An Advanced Framework for Critical Infrastructure Protection Using Computer Vision Technologies

Krishna Chandramouli\textsuperscript{1} and Ebroul Izquierdo\textsuperscript{2}

\textsuperscript{1} Venaka Media Limited, 393, Roman Road, London E3 5QS, UK  
k.chandramouli@venaka.co.uk

\textsuperscript{2} Multimedia and Vision Research Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Road, London E1 4NS, UK  
ebroul.izquierdo@qmul.ac.uk

Abstract. Over the past decade, there has been unprecedented advancements in the field of computer vision by adopting AI-based solutions. In particular, cutting edge computer vision technology based on deep-learning approaches has been deployed with an extraordinary degree of success. The ability to extract semantic concepts from continuous processing of video stream in real-time has led to the investigation of such solutions to enhance the operational security of critical infrastructure against intruders. Despite the success of computer vision technologies validated in a laboratory environment, there still exists several challenges that limit the deployment of these solutions in operational environment. Addressing these challenges, the paper presents a framework that integrates three main computer vision technologies namely (i) person detection; (ii) person re-identification and (iii) face recognition to enhance the operational security of critical infrastructure perimeter. The novelty of the proposed framework relies on the integration of key technical innovations that satisfies the operational requirements of critical infrastructure in using computer vision technologies. One such requirement relates to data privacy and citizen rights, following the implementation of General Data Protection Regulation across Europe for the successful adoption of video surveillance for infrastructure security. The video analytics solution proposed in the paper integrates privacy preserving technologies, high-level rule engine for threat identification and a knowledge model for escalating threat categories to human operator. The various components of the proposed framework has been validated using commercially available graphical processing units for detecting intruders. The performance of the proposed framework has been evaluated in operational environments of the critical infrastructure. An overall accuracy of 97\% is observed in generating alerts against malicious intruders.

Keywords: Person detection · Person re-identification (RE-ID) · Intrusion detection · Face recognition · Region based Fully Connected Network (RFCN) · Un-supervised clustering · Knowledge model · Privacy preserving technologies
1 Introduction

Modern critical infrastructures are increasingly turning into distributed, complex cyber-physical systems that require proactive protection and fast restoration against physical or cyber incidents or attacks. Addressing the challenges faced by the critical infrastructure operators especially responsible for the production and distribution of energy services, DEFENDER\footnote{\url{https://defender-project.eu/}.} project has developed several cyber-physical detectors and operational blueprints to safeguard the future European Critical Energy Infrastructure (CEI) operation against new and evolving threats. In complementary to the cyber threats, the nature of physical threats is compounded by the use of drones in addition to the human intrusion against the infrastructure with malicious intent.

Following the increasing threat of malicious activity carried out against critical infrastructure by human actors there is an exponential increase in the deployment of surveillance infrastructure such as Closed Circuit Television (CCTV) for monitoring the perimeter of critical infrastructure. Traditionally in computer vision research, the task of object detection is to classify a region for any predefined objects from the training data set. Early attempts at object classification also adopted a similar approach for the detection of an image region to be consisting of a drone or not. In this context, the application of computer vision was applied for the selection of suitable representations of objects using handcrafted features. The most successful approaches using handcrafted features require Bag of Visual Words (BoVW) was reported in [24] that includes representations of the objects with the help of local feature descriptors such as Scale Invariant Feature Transform (SIFT) [18], Speeded-Up Robust Features (SURF) [2], and Histogram of Oriented Gradients (HOG) [8]. After training a discriminative Machine Learning (ML) model such as Support Vector Machines (SVM) [6], with such representations the images are scanned for occurrence of learned objects with sliding window technique. In contrast to classical machine learning approaches, the increasing use of deep-learning algorithms has led to the development of object classification and localisation. Two main approaches have been reported in the literature that adopt two different strategies namely (i) two-stage detector, the most representative one, Faster R-CNN; and (ii) one-stage detector, such as YOLO, SSD. Two-stage detectors have been reported to indicate high localization and object recognition accuracy, whereas the one-stage detectors achieve high inference speed. The two stages of two-stage detectors can be divided by RoI (Region of Interest) pooling layer. For instance, in Faster R-CNN (Region based Convolutional Neural Network), the first stage, called a Region Proposal Network (RPN), proposes candidate object bounding boxes. The second stage, features are extracted by RoI Pooling operation from each candidate box for the following classification and bounding-box regression tasks [15]. The research presented in the article focusses on the use of two-stage detector due to the high-accuracy of the deep-learning network model.
Despite the success of computer vision technologies in addressing real-world applications within research environments, the deployment of such solutions within operational environments requires additional services that take into consideration the data privacy and citizen rights. Addressing these crucial requirements, the paper presents the activities carried out in DEFENDER project, for the development of an intelligent framework using computer vision technologies for enhancing the critical infrastructure security. The proposed framework integrates three key technologies namely (i) person detection; (ii) person re-identification (RE-ID) and (iii) facial recognition (FR) components to enrich critical infrastructure security. The overall complementarity of these technologies are leveraged to enhance the perimeter security of critical infrastructure against physical attacks. The processed outcome from these three components are subsequently analysed based on stream analytics to enhance the robustness of the detection through metadata association between individually processed frames captured from the camera sensor. Additionally, the use of privacy-preserving technologies based on Gaussian blur algorithm ensures the compliance of the framework for data privacy. The framework interfaces with the command centre to visualise the alerts generated from the critical infrastructure. The framework integrates an encrypted media repository to comply with privacy-by-design (PbD) principles.

