A Multi-modal Audience Engagement Measurement System

Miguel Sanz-Narrillos(✉), Stefano Masneri(✉), and Mikel Zorrilla(✉)

Vicomtech Foundation, Basque Research and Technology Alliance,
San Sebastian, Spain
{msanz,smasneri,mzorrilla}@vicomtech.org

Abstract. During live events the organizers often would want to deploy audience engagement systems to analyze people behaviour, perform user profiling or modify the show according to the participant feedback. Such systems usually need computer vision algorithms whose performance are severely affected by constraints such as illumination and cameras position. In this paper we present a fully automatic audience engagement system, optimized for live music events with rapidly changing illumination conditions. The system uses a multi-modal approach which combines wireless-based person detection together with computer vision algorithms for pose and face analysis. We show that such hybrid approach, while running in real-time, performs better than standard approaches that only employ computer vision techniques. The system has been tested both in a laboratory environment as well as in a concert hall and it will be deployed in distributed live events.

Keywords: Multi-modal analysis · Wireless detection · Computer vision · Audience engagement

1 Introduction

The rapid evolution of broadband and streaming technologies in the last few years has had a profound impact on the entertainment industry, which has seen a strong increase in revenues coming from internet deployed services. Such implementations have given the opportunity to reach even more customers [11], and increase exponentially the data collected from the customers, using these services to boost the user engagement. The data collected is used not only to increase the revenue through direct advertising [16], but also to understand users' behaviour, feedback and preferences to improve the general experience.

The development of new technologies is obviously a key component for industry growth, but especially in the entertainment sector this is a crucial factor as it has allowed to reach wider audiences and to make shows more appealing, interactive and enjoyable. It wasn't until very recently, for example, that live events could also be experience by an online audience in an active way.

In online events such as e-sports the organizers collect a huge amount of data from the audience, such as the number of people connected at every second, location data, or the average amount of time watching the streams.

A. P. Rocha et al. (Eds.): ICAART 2020, LNAI 12613, pp. 354–377, 2021.
https://doi.org/10.1007/978-3-030-71158-0_17

This kind of data is currently not collected in traditional live events [32]. People in charge of organizing live events (concerts, festivals, sport events) are becoming more and more interested in obtaining users' data, due to its intrinsic value and the insights that can be extracted from it, but in this case in the real world, as the data obtained can be even more important.

In this paper, we describe a passive, non-invasive system for measuring audience engagement in live events. The data collected by the system allows us to identify when and where the users are the most involved in the show, as well as what are they most interested in. As the metrics normally used to measure engagement (as movement detection, person recognition, gaze detection and emotion detection) require to accurately detect and track the different body parts of the people attending the event, the system developed include methods for performing these tasks in the challenging illumination conditions typical of live events, where off-the-shelf implementations fail.

Standard techniques for person and body keypoint detection usually feed the camera stream to a convolutional neural network (CNN) [37]. CNNs are very powerful but they also require a lot of training data to provide meaningful results. Collecting and labeling data is usually expensive and time-consuming but it is often necessary, as the detection accuracy of CNNs typically plummets when they are used to analyse data coming from different distributions.

A relevant example where standard CNN based techniques do not return good results is in live events. CNN architectures trained on existing datasets are able to accurately detect faces position or the pose and movements of the people, but in settings with constantly changing illumination conditions, such as concerts, such systems perform very poorly. A system relying only on video data would require extensive fine-tuning, performing training on additional data with abrupt changes in illumination. A further issue with such approaches is that most of the times different regions of the image to analyze contain both bright and dark spots, making it very hard for the CNN to properly generalize and detect all the people in the scene.

A more sensible way to deal with those issues is to preprocess the images captured by the cameras in order to eliminate or at least reduce the factors affecting the performances of the CNN. The preprocessing step varies depending on the input and thus it requires information from other sources, for example localization data estimated from the wireless signals emitted by the users' smartphones.

The main contribution of this paper is the description of a hybrid system which uses both computer vision and wireless signal analysis techniques for detection and tracking of people in live events and, from that, derives audience engagement measures. The use of a hybrid approach, apart from providing more user information, allows higher detection and tracking accuracy than using the two methods separately. Furthermore, the system is robust to sudden illumination changes and noisy environments without requiring additional training, making it useful for a wide range of applications.

This works improves the system described in [42] and describes in detail the development and deployment process of a complete solution for obtaining the engagement in live shows.

The code of the system and the data used during the experiments are available on Github[1]. The rest of the paper is organized as follows, Sect. 2 will cover the related work with the two techniques used for the hybridization, engagement and multi-modal systems, in Sect. 3 a comparison of the capabilities and possible obtainable metrics between different systems that perform engagement detection is done, in Sect. 4 a explanation of each method and the hybridization characteristics is explained, including the improvements and extra functionalities added, and in Sect. 5 a comparison between the results of the hybrid method, the computer vision method and the computer vision with the improvements implemented.

2 Related Work

2.1 Vision-Based Human Analysis

The detection of people in still images and video has long been one of the most studied problems in computer vision. Prior to the advent of deep learning based techniques, the standard approach was to create a human model using image keypoints and descriptors, for example Haar cascades methods [26], Support Vector Machines [5,29] or Histogram of oriented gradients [9]. In recent years, thanks to the availability of datasets such as ImageNet [12] or Microsoft COCO (Common Objects in COntext) [27] and the increase of computational CPU and GPU power, convolutional neural networks became the standard tool used for objects detection and tracking. The architectures most commonly used for this task are R-CNN and its evolutions [13,14,40], You Only Look Once (YOLO) [38,39] or Single Shot multibox Detector (SSD) [28]. More advanced architectures can provide a pixel-level segmentation of the person detected [19], while others detect the position of the joints in order to estimate the person pose [6,8,46,47]. Such algorithms rely on datasets specifically created for the task such as MPII Human Pose [2] and Leeds Sports Pose [21].

2.2 Wireless-Based Human Analysis

The standard approach for detecting and tracking people using wireless signals is to rely on the Wi-Fi and Bluetooth signals provided by a smartphone or other wireless capable devices carried by the user. One of the possible approaches relies on Received Signal Strength Indicator (RSSI) fingerprinting [49], where the communication signal strength is used to determine the distance of the device from the receptor. In order to obtain a reliable position trilateration must be used, combining the data from several receptors [33]. Other approaches rely on wireless time of flight [25], which uses the time between the emission and reception to determine the distance between the devices and from that infer the persons position. Another technique is the wireless angle of arrival [17,35], where an antenna array measures the angle of arrival of the signal instead of the ToF. In this case the angle from the device to the receptor is calculated by having an antenna array as receptor and with the difference on the reception time between each of the antennas the angle of the signal can be calculated, and with trilateration the position

[1] Indoor person localization hybrid system in live events https://bit.ly/3cYmvz2.

can be approximated. A technique that does not need the person to carry a device is the ones used in WI-SEE and WI-VI [24], where the shape of objects in the room is computed by analyzing the reflection of the Wi-Fi waves, and uses those to detect the position of the persons.

