them using the re-identification module. Then the captured video stream is processed by a top-down 2D pose estimation algorithm [42]. The extracted 2D pose joints are then converted by a 3D pose lifter [32] to a 17-joint 3D skeletal model, centered initially at the axis origin. The vector passing through the two joints on the head, one for the nose and the other for the center of the head, represents the direction of the gaze of the subject. A view frustum is then constructed around this vector. As long as the person stands sufficiently still and faces the window, an attention value is summed over time above the intersection between the view frustum and the planar point cloud corresponding to the shop window. A weighting scheme is also implemented, with a maximum around the central axis of the frustum and exponentially decreasing at the margins, similarly as in the attention spotlight model [34]. For implementation purposes, the weighting is done by dividing the view frustum into 3 concentric view frustums of increasing size. Since we are interested in the genuine interest toward scene elements, we exploit the idea of discarding the moments when people are engaged in a social interaction, looking at each other, to avoid false positive estimations. Each showcase can be divided into N tiles, respectively associated with N possible objects of interest within it. In our case, each window is divided into 3 sections of equal size. At the end of a visit session, the point with the highest attention value will indicate the section, and thus the object, of greatest interest, which will be updated with the corresponding id within the user's profile. In addition to the single most interesting object, it is also possible to obtain a ranking of interest of all the objects (of the mall or the shops individually) for each visitor. In a larger and more complex setup, with various objects of interest distributed in the scene, our method showed promising results, with a top-1 accuracy of up to 71.5%.

3 EXTRACTION OF PERSONAL STABLE TRAITS

Personal traits like personality and social status are stable social signals whose value does not change suddenly. Social signal processing (the marriage between pattern recognition and social psychology) has mainly focused on gestures, posture, gaze, physical appearance, and proxemics to extract social signals. Other features like clothing have been anyway recognized as having high potential in communicating personality and social status, being dependent on conscious choices and not as transient as a gestures.

3.1 Face Extraction of personal stable traits

Face analysis for gender, age, ethnicity and emotion has achieved impressive performance in the last years thanks to the employing of modern convolutional neural networks (CNNs). Anyway, where the four tasks are simultaneously required and a response is needed in real time, the system has to run four different CNNs, which is a quite burdensome operation. This is especially true when dealing with social robotic applications, where the software can not run over powerful servers with GPUs but instead needs to be installed on embedded systems, integrated directly on board of the robot, thus having strict constraints in terms of computational capability and available memory.

A possible solution to this problem is the use of a single *multi-task network*, trained with data from the four tasks simultaneously; the main idea is that the first layers of the network are used to learn an intermediate representation of the input data (in our case, of the facial features), that is then used for solving each of the specific classification problems [11] [48] [4] [30] [31]. The main evident advantage of this choice is that this network requires less computations and less memory with respect to the use of independent networks for each task. Anyway, it has been shown that if the tasks are sufficiently related, it is likely that the optimal intermediate representations for the various tasks will be quite similar to each other, and then a multi-task network may even achieve a better performance than the individual single-task networks, since it can exploit the sharing of the knowledge among the different tasks.

In the I-MALL system we exploit a multi-task network which simultaneously deals the four above mentioned face analysis tasks simultaneously [19]. It is worth to mention that this specific combination introduces three main problems: (i) there are no datasets in the literature properly annotated with gender, age, ethnicity and emotion labels, so the learning procedure of such a multi-task CNN must take into account the problem of missing labels; (ii) the number of available images for the various tasks may not be balanced. In particular, for our specific application, images annotated with emotion labels are in minority. If this imbalance is not considered in the learning procedure, there is a risk of penalizing the emotion recognition task in favor of others. (iii) we are mixing classification tasks having different numbers of classes (two for gender, four for ethnicity, seven for emotion) with a regression task (age estimation). Thus, the corresponding loss functions have different ranges and slopes; therefore, the loss function of the multi-task CNN must take into account this possible imbalance so as not to penalize any task during the learning [47]. We solve the problem of missing labels, dataset imbalance and loss function imbalance in the joint training