The rest of the paper is structured as follows. Section 2 presents an overview of the studies presented in the literature for person detection, person re-identification and face recognition. Subsequently, in Sect. 3 the proposed intelligent situational awareness framework is presented, which incorporates several key technical innovations including the use of privacy preserving technologies and secure encryption services. Following a detailed analysis of each of the technical innovation, the outcome of the proposed framework is presented in Sect. 4. The paper concludes with observations, remarks and roadmap for the design of physical security detector in Sect. 5.

2 Literature Review

Intelligent video surveillance has been one of the most active research areas in computer vision [29]. Within the computer vision community, the research of person detection has attracted studies from interdisciplinary scientists, interested in the design of autonomous vehicles to intelligent surveillance [14], robot navigation [10,16]. In this section the literature review of three technologies are summarised in two categories as follows.

2.1 Person Detection and Person Re-identification

Person detection is considered as an object detection problem for which the deep-learning models are trained on a number of human representations that takes into account the changes to appearance, cloths, and environment parameters among others. Prior to the recent progress in Deep-Convolutional Neural
Network (DCNN) based methods [26], researchers combined boosted decision forests with hand-crafted features to obtain pedestrian detectors [30]. In the literature, the problem of person detection has been extended to associate human representation captured from multiple-cameras. This has led to the research in single camera Multi-object Tracking (MOT) algorithms. In contrary the Multi-Target Multi-Camera Tracking (MTMCT) algorithms reported in literature are based on off-line method which requires to consider before and after frames to merge tracklets, and do post processing to merge the trajectory. In the literature, hierarchical clustering [33] and correlation clustering [22] are reported for merging the bounding box into tracklets from neighbor frames. Addressing the need to generate real-time tracker without the apriori knowledge of person tracks, an online real-time MTMCT algorithm has been reported in the literature which aims to track a person cross camera without overlap through a wide area. The framework utilises a person detection based on Openpose [4], building on a multi-camera tracker extended by a single camera tracker MOTDT [5]. In order to improve the performance, lots of research focus on local feature instead of the global feature of the whole person, such as slice [27], pose and skeleton alignment [34]. While matching local features help to improve in Person Re-ID, the challenge of pose variation remain open due to the different view from camera.

2.2 Face Recognition

Face recognition (FR) has been the prominent biometric technique for identity authentication and has been widely used in many areas, such as military, finance, public security and daily life [28]. In 2014, DeepFace [25] achieved the state of the art accuracy on the famous LFW benchmark [12], approaching human performance on the unconstrained condition for the first time (DeepFace: 97.35% vs. Human: 97.53%), by training a 9-layer model on 4 million facial images. Inspired by this work, research focus has shifted to deep-learning-based approaches, and the accuracy was dramatically boosted to above 99.80% in just three years. Deep learning technique has reshaped the research landscape of FR in almost all aspects such as algorithm designs, training/test datasets, application scenarios and even the evaluation protocol. In 2015 a system named Multi-task Cascaded Convolutional Networks (MTCNN) showed that a joint implementation for face alignment and detection could achieve higher levels of accuracy and thus has been integrated in the current implementation. Some noteworthy face recognition surveys include [3,13,23,32]. These comprehensively survey face recognition systems prior to DeepFace. Hence, these surveys do not discuss the new sophisticated deep learning approaches that emerged during the last decade. Surveys that discuss deep face recognition have singled out face recognition as an individual discipline rather than a collection of components adopted from different studies. These surveys generally discuss the face recognition pipeline: face preprocessing, network, loss function, and face classification [17,19] or discuss a single aspect of face recognition such as 3-D face recognition [28], illumination face recognition [20] or pose invariant face recognition [9]. Although these surveys
are important and provide an excellent basis for the analysis of the state-of-the-art in the field, they do not provide conclusive comparisons or analysis of the underlying network architectures.