2.3 Audience Engagement Systems

As mentioned in Sect. 1 most of the engagement systems are designed for online events because in those cases the infrastructure necessary is already available. Systems for online learning [23,31], social media [43] or news [4] already implement tools for measuring user engagement. In the case of live events the infrastructure and the system have to be built separately, although some interactions can be created with electronic devices such as lights or screens. Most current engagement systems depends on the usage of an external device to provide the information about the engagement. One example of engagement system is the glisser app [15], in which the event manager can implement questionnaires, slide sharing or a Twitter wall. In this case only the information that the person writes in the app is considered as engagement. Another approach to have a more truthful information has been the usage of electroencephalograms to measure the signals produced in the brain as in the engageMeter [18]. Such systems are not very suitable to be used in events such as concerts where multiple people are moving and user engagement has to be measured in an indirect way.

2.4 Multi-modal Systems

The usage of different techniques and methods together has been used for many years in the development of new systems to improve the final results. In the detection field this type of systems has been used in recent years for autonomous vehicles [3], combining a CNN and Lidar, person detection systems [45], which uses laser and camera data, and some datasets has been created for this type of systems such as a fall detection [30], which combines information from video and wearable sensors.

3 Engagement Measurement Systems

As the system is meant to be installed in the entertainment industry in order to measure the engagement of the people to a determined show, we are going to compare theoretically all the methods that could have useful metrics related to the engagement as number of people in the room, movements of the people or emotions. The analysis is portrayed in order to chose the methods to take part in the hybridization and it will be taken into account the accuracy, performance, hardware used and possibility to be used at distance and with crowds.

CNN. There are several networks that perform this type of detection but the technique that gives better results is the ones using pose detection, in which it detects several joints as in some cases some face points in order to locate a person, this techniques is a good candidate to be one of the bases of the hybrid system that is explained and from which results are analysed in this paper. This technique has good result when the conditions of the environment are previously known and are included in the training data, in other case the result is not as good as expected. The computing power out of the box of this type of techniques is quite high, having several GPUs to be able to perform the detection on 720p footage at real time. If we take into account the improvements included in the technique described in this paper the main problem is the conditions, that if one is not present in the preprocessing data it will not be good, so the preprocessing would need to be dependable on the conditions, reason that our system has the hybrid method. The main metrics obtained directly when this system is working in good conditions are the position of the different joints, which from that several variables such as movement or person localization can be obtained.

WiFi/Bluetooth Localization. In this case the localization of the person has huge dependency on carrying a device, as this is the thing located, and the performance is only average, having only an approximate position. Another drawback of this system is the number of devices that are detected, reason to use the filter, but the main drawback is the time between probes, that depending on the device to be tracked can change from twenty seconds to several minutes, as well as the necessity of having one of those signals activated in the devices. This method allow us to obtain the approximate localization of a device, that is normally attached to a person, so with this the number of people in a determined zone can be approximated, although the accuracy will decrease as the area of the zone decrease. The main metric obtained is the number of people inside a zone, so it could be used in cope with the CNN in order to refine the results and lower the number of false detections and people no detected.

RFID. This technique is more a barrier detector than a locator, which means that it detects when a person or an object goes through some point but it can not detect the number of people at some distance. It also requires special hardware to be in both the detected and detector devices. This technique is more suitable to maintain an stock list rather than a person locator. The main metric obtained from this method is the number of people that has crossed an invisible barrier, data that is less accurate than any of the previous methods and this even needs more specialised hardware, being in both the tracker and the detected person, than in the case of the CNN is not necessary and in the wireless is so common that is virtually not necessary.

Accelerometers. This technique can be used to measure the movement and it can perform well detecting movement and direction of movement, but the main problem with this technique is the calibration as to have reliable lectures of position we need to have an initial point. It requires to have specialized hardware in the person to be detected. The main metrics obtained from this method are in reverse as in the CNN, while in this we obtain the movement directly and the position can be computed from that data, in the CNN we obtain the position and with the historic data the movement can be computed, the main difference from those technique is the necessity of specialised hardware in the detected person to perform the detection, than also needs to be calibrated to obtain a reliable detection.

Emotion Detection by Camera. This technique make use of the detection of facial landmarks, as eye borders or lips borders, as the detection of other facial points, such as eyes or ears, this will mean that at least 100 points in the face are obtained, all this points are introduced in a neural network that has been trained with faces related to several emotions from several people, in order to reduce the bias of only having few people data, the results of this neural network are two variables, the valence and arousal, depending on the values of this two variables an emotion between the six primal ones (Happiness, sadness, fear, disgust, anger and surprise) can be obtained, the relation of the arousal and valence with the primal emotions is done by using different models as Bayesian networks, Gaussian models or Markov models and with that the emotion can be obtained. Nowadays this method can not be used with a high number of people as it needs both huge computing power and a good view of the faces to perform well. Another problem is the reliability of the data as the primal emotions are very different in each person, as well as needing to have very exaggerated expressions to be able to read the emotion. Although the metrics of this method is very unique and very related to the engagement, but the problems to the implementation surpass the increase in accuracy, the improvements of this technology has to be taken into account but nowadays the implementation in this system is not worth it.

Emotion Detection by Body Signals. This method is very related to the previous one as it is also based on obtaining the emotion of the people out of one of the primal ones from the valence and arousal, which is done the same way, the difference between this methods is the way that the variables are obtained, while in the previous one was done reading the facial points with a camera, in this case some of the body signals are measured, which normally several are taken into account to have a higher accuracy. Some of the signals that could be taken into account are the temperature, hearth rate or breath rate, in this case the reliability is greater, and technically could analyse several people at the same time, but the main drawback is the hardware that is needed, that in most of the cases needs to be in direct contact with the body. As before the metrics of this method is very unique and very related to the engagement, but the problems to the implementation surpass the increase in accuracy.

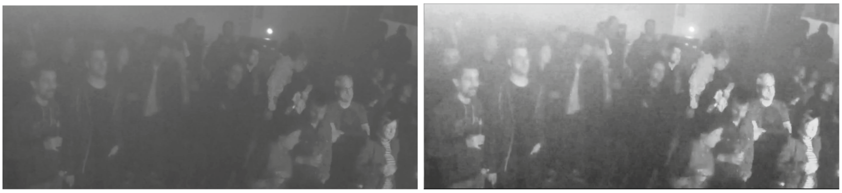
Our Hybrid Approach. As the analysis has shown the methods to be implemented in the hybrid approach are the CNN, as this is the most used technique to locate people and the metrics obtained are numerous, and the wireless detection, as the reliability is high with the localization metric. In order to obtain other metrics some additions could be done, as obtaining the gaze direction of a person, but the main additions to be done to each of the techniques is to try to reduce the drawbacks explained in this section with the hybridization, as the lightning condition of the CNN or the number of devices in the wireless method, which will be explained in the Sect. 4.1 and 4.2, while the way both methods communicate to make a real hybrid method is placed in the Sect. 4.3.