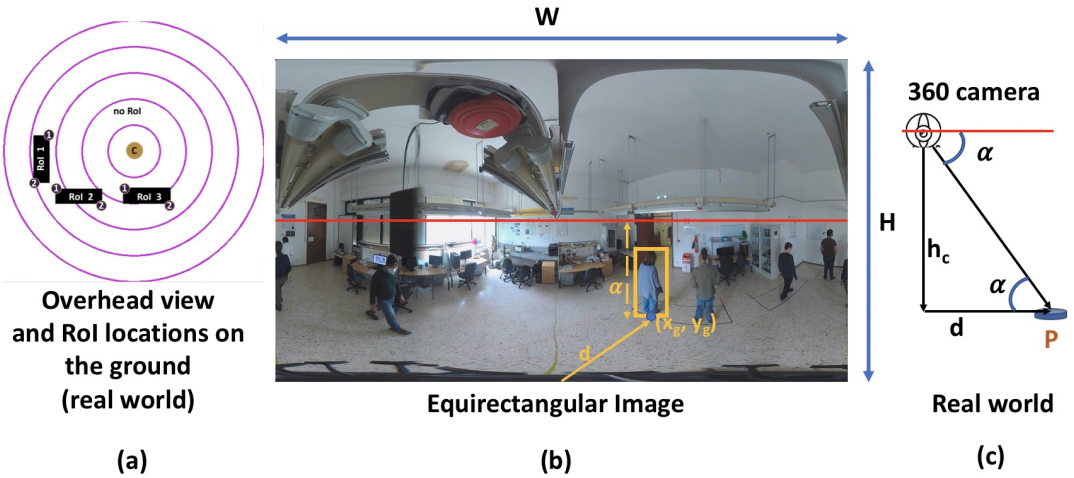

Figure 1: Image (a) represents an overhead view of the scene and the location on the ground plane of three RoIs, identified by two corners in a coordinate reference system centered on the camera C. Image (b) shows an equirectangular image of size $H \times W$. The red line is the equator of the sphere, namely the intersection of the sphere with the horizontal camera plane. The orange bounding box encloses a pedestrian whose location on the ground P is approximated by the point (in pixel coordinates) (x_g, y_g). The angle α is measured as the angular distance from the Equator line. In the real world (image (c)), the distance d on the ground-plane of the subject to the camera can be estimated by knowing the camera height h_c and the angle α.

by defining a custom learning procedure based on label masking, batch balancing and a custom weighted loss function [19]. The proposed multi-task CNN achieves an accuracy comparable with the one obtained by the corresponding single task CNNs, but reducing the overall processing time and the memory requirements by 2.5 to 4 times.

3.2 Body Clothing analysis

Analyzing the clothing style of people is a crucial step to obtaining stable traits and performing an appropriate garment recommendation. In this sense, the idea is to retrieve the clothes worn by people and perform a recommendation of clothing offered by the fashion companies inside the mall. Although the clothes worn by visitors are not necessarily indicative of the customer's purchasing intentions, they can still reveal interesting insights into the general preferences of each one. For instance, a person who wears a morning coat will likely appreciate elegant clothes. On the contrary, a person wearing large pants and a t-shirt will possibly like casual clothes. We therefore propose a method to classify in broad terms the style of the clients, in order to provide further useful information to the recommendation system. When a subject approaches the totem, a small video sequence is captured by the system and the crop of the face is used to identify the person through the re-identification module. We then use the pretrained model introduced in MovingFashion [20] to perform the detection of the top and bottom part of the outfit. The crops of the clothes are fed to a simple convolutional network that extracts a collection of attributes, derived from the DeepFashion dataset [27], which we then use in a naive manner to infer the style of the person's clothing.

4 DIGITAL SIGNAGE AND PERSONAL ASSISTANT

The personal assistant application is the core of the I-MALL digital signage. It will exploit open-world re-identification and person detection and tracking to detect the items of interest of each individual.