3 Critical Infrastructure Protection Framework

The proposed critical infrastructure protection framework interfaces directly with the three detectors developed in the project and enables the construction of high-level surveillance events such as intrusion, loitering and access control authentication for early stage identification of malicious actions against critical infrastructure. The proposed framework is presented in Fig. 1 and consists of three stages namely (i) video sequence, captured from the sensor deployed at the perimeter of the infrastructure; (ii) computer vision technologies, capable of processing multiple video streams using deep-learning network and (iii) the situational awareness components, in which the organisational policies and practices are encoded to ensure the security restrictions are not violated by intruders. In order to protect the privacy and citizen rights, the proposed framework incorporates the use of privacy preserving technologies as outlined in Sect. 3.4. The processed outcome from the situational awareness component is then integrated into the command centre to categorise the threat and also the severity with which the mitigation actions should be carried out according to the organisational policies.

![Diagram of Framework](image)

**Fig. 1. Framework**

3.1 Computer Vision Technologies

In this section, an overview of the implementation details carried out in the project for the integration of computer vision technologies is presented.
Person Detection and Person Re-identification (RE-ID). One of the challenges of MTMCT framework presented in the literature is the inability of the framework to anchor against a specific Person of Interest (POI) for modelling threat events such as loitering. In this regard, the MTMCT component has been further developed to include “unsupervised multi-camera person re-identification” framework. The overall framework design of the proposed framework is presented in Fig. 2. The implementation of the person detection component relies on the use of Region based Fully Connected Neural Network (RFCN), followed by the feature extraction of the detected person with a set of deep-learning features. The deep-learning features extracted from the identified bounding boxes are then subjected to the application of an unsupervised algorithm for clustering the people. The processing of the deep-learning features are further exploited to ensure the infrastructure operators can provide an anchor image of a POI, to retrieve the appearance of the person across several surveillance cameras.

![Fig. 2. Person re-identification](image)

The novelty of the RFCN network relies in the consideration of two-stage object detection strategy namely (i) region proposal and (ii) region classification. The scientific rationale behind the use of the two-stage proposal is elaborated in [7]. Following the extraction of the regions (RoIs), the R-FCN architecture is designed to classify the RoIs into object categories and background. In R-FCN, all learnable weight layers are convolutional and are computed on the entire image. The last convolutional layer produces a bank of $k^2$ position-sensitive score maps for each category, and thus has a $k^2(C + 1)$ - channel output layer with $C$ object categories (+1 for background). The bank of $k^2$ score maps correspond to a $k \times k$ spatial grid describing relative positions. For example, with $k \times k = 3 \times 3$, the 9 score maps encode the cases of top-left, top-center, top-right, ..., bottom-right of an object category. R-FCN ends with a position-sensitive RoI pooling layer. This layer aggregates the outputs of the last convolutional layer and generates scores for each RoI. In comparison with the literature [11], the position sensitive RoI layer in RFCN conducts selective pooling, and each of the
Situation Awareness Framework 113

$k \times k$ bin aggregates responses from only one score map out of the bank of $k \times k$ score maps. With end-to-end training, this RoI layer shepherds the last convolutional layer to learn specialized position-sensitive score maps. Subsequent to the extraction of the people, the next steps is to extract deep-learning features from the blobs that are identified as people. Thus, for the purposes of DEFENDER project, where the infrastructure operator are only concerned tracking a specific perpetrator, such as person of interest, it is vital to adopt the solution to identify the anchor points, which are provided as input to the system. To address such a need, the use of unsupervised clustering is carried out, to cluster the blobs extracted from the RFCN network. Subsequently, the features used is also able to provide the infrastructure operator to identify and select a specific person who is considered a POI for the identification across multiple cameras. The implementation of the feature extraction has been carried out using two deep-learning network models namely (i) RESNET-18 resulting in the deep-learning feature of length $1 \times 512$ and (ii) AlexNet deep learning feature resulting in the feature length of $1 \times 4096$. For the overall evaluation of the proposed unsupervised clustering framework, a set of videos across London has been captured with actors playing the role of person of interest. The experimental results of the component included a cluster size of 50 for each of the video footage, and the aggregated result of the people detector was clustered using K-Means using both the RESNET-18 and AlexNet deep-learning features. The results achieved 96% accuracy in each cluster for the 4-actors embedded within the content capture section for the validation of the component efficiency. Subsequent analysis will be carried out for the evaluation of the retrieval performance for each of the anchor as selected by the infrastructure operator.