4 Methods

4.1 CNN-Based Detection and Tracking

CNN architectures are the de facto standard for tasks like people detection and tracking. Usually, one or more cameras are used as input source for the CNN, while the output consists of a vector of bounding boxes describing the people position or, depending on the architecture, the position of specific joints. The reference technique used in this work is the one described in [34], a technique for person detection that detects the person position as well as that of different body parts such as face, shoulders or hips. This technique provides average accuracy in densely packed scenes (with more than 30 people in the same scene). The base implementation used in this project is based on [48] and in order to improve its detection accuracy we implemented new functionalities related to preprocessing of the input, tracking additional keypoints and improving the performance.

The addition of a preprocessing step is the contribution that improved the detection accuracy the most. Using this module, we were able to accurately detect people and their body parts without fine-tuning the architecture using a dedicated training set. The preprocessing is done in three steps. First, the input frame is sliced into several rectangular area. Each area is then processed separately (and in parallel) to the others. Then, the contrast of each area is improved by applying the CLAHE (Contrast Limited Adaptive Histogram Equalization) algorithm, to make the borders of the objects more noticeable. Finally, gamma correction and normalization is applied to each image slice. This helps in cases where some slices are illuminated while others are not, as it is often the case during concerts or other live events. The results of preprocessing can be appreciated in Fig. 1, where while in the first part the illumination is on the low-right part of the frame the preprocessing intensifies the borders and regulate the light in the overall picture.



(a) Original Frame **(b)** After CLAHE and Gamma correction

Fig. 1. Preprocessing of input frames.

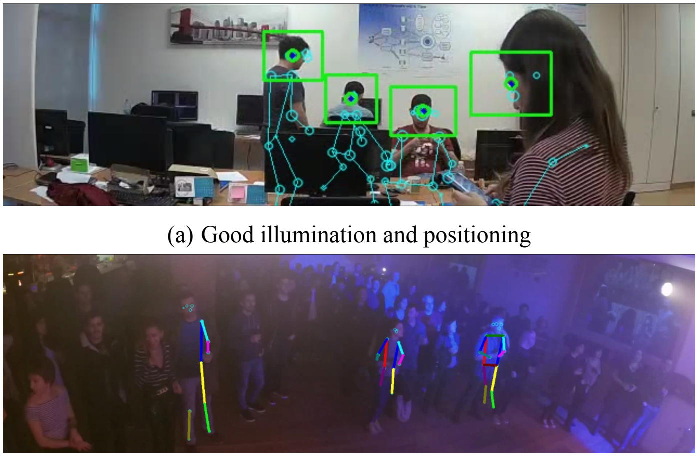
Apart from the position of the person and its keypoint, our system detects gaze and movement. The gaze detection tool infers approximately the gaze direction (front, right or left) from the position of the eyes and nose detected by the CNN. It is calculated analysing the angle formed by the segments connecting the two eyes and the nose with the midpoint between the eyes. Movement data, on the other hand, is computed by performing tracking of the different keypoints so that we could evaluate average and maximum speed of a person over time.

Changes in performance has been implemented to reduce the processing of all the parts that are nor necessary as reducing the number of joints to detect, in cases that is impossible to see them as the knees in crowded concerts. This project also allows to eliminate zones from the image that are impossible to contain people as ceilings.

The part that improves the performance more is the implementation of tracking which puts a tracker in the detected people, this allow the system to only perform the detection in one out of several frames instead of on everyone.

The dataset used for training this CNN is COCO, which contains more than 200K labeled images although not all of them are fully annotated. The images are usually taken with good to excellent illumination conditions and no preprocessing (except for the standard data augmentation procedures) was applied during training.

The implementation in [48] shows a steep decline in the accuracy of detections when the illumination conditions of the scene are not represented in the training set. An example of such performance decay can be seen in Fig. 2, where the detection is perfect in the upper figure, while in the bottom image a very small percentage of the people gets detected. This is caused by not having the network trained with all the possible illumination conditions, making it almost unusable without the aforementioned preprocessing step.



(a) Good illumination and positioning

(b) Illumination changing and strange positioning

Fig. 2. Pose and person detection under different illumination conditions [42].

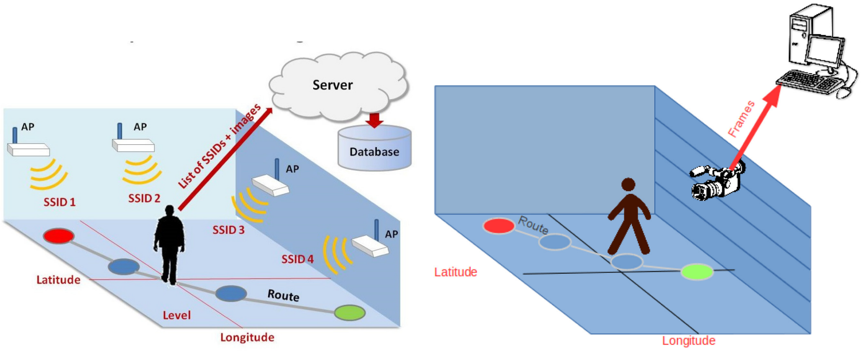
Comparing the two images in Fig. 2 we can easily appreciate that the main difference between them is the illumination conditions as the image on top has higher brightness and contrast than the one on the bottom. In audience monitoring applications (such as during live events) it is highly likely that the illumination conditions change over time, and often different parts of the scene have different brightness and contrast. In this case any person detection algorithm is doomed to fail unless the input frames are preprocessed so that they provide the same illumination conditions across the whole image as well as over time. Section 4.3 describes in detail how our implementation chooses the parameters used to preprocess the input frames before feeding them to the neural network.

4.2 Wireless Data

Nowadays almost every person carries at least one device capable of receiving wireless signals such as Wi-Fi or Bluetooth. These types of signals have already been used to perform device localization tasks [1,10,24] because they offer several advantages compared to computer vision based applications, as they require much less computation capability and they are not affected by problems like occlusions.

A device can be localized even though it is not connected to a network, the only requirement is that it is in the range of the router. This occurs because the devices periodically scan the environment to check for available networks, while at the same time the access points (APs) send broadcast messages to make their network discoverable. When a device detects that it is in the range of an available network, it exchanges with the AP the MAC addresses and the connection properties. By analysing such properties and the intensities of the signal received by the AP, we can estimate the distance between them.

Knowing the distance between the device and the detector is not enough to locate the device in a room, but with trilateration (using three or more detectors) the position of the object can be calculated. The process for the localization can be seen in Fig. 3a, where the persons position is approximated by four trackers (in our case the APs). Our implementation is based on Find3 [44], with some modifications allowing us to perform device filtering and to deal with devices performing MAC randomization.



(a) Wireless signals method

(b) Computer vision method

Fig. 3. Device positioning and communication diagram [42].

We modified the original implementation of Find3, adding two new functionalities to make its adoption more suitable for usage in live events settings. The first add-on is device filtering and it was developed to improve the performances of the system, while the second is zone differentiation, which allows us to know in which part of the room each person is located.