4.1 Extraction of personal temporary feelings

Understanding the feeling of an individual while looking a specific advertising content, a shopping window or a particular product in the shelf is very important for the retailers in order to understand the degree of appreciation of a particular product. Most of face expression analysis solutions in the literature consider still image faces acquired in controlled conditions by a traditional surveillance camera and detect the traditional seven classes of emotion Joy, Sadness, Anger, Fear, Surprise, Contempt, and Disgust [17]. Few of them have addressed emotion in faces extracted by videos in real environments with overexposure or underexposure lighting conditions, or blurring due to movements. There is also little attention to computational requirements and real time constraints, as required in real contexts. We exploit two different approaches based on body and face analysis. Our body emotion recognition system integrates spatial pose and temporal relations between movements. The model is trained on videos of humans in a shopping mall-like environment observing different recommended products. The users can express themselves without restrictions and just provide their interest level towards the observed fashion items. As a pre-processing step we use the OpenPose [8] to extract pose information, which is then processed in sequence thanks to an LSTM. A detailed overview of the body emotion recognition system has been presented in [6]. Similarly, we adopt a face expression analysis module to understand

implicit reactions of the users. As for body emotion, we trained our model on data of users observing fashion recommendations in order to infer personal preference from video footage alone. Our model is based on a pixel-based branch and on a landmark-based branch, which are then fused thanks to a mutual learning strategy. Further details are provided in [3].

4.2 Social Robot

MIVIABot is the social robot in charge of interacting in a smart way with the human, suggesting the specific product he/she may be interested in depending on its biometric characteristics. MIVIABot is based on Pepper robotic platform by Aldebaran and its hardware capabilities have been extended with a microphone ReSpeaker 4 Mic Array v2.0, an Intel RealSense Depth Camera D435 and a processing unit based on an NVIDIA Jetson Xavier NX. The software architecture, shown in Fig. 2, has been designed and developed by using Robotic Operating System (ROS), the de-facto standard framework for robot programming. All the software is deployed on board of the Jetson, fully controlling the robot. One of the main advantages deriving from the combination between the chosen hardware setup and the designed software architecture is that the system is robot-agnostic, in the sense that it is possible to change the social robot and/or sensory equipment by just changing the two acquisition nodes (*Video Input* and *Audio Input*).

In the video processing pipeline, the sequence of frames is analyzed by two ROS nodes, namely *Object Detection* and *Face Detection*. The former is responsible for localizing and recognizing the objects in the environment in which the social robot has to work. The latter has to identify the position of the people in the scene that are willing to interact with the robot; the face detected is passed to a *Face Embedding* and *Soft-Biometrics Recognition* nodes. The first node extracts a vector representation of the detected faces in order to allow similarity-based tasks like People Tracking [15] and Face Re-Identification [33]. In parallel, a ROS node extracts from each face the soft-biometrics, i.e. age, gender, ethnicity and emotion, by means of the multitask network described before. After that all the above data have been extracted, a ROS node synchronizes them in order to manage the different inference times of the involved ROS nodes. The audio processing pipeline is activated by a *Voice Activity Detection* (VAD) [25] node, responsible for identifying utterances in the audio stream and retrieving the related signals. The signal is processed by other two additional ROS nodes: *Speaker Biometrics* and *Speech Recognition*. The first node extracts all the information related to the speaker's soft biometrics (i.e. age [44], gender [40] and emotion [1]) and a vector representation of its identity-related features. The Speech Recognition [41] node, on the other hand, is responsible for transcribing speech into text in order to allow a spoken natural language interaction with the social robot. The VAD also activates the *Sound Source Localization* node [36] [39] in order to identify the position (w.r.t. the social robot) of the source of the detected sound. In this way, it is possible for the robot to understand the position of the interlocutor even if they are not in the camera's field of view. Once high-level information has been extracted from the two modalities in an independent manner, they are aggregated in the *Multimodal Sensors Aggregation* node. The aggregation allows to get a representation of the scene state

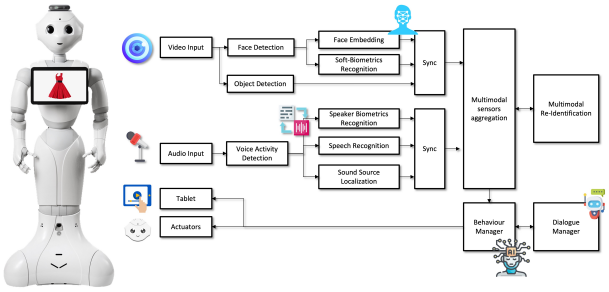

Figure 2: MIVIABot architecture. Data acquired by visual and audio analysis modules are combined and exploited in order to personalize the behavior of the robot with respect to the person interacting with it.

which is independent of the way in which it has been acquired. For instance, the robot can identify the interlocutor position with respect to himself exploiting either the Direction of Arrival from the Sound Source Localization node or the Face Detection results. Furthermore, the redundant information acquired from multiple sensors are aggregated to obtain a more robust awareness of the environment.