Facial Recognition Component. The implementation of face recognition component within DEFENDER project consists of two modules namely (i) face detection and (ii) face recognition. The face detection mechanism begins with a new image representation named “Integral Image”. The integral image is computed on an image using some pixel based operations. This integral image enables fast evaluation of features and hence is used compute a set of features which are reminiscent of Haar Basis functions. However, the number of haar-like features computed on an image sub-window is very large, far larger than the number of pixels itself, and hence requires filtering to obtain only a small set of critical features. This is carried out using an Adaboost based learning process which excludes a majority of available features and narrows the feature-set down to include only the critical features. The process itself is a slight modification introduced to the AdaBoost procedure: Each step in the boosting process involves a weak classifier that depends only on a single feature, and this is modified to be used as a feature selection process. Next, the features are fed into a set of complex classifiers, which are combined in a cascade structure. This method of combination is said to enable an increase in speed by focusing attention on promising regions of the image. The face recognition system integrated in the DEFENDER detector is based on the specification of Facenet. Instead of training
a face recognition system in the form of a conventional classifier, FaceNet implements a system which directly maps the input face thumbnails to the compact Euclidean space. The Euclidean space is generated such that the L2 distance between all faces of the same identity is small, whereas the L2 distance between a pair of face images from different identities is large. This is enabled by triplet loss which, by definition, aims at minimizing the distance between pairs of same identity while maximizing the distance between pairs of different identities. For more details on the implementation of the face recognition, readers are referred to [31]. Additional details on the video analytics components integrated within the platform can be found in [1].

3.2 Temporal Smoothing

The computation of media streams captured from the detectors are often subjected to frame level processing as supported by the specification of the camera. One of the challenges in performing a frame level analysis results in the lack of ability to model the global situational awareness of the environment. To address this challenge, the proposed security framework implements a stream analytics solution with the use of latency processing that allows for buffering the input media stream for a period between 0.15 to 0.6 s (for video streams captured at 60 fps and 25 fps respectively) prior to the deployment of the computational module. The stream analytics platform allows for the construction of global situational awareness through the consolidation of the media sources collected in the buffer. The initial latency period of the detector does not affect the performance of the detector, rather enhances the reliability and severity measure of the alerts generated.

3.3 Encrypted Media Repository

Following the implementation of GDPR across Europe, data privacy and protection has become an inherent necessity to adopt privacy-by-design methodologies for system implementation. Therefore, the project has adopted the use of encryption solution to process the media data captured from the detectors. An overview of the system adopted within the project has been presented in Fig. 3. The encryption is carried out using AES 256 Crypt specification\(^2\). All the media data extracted from the computational units are encrypted using a pre-determined password as configured within the system deployment.

3.4 Privacy Preserving Technologies

In the literature, there are seven privacy preserving techniques reported in the literature including blur, pixelating, emboss, solid silhouette, skeleton, 3D avatar

\(^2\)https://www.aescrypt.com/.
and invisibility [21]. The new capabilities of such systems provide the computational tools the ability to collect and index a huge amount of private information about each individual, approaching the perimeter of the critical infrastructure. However, based on the privacy by design methodology adopted within the project, the framework transfers no personal data should be processed or made available to the command centre until a threat has been identified, which requires neutralisation. To this end, the use of privacy-by-design methodology incorporated within the media processing framework adopts the use of Gaussian blur to mask the identity of the person against the extraction of usable features. For the person re-identification component, the use of feature extraction module based on AlexNet protects the identity of the person without compromising the computational ability of the platform.