As we mentioned before, the number of wireless capable devices has dramatically increased in the last few years, so when an AP looks for available devices, it could detect not only mobile phones or smartwatches, but also other device types such as printers, TVs, smart balances etc. This is a problem as we are only interested in a device if it's being carried by a person, otherwise the estimation of the number of people would be skewed. For this reason we filter the list of available devices twice. First, by removing all the device that are not detected by all the APs, as we assume that such devices lie outside our region of interest (as we show in Fig. 4).

Then, we filter a second time according to the device manufacturer. This information is encoded in the MAC address of the device, since the Organization Unique Identifier (OUI) is a 24-bit identifier encoded in the first three octets of a MAC address. By using the company ID list provided on the IEEE website [20], we filter the results by excluding all the devices not associated with companies producing smartphones or smartwatches.

This filtering based on MAC address could pose an issue though, as in recent version of the Android and iOS operating systems the devices implement MAC randomization. For security reasons, before a connection to the network is established, the devices share with the AP a random MAC address, and communicate the real one only once the habdshake is completed and the connection to the network is established. In Apple devices running iOS 8 or above the MAC address is completely random and changes periodically, thus making MAC filtering uneffective. For devices running Android 10 (currently the most recent OS version), device filtering still works since the randomized MAC addresses are chosen from a known range bought by Google and available in the IEEE list.

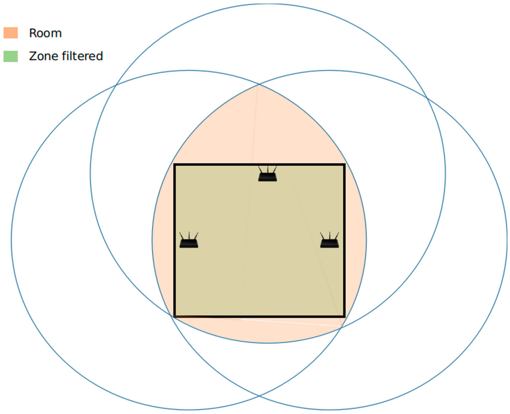


Fig. 4. Filter zone of interest.

We implemented the two filters both in the server and on the detectors. We first developed a Python version of the filters and then, to improve the performances, we also ported the code to Go. Table 1, summarizes the results obtained where, unsurprisingly, the fastest implementation of the filters is the one written in Go and running on the server. The results of the table are the average between 20 detections with an average of 35 MACs in the input and 10 when the filtering is done.

Table 1. Time (s) spent in the MAC comparison.

Server with python	Server with Go	Scanner with python	Scanner with Go
0.16454	0.03668	28.36	2.5

Another addition introduced over the original implementation is the possibility to classify the position of the devices into different areas (customizable by the end user). The classification can be performed in three different ways:

- Manual: this method is based on having the RSSI measured in the room by one device before the detection then with that having a map of the RSSI in the room, and compare the measurements of each device with that map to position each of the devices. This method has low computational requirements, but changes in the environment, such as moving obstacles or reflections, can severely affect its accuracy.
- Weighted: this method tries to improve the precision by using a group of measurements instead of only the most recent. We tried several weighting strategies and we found out that logarithmic weight averaging offered the best accuracy. This method improves the results but also the computing power needed.
- Automatic: the main focus of this method is to make the zone differentiation non dependable on the conditions of the room. This is achieved by using reference devices in selected locations of the room to obtain the measurements of known positions and compare them with the ones of unknown devices. Reference devices can be distinguished by their MAC address as long as they remain connected to the network. This method outperforms both the manual and weighted approach in terms of accuracy and it can even be combined with the weighted method for a further boost in accuracy.

4.3 Hybridization

Once the results from vision and wireless based detection systems are available, the hybridization step is responsible for processing and combining them in order to obtain a higher accuracy. The main idea is that the data provided by the wireless person detection system can represent a rough estimation of the number of people in the scene and, comparing it with the previous result from the vision system, it can steer the preprocessing of the frames to improve the subsequent vision-based detection and tracking results of the system.

Consider for example the image in Fig. 2. In this case the wireless detection system could estimate that there are more than 30 people in the range of the router, while the vision-based system only detects 3 people (due to poor illumination condition, varying image contrast in different region of the image, etc.). The work-flow of the hybridization is summed up in Fig. 5.

Apart from the detection and the tracking data from both systems, the hybridization system takes as input a function which maps the 3D regions in which the wireless detection splits the room and the 2D regions in the camera frames where the preprocessing will be applied.

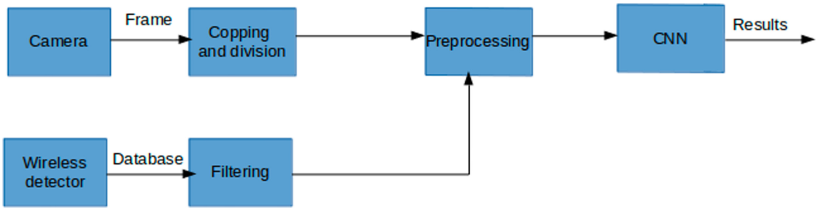


Fig. 5. Workflow for the hybrid system.

Preprocessing. The aim of the preprocessing is twofold, as it should both speed-up the detection times and modify the input images with the aim of maximizing the detection accuracy.

The preprocessing performed is composed by several steps. First, the image is cropped to remove the parts of the frame where no person could appear (see Fig. 6 for an example). The cropping process is performed manually, as it depends on the camera positioning, and it is a one-time operation which is then applied to every frame of the video. This improves the performance since there is a lower quantity of pixels to evaluate and the neural network is able to process more frames in a single pass. Then, the input frame is divided into different slices. Figure 7 shows an example where the input frame, after cropping, is split into six parts. Each slice will be then preprocessed separately by applying different brightness and contrast changes. In this way the system is able to cope with the fact that different parts of the frame may have different color

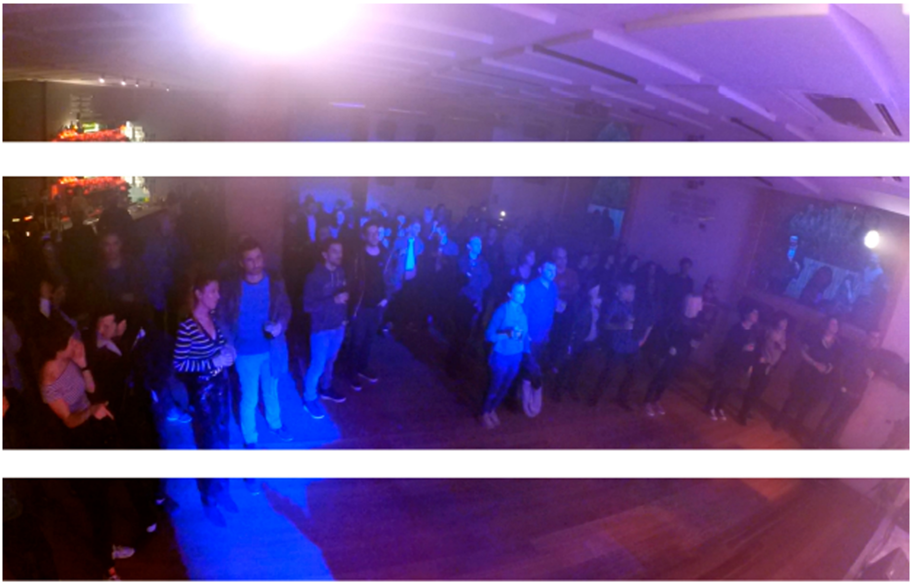


Fig. 6. Elimination of non-person parts of the frame [42].