Finally, the data extracted by the audio and video pipelines are used by the *Behavior Manager* node, based on a finite state automaton, which is responsible for starting and keeping the engagement with the interlocutor. This node also relies on the *Dialogue Manager* node for understanding person' intents through Natural Language Processing algorithms and for retrieving the related natural language response. The Dialogue Manager node is based on Albert transformer model, a light version of the well known BERT architecture, optimized in terms of number of parameters and memory requirements. The Behavior Manager node is also in charge of the contents management on the tablet: a set of rules is previously defined, and the proper content is shown on the tablet depending on the person interacting with the robot.

4.3 Digital Recommendation System: Totem

The personal shopping assistant application is the core of the I-Mall digital signage. It exploits open-world re-identification and person detection and tracking in the mall to detect the items of interest of each individual. Combined with the analysis of personal traits, both clothing and behaviors, it supports decision on the appropriate personalized content to display on the digital signage when the person is in front of the terminal. The interactivity of the person at the digital signage terminal and the analysis of his clothing and emotional status drive the display of content that will eventually fit the person's interests. On this basis the application suggests the appropriate products considering item characteristics, style, color. Neighborhood-based and matrix factorization-based collaborative filtering have been used in order to compute recommendations. In fact, the outcomes of open-world re-identification and person detection and tracking, extraction of personal traits and

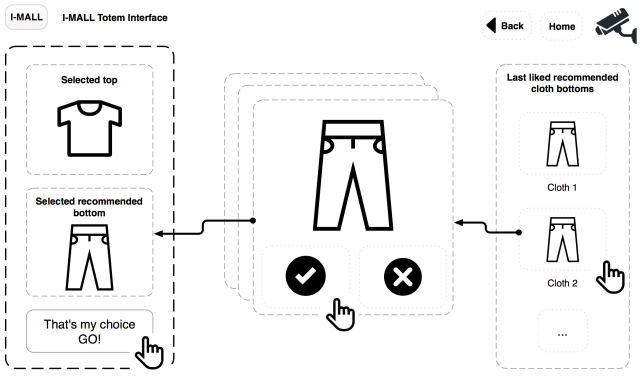

Figure 3: The digital signage interface functional mockup. Bottom garment recommendation is displayed. Recommendations are provided taking into account user's likes as well as user's detected clothing style and emotional status exploiting iterative recommendation with reinforcement learning.

feelings make digital signage contents both personalized and dynamic, namely variable depending on the time at which the digital signage is observed during the user stay in the mall.

4.3.1 Digital signage totem interface. A digital signage system installed in the mall allows the mall visitor to take advantage of personalized clothes recommendations. These recommendations are based on the users's behavior detected by the I-MALL system during his or her stay in the mall as well as on the user behavior analysis when in front of the totem. The digital signage displays a multimedia application that can re-identify, through a dedicated camera, the user and retrieve the profiling data stored by the system in its knowledge-base. A special location, outlined on the floor, is made available to the user that can take place in front of the digital signage in order to be re-identified. The I-MALL system re-identifies the user returning his/her demographic information (i.e. gender and age range). Additional information is provided by the clothing analysis module regarding his/her estimated preferred clothing style. The analysis is performed on the clothes worn at the moment by the user. Furthermore his/her emotional status is estimated by the emotional status module. The application then shows a gallery of the user's inferred favorite clothing items balancing all these information. Specifically, this first recommendation is based on the user's behavior analysis (i.e. clothing and emotional status) and estimated observation time of individual garments during his/her visit. The suggested items can be either the items actually observed or items similar in terms of garment characteristics. The suggested items are sorted by an estimated preference index [13, 14, 16]. Once the user selects an item of interest he or she is redirected to the top-bottom garment recommendation and presentation interface. If the user has selected a top garment he or she will find the selected item on the left and in the center a set of bottom garment recommendations presented in a stacked container (see Fig. 3). Iterative collaborative filtering is used to refine this top-bottom recommendation through a reinforcement learning model which exploits information on the user's detected clothing style as well as his/her emotional status while interacting with the interface. Once the user finds a satisfactory top-bottom clothes combination he or she is provided with routing directions inside the mall in order to eventually purchase the items of interest.