### 3.5 Knowledge Model

The knowledge model represents a set of high-level business rules that encodes the notion of abnormal behaviour at the perimeter of critical infrastructure based on temporal association of people detected using the surveillance infrastructure. The syntax adopted for the rule definition is based on JavaScript Object Notation (JSON) representation that systematically formalises the attributes of the detector, for anomaly detection. Each of the extracted person object from the perimeter is indexed against a unique identifier. The encrypted media repository creates a new index for every new person being detected. Internally, a large-index of temporal occurrence of the individual people are stored. The occurrence index consists of three categories of time stamps namely (i) past, (ii) current and (iii) ignored. For each of the new person being detected, a similarity metric is applied to associate the new person to existing index of person being treated. In addition, the high-level event representation syntax for event detection such as loitering, reconnaissance are also encoded in the JSON format based on the timestamps. The detector outcomes following the media processing are also exported in JSON format as specified in the knowledge base. The threat level severity
are pre-determined based on the proximity of the threat against breaching the perimeter of the critical infrastructure.

### 3.6 Threat Evaluator

The threat evaluator module receives input from privacy preserving technology output and the organisational guidelines on the threat models and severity. In addition, the module will also interact with the encrypted media repository to present the decrypted media data to the command centre upon the detection of the threat. The module evaluates the spatial constraints configured within the platform to determine the threat level. As an instance, the intruder detector has two levels of severity based on the proximity of threat to the critical infrastructure. The severity levels are appropriately identified and the evolution of the threat in time will be continuously monitored through alerts shared with the command centre. The spatial configuration of the infrastructure environment are coded into spatial coordinates as viewed by the sensor. The 2D image coordinates are internally mapped against the real-world distance measures A visual description of the spatial mapping is presented in Fig. 5. For the determination of high-level threat analyser such as reconnaissance or loitering, specification of temporal rules have been defined within specific time windows to identify malicious perpetrators. These rules are quantified through a JSON syntax as outlined in Sect. 3.5. The anonymous identity labels assigned to the individual extracted from the detectors are used to visually cluster and enable correlation between the repeated appearance of the same individual near the vicinity of the critical infrastructure. The command centre provides a unified interface for the collection of media captured from distributed availability of the detectors. The web-interface allows for the easy navigation for the selection of different detector output that are spatio-temporal indexed.

### 3.7 Command Centre

The command centre is a central interface that interfaces with each of the detectors and integrates the different modules within the proposed framework. Upon the installation of the detector at the perimeter, the detector is configured with the command centre through the specification of IP address through which the detector will communicate with the command centre. The detector installation at the Erchie trial site for the intruder detection is presented in Fig. 4. The registration of the detector installation is carried out using a RESTful interface and JSON metadata consumed by the command centre. Subsequently, the evidence collected from the detector both raw data and the processed output are both transmitted to the threat evaluator module, which upon the determination of the data sources, decides to present the privacy protected results or the raw data based on the severity of the threat level.
4 Experimental Results

The overall experimental results of the proposed security framework have been evaluated within the context of operational environments of the critical infrastructure assets being protected against external threats. The results summarised below are obtained from the two pilot trials carried out in DEFENDER namely Erchie and Okrogolo. While the human intruder detection has been extensively evaluated in the Erchie trial, the face recognition component has been evaluated as a part of the Okrogolo trial. The voluntary participation of the actors has been used to evaluate the system performance. The evaluation of the system performance included the computational latency required for the detector to send notifications to the command centre on the appearance of intruders and the evolution of the threats sequence in time. To facilitate the deployment of the countermeasures against the threat the alert sent to the command centre based on the event are separated by 20s. Based on the evolution of the threat, from the proximity of the infrastructure to the approach of the perimeter, the status flag embedded within the alerts are changed from LOW, MEDIUM, HIGH and VERY HIGH. The experimental evaluation of the detector carried out in the DEFENDER field trial, yielded an accuracy of 96.7% in the person detection. The spatial depiction of the results obtained from the trail is presented in Fig. 5.

In order to evaluate the robustness of the framework, a continuous experimentation process was adopted in which the detector was operated for long periods of time for the detection of different threat levels. A summary of the evaluation results are presented in Table 1. The long-term durability of the physical security
detectors were evaluated against the ability of the detector to identify the appropriate threat levels based on the critical infrastructure perimeter configuration. A total of 4 participants volunteered to take part in the trial for the evaluation of the detector. The infrastructure intrusion attack scenario was orchestrated with several approaches to the perimeter being considered. During the operation of the trial, a total of 267 events were identified resulting in a total of 350 threats. The report of additional 83 alerts were attributed to the dual detection of person intrusion due to mis-classification of non-human objects as intruders. For each of the intruder detected, an alert was generated and transmitted to the command centre.