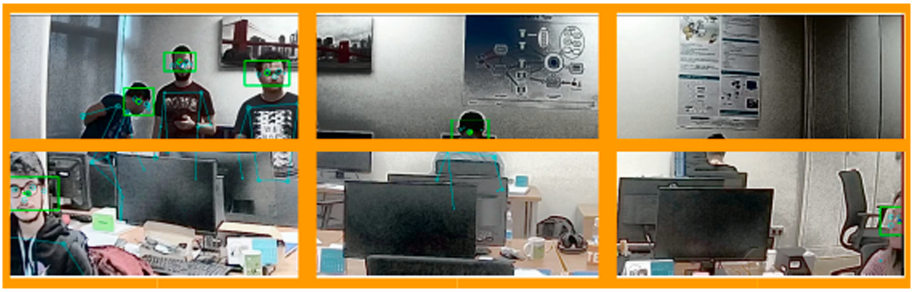


Fig. 7. Preprocessing slicing of the frame [42].

and brightness statistics. There is a 5% overlap between each slice (represented by the orange lines in Fig. 7) to counter the fact that people moving in the scene from one slice to the adjacent one may be lost when crossing from one slice to the other. The way the frame is split into different slices depends on the camera position as well as on the geometry of the regions identified by the wireless detection system.

The processing of each slice is done by applying contrast stretching using the CLAHE transformation [36], followed later by Gamma correction [41] to reduce or increase the number of bits of luminance and so dynamically increase or decrease the processing power needed. The parameters used for performing CLAHE and gamma correction are dynamically chosen by comparing the detection results of the wireless and vision-based system.

Tracking Strategies. In order to speed-up the processing times of the vision system, the detection step is performed once every 10 frames, while in the remaining frames people are only tracked using MedianFlow [22].

To avoid tracking false detections indefinitely, the tracking is periodically reset, while correct assignments keep being tracked by performing a simple nearest-neighbor assignment from previous frames. Figure 8 shows a visualization of the tracking of a person's face: the green rectangle shows the current position of the face, while a curve shows the path followed by the face center. The most recent positions (the latest 20 frames) are drawn in blue, while older positions are shown in green and, for positions older than 50 frames, in red.

The wireless detection system does not implement a tracking mechanism, but data from previous measurement is used to increase the robustness of the detection mechanisms. Previous measurements are exponentially weighted, with a higher weight associated to more recent measures.

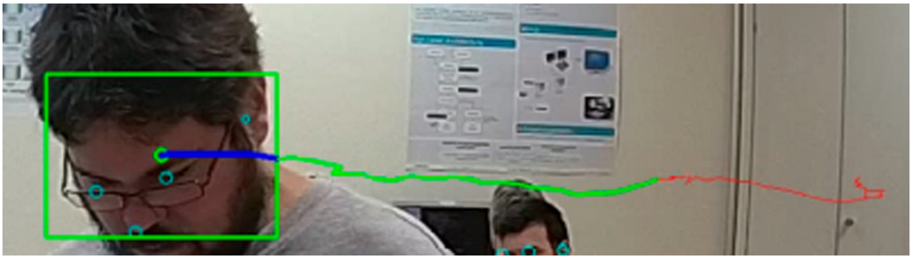


Fig. 8. Tracking path drawing in the frame [42]. (Color figure online)

Zone Relation. The wireless method divides the room in several zones, while the computer vision method divides the frame in several slices. In the tests we performed, we used three zones for the wireless system and six for the computer vision one. Before the processing starts, a function maps the zones from the camera to a zone in the 3D space. The mapping is not perfect but, as the precision of the wireless technique is in the range of centimeters, the relation does not need to be exact.

The number of zones in the wireless method depend on the accuracy needed and the conditions of the room, such as size and shape. It is possible to have a different number of zones and trackers: in our tests we used 2 trackers for classifying into 3 zones.

Depending on how the image is split, it may happen that if the person is very close to the camera, or the person does not wear his device, the computer vision system detects one person in one zone while the wireless method detects it in another. Some of that issues can be avoided with a good camera positioning, which is at a medium distance from the people and at a height of 2.5 m approximately. If the camera cannot be moved, the detection difference between the methods can be changed. This difference compares the total detections between the methods and in the case that is greater than a threshold the preprocessing conditions (gamma and contrast), are changed.

4.4 Engagement

The information collected from the combination of computer vision and wireless signals techniques can be used to measure the engagement of the people attending the live event. We are aware that the information available is only a proxy of the user engagement and that our measurements are not perfect. For this reason we are more interested in average user behaviour patterns and we decided to analyze the results also from a qualitative point of view. Below is a list of the parameters used to estimate the user engagement:

- Number of people in the room: This is the simplest metric we collect. It is computed by averaging the estimation from the vision and wireless based system, and gives a rough indication of how successful a specific event has been. It is useful when compared with data from previous events or for events happening in other rooms or venues.

- Number of people in each zone: As explained in Sect. 4.3, we can also estimate the number of people in specific areas. Analyzing the change of population per area over time (for example, the ratio of people close to the stage vs. people in the bar area) can provide information about what parts of the show were the most engaging.
- Movement patterns: since people can be tracked, we can analyze their movement patterns and study the dynamic of people movements in the venue over time.
- Movement of a person thought the entire session: This is related to the previous metric, but in this case we focus on a single person instead of checking group dynamics. This metric can provide detailed information about the engagement of the user, especially when comparing it with the information about the event (lists of songs played, performers etc.)
- Movement of the different limbs: This metric is a proxy indicator for how much a person has been dancing. In many cases when people are dancing we noticed a small movement across the venue but a high movements of the joints detected by the system.
- Direction where a person is looking to: Gaze information, especially its variation over time, is also very relevant for determining the user engagement. By analyzing when the people change their gaze direction towards the stage we could detect the most interesting moment of the event.

5 Results

We conducted fifteen tests in a controlled environment, changing the following variables:

Number of people on camera: controls the number of people that can be seen in the image retrieved from the camera. This variable can take the values from four to eleven in the test. It has been included to see if the system loses precision when increasing the number of people in the room.

Separation between the people: controls the distance between the people in the room. It is treated as a binary variable as people could be either close (distance is less than 30 cm) or separated (distance is greater than 70 cm). This variable has been included to see the impact of occlusions in the vision-based system and to measure the reliability of the tracking system.