5 SYSTEM ARCHITECTURE

In Fig. 4 an high-level overview of the system architecture is shown. The system is composed by several modules communicating with the knowledge-base through an API gateway. Communication and messaging between modules can be bi-directional. A service-based approach has been followed, agnostic with respect to languages and protocols. The I-MALL Api gateway is the layer providing the basic

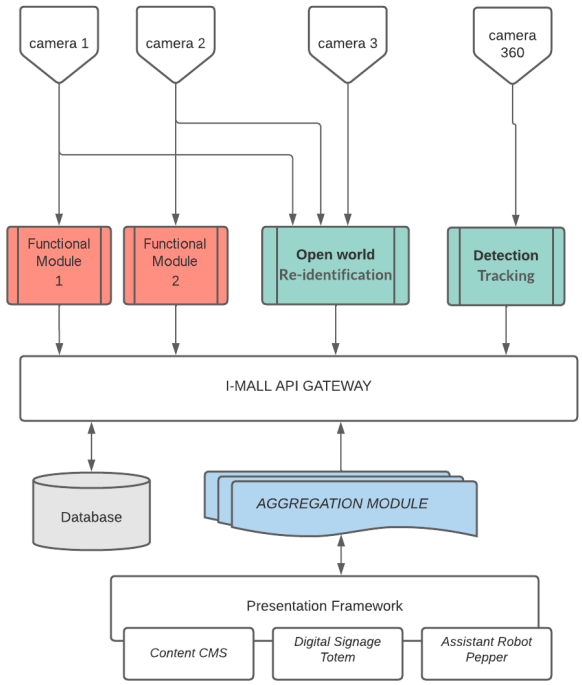

Figure 4: I-MALL system modules architecture.

services for communication between the various modules of all the overall system. It is exploited to acquire/send data between the application modules and to interact with the Aggregation module and the Presentation Framework. A protocol-driven approach is followed, allowing loose coupling between system components. The I-MALL Api gateway takes care of the successful exchange of messages between the several services, without them having to worry about routing and guaranteeing delivery. The transport protocol is HTTP/REST, with JSON as the data format. Thanks to this architecture design all the modules in the system work independently. Other modules can be easily plugged-in as long as they are compliant to data input specifications (JSON/REST API). Deployment of the I-MALL system has been done with Docker[1] technology exploiting images and containers.

[1]https://www.docker.com/

Table 1: I-MALL system evaluation: person re-identified

Re-identification task	Matches	Number of people	Accuracy
subjects involved	9	10	90%

5.1 Aggregation Module

In the context of the I-MALL architecture a strategic role is played by the Aggregation Module, which is responsible for the fusion of data coming from the computations of all the others functional modules. This fusion is done on the basis of the analysis performed by the re-identification module that, given descriptors of an image face crop as input, returns back the identifier of the customer previously enrolled in the system (see Fig. 5). Exploiting this identifier the user Extended Profile and the history of his/her visit is updated.