**Table 1.** Robustness evaluation of the proposed framework for intrusion detection

| Parameter of evaluation                        | Metric          |
|-----------------------------------------------|-----------------|
| Total duration of the deployment              | 21,600 s        |
| Total amount of video data processed          | 648,000 frames  |
| Total number of participants in the trial     | 4               |
| No. of incidents identified                   | 267             |
| No. of alerts generated                       | 350             |
| No. of repeated alerts generated with high severity | 47             |
| Computational latency of the component        | 25 m            |
Table 2. Robustness evaluation of the proposed framework for face recognition

| Parameter of evaluation                                      | Metric |
|--------------------------------------------------------------|--------|
| Total number of participants in the trial                    | 14     |
| % accurate detections with single person in the frame        | 97.8   |
| % detections with multiple persons in the frame              | 96.7   |
| Computational latency of the component                       | 35 m   |

The face recognition component has been evaluated against changing environmental conditions with a total of 40 participant (consisting of 14 nationalities and members from 9 ethnic background) features annotated within the database. The media data captured from the street level camera has been integrated with the module. In contrast to the scientific review of face recognition solutions, the evaluation metrics applied here includes the distance measurement at which faces can be detected and the distance at which the recognition takes place. The integrated face recognition component has shown to deliver reliable performance between 10 to 15 m distance between the detector and the subject annotated in the database. The use of MTCNN face detector delivers performance at 20 m and beyond depending on the availability of facial characteristics and features. The results of continuous evaluation of experimental results are summarised in Table 2. A total of 14 voluntary participants were evaluated within the context of operational environment. An overall accuracy of 97.8% were observed.

Fig. 6. Face Recognition in operational environment for authentication
when a single person detection was carried out. Subsequently, the multi-person recognition up to 10 participants has yielded an average of 96.7% overall accuracy. The decreased efficiency of the algorithmic performance is attributed to the configuration of thresholds required to against the L2 norm of the algorithm output.

5 Conclusion and Future Work

The paper has summarised the integration of three computer vision technologies within DEFENDER project. The paper outlined the implementation of additional computational components to enhance the operational capacity of the laboratory validated solution. The real-world deployment of the solution has been extensively evaluated in Erchie and as a part of the Okrogolo trails. The security framework brings together several key innovations to deliver real-time operational insight to the infrastructure operators for deploying mitigating actions against perceived threats. The novelty of the proposed solution relies in the use of privacy-by-design methodology for protecting the identity and rights of citizen. The structured encoding of organisational policies for identifying threats are considered by the threat evaluator to deliver alerts to the command centre of the infrastructure operator. The future work will review the design of detectors and enhance the communication protocol for enabling bi-directional transfer of control and media signals between the command centre and the media detectors. In addition, the performance of the detector will be continuously reviewed and kept abreast with latest scientific innovations results reported in the literature. Finally, the structure of the knowledge model will adopt the use of ontology and semantic representation for encoding security specification of critical infrastructure.

Acknowledgement. The research activities leading to this publication has been partly funded by the European Union Horizon 2020 Research and Innovation program under MAGNETO RIA project (grant agreement No. 786629) and DEFENDER IA project (grant agreement No. 740898).