Wi-Fi connection: controls if the mobile device of the people are connected to the same network as the scanning devices, allowing the system to know the real MAC address of the device and to retrieve more data from it.

Illumination: controls the state of the lights on the room, either turned on or changing over time. This variable has been included to see if both the preprocessing with segmentation and the hybrid approach can reduce the effect of the change of illumination in the computer vision techniques.

Number of people moving: controls the quantity of people moving from one zone to another. This variable is expressed in percentage of the total people in the image.

Table 2 shows the different conditions under which the fifteen tests were ran.

In order to simplify the testing and the further proving of results, we ran the test in offline mode, that is we first recorded the electromagnetic environment and the room

Table 2. Test variables [42].

Test	People	Separation	Wi-Fi	Lights	People moving
1	11	30 cm	✗	Turn on	3
2	11	30 cm	✗	Turn on	5
3	11	30 cm	✗	Changing	5
4	11	70 cm	✗	Changing	3
5	6	70 cm	✗	Turn on	2
6	11	30 cm	✓	Changing	5
7	11	30 cm	✓	Turn on	3
8	11	70 cm	✓	Changing	3
9	6	30 cm	✓	Changing	2
10	11	30 cm	✓	Turn on	5
11	11	70 cm	✓	Changing	7
12	11	70 cm	✓	Turn on	4
13	6	70 cm	✓	Changing	1
14	6	70 cm	✓	Turn on	1
15	11	30 cm	✓	Turn on	11

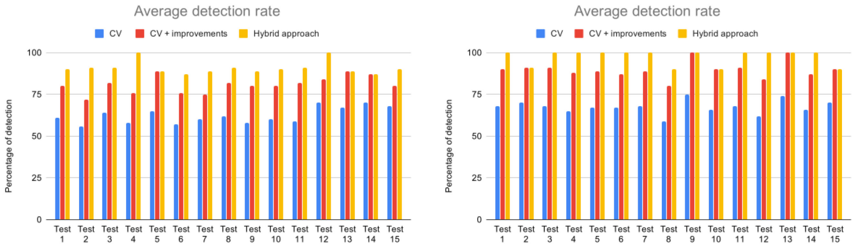
with the camera, and then later we processed the data. The video was taken in two modalities, a low-quality one (360p resolution, 10 fps and 400 kbps bitrate) and a high-quality one (1080p, 10 fps and 5 Mbps) to compare security camera quality to consumer grade cameras. Each test lasted five minutes, both for video and recording of the electromagnetic environment. As expected, using low quality videos the detection rate decreases, having more false detections and less people detected. Strangely, we noticed that double detections, person being detected two times in the same frame were more probable with the high quality video. This double detections happens when the system does not detect that two detected joints are from the same person and attributes them to different people, by supposing that the rest of the person is not detected because is being covered.

In the tests we compare, when possible, the out of the box computer vision algorithm without any of the improvements that has been exposed in this text, the computer vision algorithm with all the improvements that has been covered and the hybrid method with also the improvements exposed. In the case of analysing the tracking we only use the computer vision with the improvements, as this is the only one that implements the tracking mechanism.

The tests measured the following:

- True positive detections: measures the number of persons correctly detected at each frame. This variable is related to the maximum number of people that the system is able to track.

- Number of false detections: measures false detection at each frame. This variable will take into account both the false negatives (missing detections) and the false positives (detecting a person when it is not there, or detecting the same person twice).
- Tracking: This variable takes into account the movement of the people across different zones in the room and their location. This variable will measure if the system can track the movement of a person through the time.
- Processing time: This variable analyses the average time that is necessary for the processing of a frame in the video.



(a) Low-quality video

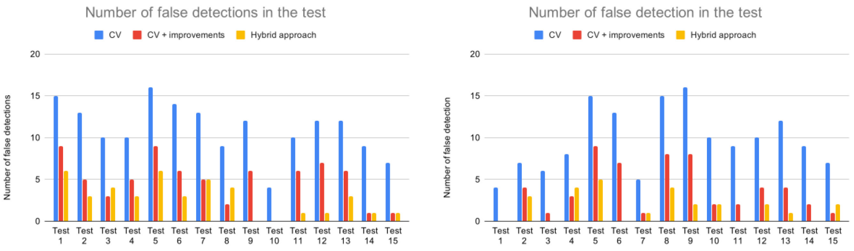
(b) High-quality video

Fig. 9. Average person detection rate.

In Fig. 9 we show the average (per frame) percentage of people detected on the videos in each of the tests, while in Fig. 10 we report the average number of false detections, both for the low and high bit-rate videos.

Figure 9a shows that, for the low-quality video, the hybrid approach in most cases performs better than the vision-only system (and in two cases correctly detects all the people in the scene), while in three cases it shows the same performance. Figure 10a shows a strong improvement in terms of false detections across almost every test, and no false detections at all in one case. In both cases it can be seen that the performance with computer vision is worse when the improvements and add-ons mentioned in this paper are not included, although the conditions of the test were not as changing as in the entertainment shows both detection rates, for low and high quality videos, has increased.

Similar conclusions can be drawn when analyzing the results on the high-quality video. Figure 9b shows that the hybrid system improves over the vision-only method and in 11 cases, reaching 100% detection rate. Figure 10b shows a similar trend: with the exceptions of tests #4 and #15 (where one of the participants is detected twice by the system), the false detections are lower when using the hybrid approach. In this case the difference of the computer vision before and after the improvements are noticeable, reducing the double detection in more than half in the majority of the cases and improving the detection rate by 10%.



(a) Low-quality video

(b) High-quality video

Fig. 10. Average false detections.

We measure the quality of the tracking using the *tracking length* metric [7]. In Fig. 11 we show the tracking length when using the hybrid approach (red) and the vision-only based system. The tracking length is fairly consistent across the different tests, and the results clearly show that the hybrid approach improves over the vision-only system, with an approximate gain of 25%. As the tracking is an addition included in the computer vision there is no data for the computer vision without improvements or add-ons.

Figure 12 and Fig. 13b show some examples of the difference in detection and tracking quality between the vision-only system and the hybrid one according to different metrics:

- Tracking performance - Fig. 12a shows that the movement of the person is recorded for much longer time when using the hybrid approach.
- Number of detections - Fig. 12b shows how the hybrid approach is able to detect more people and how the vision-only approach may fail, by detecting a group of people as a single person.
- False and double detections - Fig. 13a shows that the double detection of the person does not take place on the hybrid method.
- Body parts detections - Fig. 13b shows that, even if both methods fail to detect the person, the hybrid method is able to detect at least some body parts

We measured the difference in processing time between the hybrid system and the computer vision with and without improvements techniques. The results, displayed in the Table 3, show that the hybrid approach is marginally slower than the vision-only based method when it has the improvements added, but without them it is many times better, this is due to the tracking functionality, which reduces the number of complete frames that has to be analysed completely. Performance were measured on an Intel i5 PC with 16 GB of RAM and a Nvidia 1080 GPU, taking the average over 20 runs.