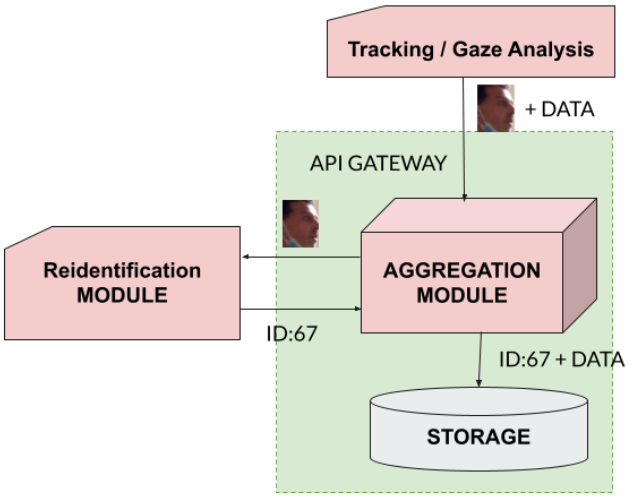

Figure 5: Aggregation Module functional schema

6 TEST AND DEMO

Since the beginning of the project it was planned a system evaluation at the VeronaFiere Centre with an a real setup that should have been done during one of the facility scheduled events. Arrangements had been made and a letter of agreement for authorisation was attached to the project documentation. However, the situation arisen as a result of the COVID-19 epidemic led to a delay of the activities of the original plan for the impossibility of testing the project at public events. Because of that, a new setup was created at the University of Palermo in order to reproduce a real situation as regard to spaces, installed cameras and the behavior of users in a mall facility. The I-MALL test setup involved the installation of the system in an area of 8.5m x 4m. As shown in Fig. 6 the setup was composed of: 3 shops (N), 1 totem, 1 robot (Rob), 4 cameras (3 CN + 1 CT), 1 camera 360 (C360).

Below is a summary of the data from the test session, which lasted 31 minutes and involved 10 people. The system was tested by trying to re-create partial or total occlusion in order to verify its robustness. This was done considering the fact that occlusions could

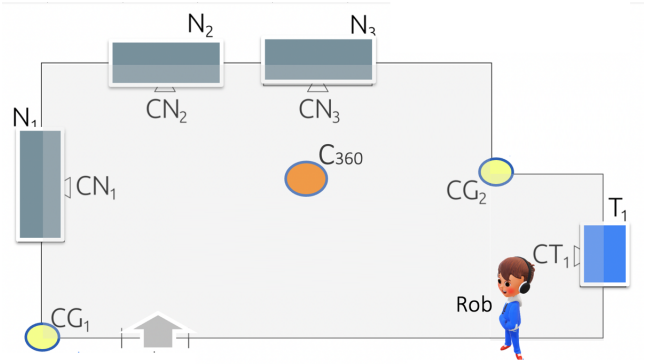

Figure 6: Demo setup.

Table 2: I-MALL system evaluation: matches vs. detections

Re-identification task	Matches	Detections	Accuracy
Total	**569**	**574**	**99,12%**
CN1 cam	79	79	100%
CN2 cam	277	282	98,23%
Totem cam	75	75	100%
360 cam	69	69	100%

likely happen frequently in a crowded mall. The re-identification task was 90% accurate with regard to the re-identification of the subjects involved in the simulation (See Tab. 1). In the only case in which the system failed this was due to some large occlusions in the acquisition phase. The accuracy of the re-identification of all the face detections captured by the system was 99, 12%. Re-identification errors occurred at CN2 cam location due, also in this case, to face person occlusions in the acquired images (See Tab. 2).

7 PRIVACY COMPLIANCE GDPR

The I-MALL project is conceived in the European Union, under the Regulation n. 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to processing of personal data and free movement of such data, and Directive 95/46/EC General Data Protection Regulation (GDPR). These are the regulatory sources that ruled the design and processing of information as regard to all the system functional modules, having as goals the minimization of the processing and the pseudonymization of personal data, transparency with regard to the functions and processing of personal data, security. According to the above principles perspective, I-MALL functionalities have been designed to be compliant with the regulations for privacy, so that the application is deployable in a real scenario. In particular: (i) Face and body of individuals once detected are immediately processed and encoded into numerical (non-decodable) descriptors. Such descriptors are aggregated into customers' identities that are identified by progressive anonymized numbers. Locations visited, interests, behaviors and feelings are associated to such anonymized identities. The appearance of the individual is not memorized at any stage of processing. (ii) All personal data collected create a temporary record of the visit in the mall. They will be maintained

just to provide personalized assistance and give appropriate feedbacks at the digital signage terminals. They will be erased from the system as soon the person leaves the mall.

8 CONCLUSIONS

In this article we have presented a system of re-identification of people in the context of a crowded environment such as a shopping mall. Visitors are identified, tracked and re-identified through computer vision at different locations. The system builds a profile of interest for each user based on the analysis of his/her behavior which includes the analysis of the face, of the customer emotional state, the direction of the gaze, the style in which he/she is dressed. This information is used to improve recommendation systems by context-awareness and human centered design through personalisation. The Shopping Mall context is just an exemplar context. We believe that the results of the project are general enough to find application in many different real contexts beyond that used for the experiments. The trend towards environment sensorization and personalized on-demand services will even more enhance their relevance for the future.

9 ACKNOWLEDGMENT

This work was partially supported by the Italian MIUR within PRIN 2017, Project Grant 20172BH297: I-MALL - improving the customer experience in stores by intelligent computer vision.