References

1. Arachchilage, S.W., Izquierdo, E.: A framework for real-time face-recognition. In: 2019 IEEE Visual Communications and Image Processing (VCIP), pp. 1–4 (2019)
2. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-Up Robust Features (SURF). Comput. Vis. Image Underst. (2008). https://doi.org/10.1016/j.cviu.2007.09.014
3. Bowyer, K.W., Chang, K., Flynn, P.: A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. Comput. Vis. Image Underst. (2006). https://doi.org/10.1016/j.cviu.2005.05.005
4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 (2017). https://doi.org/10.1109/CVPR.2017.143
5. Chen, L., Ai, H., Zhuang, Z., Shang, C.: Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In: Proceedings - IEEE International Conference on Multimedia and Expo (2018). https://doi.org/10.1109/ICME.2018.8486597
6. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learn. (1995). https://doi.org/10.1023/A:1022627411411
7. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: Object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems (2016)
8. 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
9. Ding, C., Tao, D.: A comprehensive survey on pose-invariant face recognition. ACM Trans. Intell. Syst. Technol. (2016). https://doi.org/10.1145/2845089
10. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. Int. J. Robot. Res. (2013). https://doi.org/10.1177/0278364913491297
11. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2015). https://doi.org/10.1109/TPAMI.2015.2389824
12. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report (2007)
13. JafriRabia, R., A.: J. Inf. Process. Syst. 5(2), 41–68
14. Jiang, Z., Huynh, D.Q.: Multiple pedestrian tracking from monocular videos in an interacting multiple model framework. IEEE Trans. Image Process. (2018). https://doi.org/10.1109/TIP.2017.2779856
15. Jiao, L., et al.: A survey of deep learning-based object detection. CoRR abs/1907.09408 (2019), http://arxiv.org/abs/1907.09408
16. Khan, A., Rinner, B., Cavallaro, A.: Cooperative robots to observe moving targets: Review. IEEE Trans. Cybernet. (2018). https://doi.org/10.1109/TCYB.2016.2628161
17. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision (2015). https://doi.org/10.1109/ICCV.2015.425
18. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the IEEE International Conference on Computer Vision (1999). https://doi.org/10.1109/iccv.1999.790410
19. Masi, I., Wu, Y., Hassner, T., Natarajan, P.: Deep face recognition: a survey. In: Proceedings - 31st Conference on Graphics, Patterns and Images, SIBGRAPI 2018 (2019). https://doi.org/10.1109/SIBGRAPI.2018.00067
20. Ochoa-Villegas, M.A., Nolazo-Flores, J.A., Barron-Cano, O., Kakadiaris, I.A.: Addressing the illumination challenge in two-dimensional face recognition: a survey. IET Comput. Vis. 9(6), 978–992 (2015). https://doi.org/10.1049/iet-cvi.2014.0086
21. Padilla-López, J.R., Chaaraoui, A.A., Flórez-Revuelta, F.: Visual privacy protection methods: a survey. Expert Syst. Appl. 42(9), 4177–4195 (2015). https://doi.org/10.1016/j.eswa.2015.01.041
22. Ristani, E., Tomasi, C.: Features for multi-target multi-camera tracking and re-identification. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2018). https://doi.org/10.1109/CVPR.2018.00632
23. Scheenstra, A., Ruifrok, A., Veltkamp, R.C.: A survey of 3d face recognition methods. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) Audio- and Video-Based Biometric Person Authentication, pp. 891–899. Springer, Berlin Heidelberg, Berlin, Heidelberg (2005)
24. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: Proceedings of the IEEE International Conference on Computer Vision (2003). https://doi.org/10.1109/iccv.2003.1238663
25. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)
26. Tian, Y., Luo, P., Wang, X., Tang, X.: Deep learning strong parts for pedestrian detection. In: Proceedings of the IEEE International Conference on Computer Vision (2015). https://doi.org/10.1109/ICCV.2015.221
27. Varior, R.R., Haloi, M., Wang, G.: Gated siamese convolutional neural network architecture for human re-identification. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2016). https://doi.org/10.1007/978-3-319-46484-8_48
28. Wang, M., Deng, W.: Deep face recognition: A survey. CoRR abs/1804.06655 (2018). http://arxiv.org/abs/1804.06655
29. Wang, X.: Intelligent multi-camera video surveillance: a review. Pattern Recogn. Lett. (2013). https://doi.org/10.1016/j.patrec.2012.07.005
30. Zhang, S., Benenson, R., Schiele, B.: Filtered channel features for pedestrian detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2015). https://doi.org/10.1109/CVPR.2015.7298784
31. Zhang, X., Chandramouli, K., Gabrijelcic, D., Zahariadis, T., Giunta, G.: Physical security detectors for critical infrastructures against new-age threat of drones and human intrusion. In: 2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW), pp. 1–4 (2020)
32. Zhang, X., Gao, Y.: Face recognition across pose: a review. Pattern Recogn. (2009). https://doi.org/10.1016/j.patcog.2009.04.017
33. Zhang, Z., Wu, J., Zhang, X., Zhang, C.: Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project. CoRR abs/1712.09531 (2017), http://arxiv.org/abs/1712.09531
34. Zheng, L., Huang, Y., Lu, H., Yang, Y.: Pose-invariant embedding for deep person re-identification. IEEE Trans. Image Process. (2019). https://doi.org/10.1109/TIP.2019.2910414