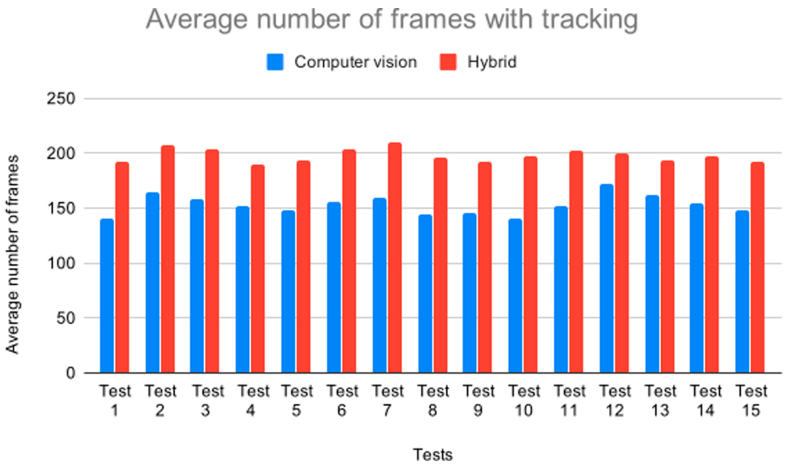
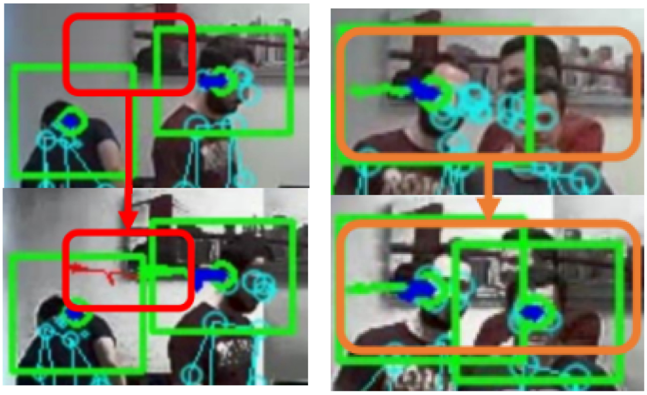


Fig. 11. Average tracking length across all tests. (Color figure online)

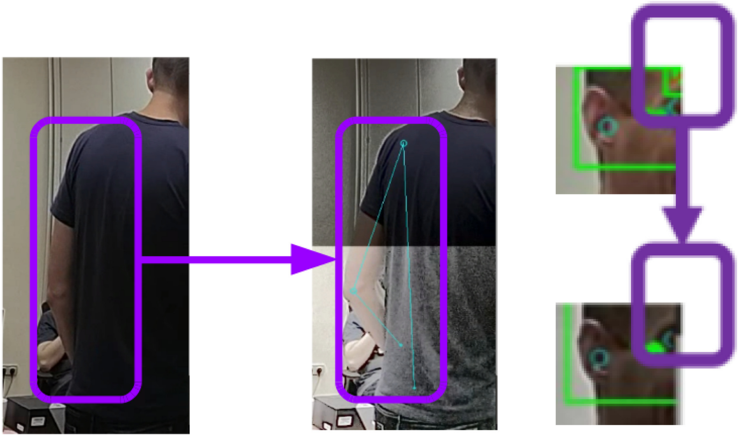


(a) Tracking comparison (b) Detection comparison

Fig. 12. Comparison between computer vision technique (Left) and hybrid approach (Right) [42].

Table 3. Average processing times over 20 runs.

Quality	CV-only (fps)	CV-improved (fps)	Hybrid (fps)
High	0.46	2.73	2.51
Low	0.67	3.79	3.70



(a) Body parts detection comparison

(b) False detection comparison

Fig. 13. Comparison between CV (Left) and hybrid approach (Right) [42].

Finally, we also measured the ability of the system to determine the location of the people in the different zones of the room. We measured the average number of people in each zone across every test video, and compared it to the localization results when using only the vision system with and without the improvements, only the Wi-Fi method, or the hybrid approach. As detailed in Table 4, the hybrid approach is the one that matches the ground truth data more closely.

Table 4. Average number of people detected per zone, using CV-only, Wi-Fi only, hybrid methods and ground truth [42].

Zone	CV	CV improved	Wi-Fi	Hybrid	GT
1	3.2	3.4	3.6	**3.7**	3.9
2	0.7	1.0	1.4	**1.1**	1.2
3	2.8	2.9	3.3	**3.2**	3.2

6 Conclusions and Future Work

In this work we presented an out of the box solution for person detection, tracking and counting based on computer vision and WiFi signal analysis techniques. The system presented improves over a previous prototype [42] by modifying the preprocessing of the data in the computer vision based task. We have compared the hybrid solution, using both CV and WiFi signal techniques, with the improved CV-only and the WiFi-only implementations.

The comparison demonstrates that the hybrid method presented here performs better than the other methods, as it provides higher detection and tracking accuracy with only a small detriment in processing times. We also applied the improved preprocessing step to the hybrid solution. In this case we noticed a negligible improvement when testing the system in lab conditions, and slightly improved tracking accuracy in complex setups such as live shows with many people and poor lighting conditions.

In the future we plan to improve the hybrid system in terms of both performance and functionalities. Performance-wise, gains could be obtained by switching to newer, better models or implementing improved methods for localization via wireless signals. Tracking could improve using face-identification techniques which allow to resume tracking after occlusions. We believe that is possible to achieve real-time (25 fps or better) without compromising the detection accuracy. Regarding functionalities, one obvious improvement is to increase the amount of information extracted, for example running emotion analysis (which correlates directly with user engagement) or action recognition. More interestingly, new functionalities can be added by strengthening the interaction between the vision and wireless systems: for example, the hybrid implementation could use the data from the camera to tweak the parameters used to perform wireless localization, or to reshape the zones of interest based on people movements.