REFERENCES

[1] Hadhami Aouani and Yassine Ben Ayed. 2020. Speech emotion recognition with deep learning. *Procedia Computer Science* 176 (2020), 251–260.
[2] Jun Bao, Buyu Liu, and Jun Yu. 2022. ESCNet: Gaze Target Detection With the Understanding of 3D Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14126–14135.
[3] Federico Becattini, Xuemeng Song, Claudio Baecchi, Shi-Ting Fang, Claudio Ferrari, Liqiang Nie, and Alberto Del Bimbo. 2021. PLM-IPE: A Pixel-Landmark Mutual Enhanced Framework for Implicit Preference Estimation. In *ACM Multimedia Asia*. 1–5.
[4] Ilyes Bendjoudi, Frederic Vanderhaegen, Denis Hamad, and Fadi Dornaika. 2021. Multi-label, multi-task CNN approach for context-based emotion recognition. *Information Fusion* 76 (2021), 422–428.
[5] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*. IEEE, 3464–3468.
[6] Wolmer Bigi, Claudio Baecchi, and Alberto Del Bimbo. 2020. Automatic interest recognition from posture and behaviour. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2472–2480.
[7] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* (2020).
[8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.
[9] Marisa Carrasco. 2011. Visual attention: The past 25 years. *Vision research* 51, 13 (2011), 1484–1525.
[10] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. 2020. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5396–5406.
[11] Michael Crawshaw. 2020. Multi-Task Learning with Deep Neural Networks: A Survey. *arXiv preprint arXiv:2009.09796* (2020).
[12] Giovanni Cuffaro, Federico Becattini, Claudio Baecchi, Lorenzo Seidenari, and Alberto Del Bimbo. 2016. Segmentation free object discovery in video. In *European Conference on Computer Vision*. Springer, 25–31.
[13] Lavinia De Divitiis, Federico Becattini, Claudio Baecchi, and Alberto Del Bimbo. 2022. Disentangling Features for Fashion Recommendation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2022).
[14] Lavinia De Divitiis, Federico Becattini, Claudio Baecchi, and Alberto Del Bimbo. 2021. Style-Based Outfit Recommendation. In *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE, 1–4.
[15] Rosario Di Lascio, Pasquale Foggia, Gennaro Percannella, Alessia Saggese, and Mario Vento. 2013. A real time algorithm for people tracking using contextual reasoning. *Computer Vision and Image Understanding* 117, 8 (2013), 892–908.
[16] Lavinia De Divitiis, Federico Becattini, Claudio Baecchi, and Alberto Del Bimbo. 2021. Garment recommendation with memory augmented neural networks. In *International Conference on Pattern Recognition*. Springer, 282–295.
[17] Paul Ekman and Karl G Heider. 1988. The universality of a contempt expression: A replication. *Motivation and emotion* 12, 3 (1988), 303–308.
[18] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. 2021. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11390–11399.
[19] Pasquale Foggia, Antonio Greco, Alessia Saggese, and Mario Vento. 2022. Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognitio. *submitted to Engineering Applications of Artificial Intelligence* (2022).
[20] Marco Godi, Christian Joppi, Geri Skenderi, and Marco Cristani. 2022. MovingFashion: a Benchmark for the Video-to-Shop Challenge. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1678–1686.
[21] Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401* (2014).
[22] Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. 2022. A Modular Multimodal Architecture for Gaze Target Prediction: Application to Privacy-Sensitive Settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5041–5050.
[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[24] Zhengxi Hu, Dingye Yang, Shilei Cheng, Lei Zhou, Shichao Wu, and Jingtai Liu. 2022. We Know Where They Are Looking at From the RGB-D Camera: Gaze Following in 3D. *IEEE Transactions on Instrumentation and Measurement* 71 (2022), 1–14.
[25] Juntae Kim and Minsoo Hahn. 2018. Voice activity detection using an adaptive context attention model. *IEEE Signal Processing Letters* 25, 8 (2018), 1181–1185.
[26] Yunhao Li, Wei Shen, Zhongpai Gao, Yucheng Zhu, Guangtao Zhai, and Guodong Guo. 2021. Looking here or there? gaze following in 360-degree images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3742–3751.
[27] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1096–1104.
[28] Liliana Lo Presti, Giuseppe Mazzola, Guido Averna, Edoardo Ardizzone, and Marco La Cascia. 2022. Depth-Aware Multi-object Tracking in Spherical Videos. In *International Conference on Image Analysis and Processing*. Springer, 362–374.
[29] Giuseppe Mazzola, Liliana Lo Presti, Edoardo Ardizzone, and Marco La Cascia. 2021. A Dataset of Annotated Omnidirectional Videos for Distancing Applications. *Journal of Imaging* 7, 8 (2021), 158.
[30] Zuheng Ming, Joseph Chazalon, Muhammad Muzzamil Luqman, Muriel Visani, and Jean-Christophe Burie. 2018. FaceLiveNet: End-to-end networks combining face verification with interactive facial expression-based liveness detection. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 3507–3512.
[31] Zuheng Ming, Junshi Xia, Muhammad Muzzamil Luqman, Jean-Christophe Burie, and Kaixing Zhao. 2019. Dynamic multi-task learning for face recognition with facial expression. *arXiv preprint arXiv:1911.03281* (2019).
[32] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[33] Federico Pernici, Federico Bartoli, Matteo Bruni, and Alberto Del Bimbo. 2018. Memory based online learning of deep representations from video streams. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2324–2334.
[34] Michael I Posner, Charles R Snyder, and Brian J Davidson. 1980. Attention and the detection of signals. *Journal of experimental psychology: General* 109, 2 (1980), 160.
[35] Delong Qi, Weijun Tan, Qi Yao, and Jingfeng Liu. 2021. YOLO5Face: why reinventing a face detector. *arXiv preprint arXiv:2105.12931* (2021).
[36] Caleb Rascon, Ivan Meza, Gibran Fuentes, Lisset Salinas, and Luis A Pineda. 2015. Integration of the multi-DOA estimation functionality to human-robot interaction. *International Journal of Advanced Robotic Systems* 12, 2 (2015), 8.
[37] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking?. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc.
[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[39] Alessia Saggese, Nicola Strisciuglio, Mario Vento, and Nicolai Petkov. 2017. A real-time system for audio source localization with cheap sensor device. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–7.