References

1. Altini, M., Brunelli, D., Farella, E., Benini, L.: Bluetooth indoor localization with multiple neural networks. In: ISWPC 2010 - IEEE 5th International Symposium on Wireless Pervasive Computing, 1 June 2010, pp. 295–300 (2010). https://doi.org/10.1109/ISWPC.2010.5483748
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: new benchmark and state of the art analysis. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2014). https://doi.org/10.1109/CVPR.2014.471
3. Asvadi, A., Garrote, L., Premebida, C., Peixoto, P., Nunes, U.J.: Multimodal vehicle detection: fusing 3D-LIDAR and color camera data. Pattern Recogn. Lett. **115**, 20–29 (2018). https://doi.org/10.1016/j.patrec.2017.09.038
4. Bodd, B.: Means, not an end (of the world) the customization of news personalization by European news media. SSRN Electron. J. (2018). https://doi.org/10.2139/ssrn.3141810
5. Bourdev, L., Malik, J.: Poselets: body part detectors trained using 3D human pose annotations. In: 2009 IEEE 12th International Conference on Computer Vision (2010). https://doi.org/10.1109/iccv.2009.5459303
6. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In: arXiv preprint arXiv:1812.08008 (2018)
7. Čehovin, L., Leonardis, A., Kristan, M.: Visual object tracking performance measures revisited. IEEE Trans. Image Process. **25**(3), 1261–1274 (2016)
8. Chang, J.Y., Moon, G., Lee, K.M.: V2V-PoseNet: voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2018). https://doi.org/10.1109/CVPR.2018.00533
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005 (2005). https://doi.org/10.1109/CVPR.2005.177

10. Dari, Y.E., Suyoto, S.S., Pranowo, P.P.: CAPTURE: a mobile based indoor positioning system using wireless indoor positioning system. Int. J. Interact. Mob. Technol. (JIM) **12**(1), 61 (2018). https://doi.org/10.3991/ijim.v12i1.7632
11. Deloitte: 2018 Media and Entertainment Industry Trends — Deloitte US (2018)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
13. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
14. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2014). https://doi.org/10.1109/CVPR.2014.81
15. Glisser (2019). https://www.glisser.com/features/. Accessed 24 Sept 2019
16. Granados, N.: Digital Video and Social Media Will Drive Entertainment Industry Growth in 2019 (2018)
17. Gupta, P., Kar, S.P.: MUSIC and improved MUSIC algorithm to estimate direction of arrival. In: 2015 International Conference on Communication and Signal Processing, ICCSP 2015 (2015). https://doi.org/10.1109/ICCSP.2015.7322593
18. Hassib, M., Schneegass, S., Eiglsperger, P., Henze, N., Schmidt, A., Alt, F.: EngageMeter: a system for implicit audience engagement sensing using electroencephalography. In: Conference on Human Factors in Computing Systems - Proceedings (2017). https://doi.org/10.1145/3025453.3025669
19. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
20. IEEE OUI MAC registries (2019). https://regauth.standards.ieee.org/standards-ra-web/pub/view.html#registries. Accessed 24 Sept 2019
21. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: Proceedings of the British Machine Vision Conference (2010) https://doi.org/10.5244/C.24.12
22. Kalal, Z., Mikolajczyk, K., Matas, J.: Forward-backward error: automatic detection of tracking failures. In: 2010 20th International Conference on Pattern Recognition, pp. 2756–2759. IEEE (2010)
23. Khalil, M., Ebner, M.: Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories. J. Comput. High. Educ. **29**(1), 114–132 (2016). https://doi.org/10.1007/s12528-016-9126-9
24. Nanani, G.K., Prasad Kantipudi, M.V.V.: A study of WI-FI based system for moving object detection through the wall. Int. J. Comput. Appl. **79**(7), 15–18 (2013). https://doi.org/10.5120/13753-1589
25. Lanzisera, S., Zats, D., Pister, K.S.: Radio frequency time-of-flight distance measurement for low-cost wireless sensor localization. IEEE Sens. J. (2011). https://doi.org/10.1109/JSEN.2010.2072496
26. Lienhart, R., Maydt, J.: An extended set of Haar-like features for rapid object detection. In: Proceedings, International Conference on Image Processing (2003). https://doi.org/10.1109/icip.2002.1038171
27. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
28. Liu, W., et al.: SSD: single shot MultiBox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2

29. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-SVMs for object detection and beyond. In: Proceedings of the IEEE International Conference on Computer Vision (2011). https://doi.org/10.1109/ICCV.2011.6126229
30. Martínez-Villaseñor, L., Ponce, H., Brieva, J., Moya-Albor, E., Núñez-Martínez, J., Peñafort-Asturiano, C.: UP-fall detection dataset: a multimodal approach. Sensors (2019). https://doi.org/10.3390/s19091988
31. Meyer, K.A.: Student engagement in online learning: what works and why. ASHE High. Educ. Rep. (2014). https://doi.org/10.1002/aehe.20018
32. Mitchell, J.: Hollywood's Latest Blockbuster: Big Data and The Innovator's Curse (2014)
33. Oguejiofor, O.S., Okorogu, V.N., Adewale, A., Osuesu, B.O.: Outdoor localization system using RSSI measurement of wireless sensor network. Int. J. Innov. Technol. Explor. Eng. **2**, 1–6 (2013)
34. Papandreou, G., Zhu, T., Chen, L.-C., Gidaris, S., Tompson, J., Murphy, K.: PersonLab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018, Part XIV. LNCS, vol. 11218, pp. 282–299. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_17
35. Peng, R., Sichitiu, M.L.: Angle of arrival localization for wireless sensor networks. In: 2006 3rd Annual IEEE Communications Society on Sensor and Adhoc Communications and Networks, Secon 2006 (2007). https://doi.org/10.1109/SAHCN.2006.288442
36. Pizer, S.M., Johnston, R.E., Ericksen, J.P., Yankaskas, B.C., Muller, K.E.: Contrast-limited adaptive histogram equalization: speed and effectiveness. In: Proceedings of the First Conference on Visualization in Biomedical Computing (1990)
37. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: an astounding baseline for recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (2014). https://doi.org/10.1109/CVPRW.2014.131
38. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016). https://doi.org/10.1109/CVPR.2016.91
39. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. CoRR abs/1804.02767 (2018). http://arxiv.org/abs/1804.02767
40. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
41. Richter, R., Kellenberger, T., Kaufmann, H.: Comparison of topographic correction methods. Remote Sens. (2009). https://doi.org/10.3390/rs1030184
42. Sanz Narrillos, M., Masneri., S., Zorrilla, M.: Combining video and wireless signals for enhanced audience analysis. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART, pp. 151–161. INSTICC, SciTePress (2020). https://doi.org/10.5220/0008963101510161
43. Schivinski, B., Christodoulides, G., Dabrowski, D.: Measuring consumers' engagement with brand-related social-media content: development and validation of a scale that identifies levels of social-media engagement with brands. J. Advert. Res. (2016). https://doi.org/10.2501/JAR-2016-004
44. Schollz, Z.: High-precision indoor positioning framework, version 3 (2019). https://github.com/schollz/find3. Accessed 10 Apr 2019
45. Spinello, L., Triebel, R., Siegwart, R.: Multimodal people detection and tracking in crowded scenes. In: Proceedings of the National Conference on Artificial Intelligence (2008)
46. Su, Z., Ye, M., Zhang, G., Dai, L., Sheng, J.: Cascade feature aggregation for human pose estimation. In: CVPR (2019)

47. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep High-Resolution Representation Learning for Human Pose Estimation (2019). http://arxiv.org/abs/1902.09212
48. Wightman, R.: posenet-pytorch (2018). https://github.com/rwightman/posenet-pytorch
49. Yiu, S., Dashti, M., Claussen, H., Perez-Cruz, F.: Wireless RSSI fingerprinting localization (2017). https://doi.org/10.1016/j.sigpro.2016.07.005