[40] Héctor A Sánchez-Hevia, Roberto Gil-Pita, Manuel Utrilla-Manso, and Manuel Rosa-Zurera. 2022. Age group classification and gender recognition from speech with temporal convolutional neural networks. *Multimedia Tools and Applications* (2022), 1–18.

[41] Rainer Stiefelhagen, Christian Fugen, R Gieselmann, Hartwig Holzapfel, Kai Nickel, and Alex Waibel. 2004. Natural human-robot interaction using speech, head pose and gestures. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, Vol. 3. IEEE, 2422–2427.

[42] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[43] Zhihong Sun, Jun Chen, Liang Chao, Weijian Ruan, and Mithun Mukherjee. 2020. A survey of multiple pedestrian tracking based on tracking-by-detection framework. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 5 (2020), 1819–1833.

[44] Naohiro Tawara, Atsunori Ogawa, Yuki Kitagishi, and Hosana Kamiyama. 2021. Age-VOX-Celeb: Multi-Modal Corpus for Facial and Speech Estimation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6963–6967.

[45] Binglu Wang, Tao Hu, Baoshan Li, Xiaojuan Chen, and Zhijie Zhang. 2022. GaTector: A Unified Framework for Gaze Object Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19588–19597.

[46] Nicolai Wojke and Alex Bewley. 2018. Deep Cosine Metric Learning for Person Re-identification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 748–756. https://doi.org/10.1109/WACV.2018.00087

[47] Wenwen Zhang, Kunfeng Wang, Yutong Wang, Lan Yan, and Fei-Yue Wang. 2021. A loss-balanced multi-task model for simultaneous detection and segmentation. *Neurocomputing* 428 (2021), 65–78.

[48] Weiwei Zhang, Guang Yang, Nan Zhang, Lei Xu, Xiaoqing Wang, Yanping Zhang, Heye Zhang, Javier Del Ser, and Victor Hugo C de Albuquerque. 2021. Multi-task learning with multi-view weighted fusion attention for artery-specific calcification analysis. *Information Fusion* 71 (2021), 64–76.