Title:	Document Version:
D5.4 Human-centred computer vision for crowd protection	1.1

Project Number:	Project Acronym:	Project Title:
H2020-740466	LETSCROWD	Law Enforcement agencies human factor methods and Toolkit for the Security and protection of CROWDs in mass gatherings

Contractual Delivery Date:	Actual Delivery Date:	Deliverable Type*-Security*:
M12 (April 2018)	M12 (April 2018)	R-PU

*Type: P: Prototype; R: Report; D: Demonstrator; O: Other.

**Security Class: PU: Public; PP: Restricted to other programme participants (including the Commission); RE: Restricted to a group defined by the consortium (including the Commission); CO: Confidential, only for members of the consortium (including the Commission).

Responsible:	Organisation:	Contributing WP:
Fabio Roli	UNICA	WP5

Authors (organisation):

Fabio Roli (UNICA)
Giorgio Fumera (UNICA)

Abstract:

This document is a progress report of task T5.4, Human-centred computer vision (HCV) tool, which is part of work package WP5, aimed at developing a human-centred supporting toolkit for law enforcement agencies. It describes the activities already carried out (analysis of background research, and definition and analysis of requirements) and the ongoing activities (design and implementation) toward the development of the HCV tool. It also addresses the integration with the crowd modelling and dynamic risk assessment tools developed in other tasks and work packages.

Keywords:

Computer vision, video surveillance, appearance-based person re-identification, attribute-based people search, crowd monitoring, crowd behaviour analysis, detection of anomalous crowd behaviours, human-in-the-loop computer vision

Revision History

Revision	Date	Description	Author (Organisation)
V0.1	09.04.2018	First draft	Giorgio Fumera (UNICA), Fabio Roli (UNICA)
V0.2	18.04.2018	Version with comments from the peer review (DBL)	Alessia Golfetti (DBL), Sabina Giorgi (DBL)
V0.3	02.05.2018	Second draft, addressing peer review comments by PROPRS and DBL	Giorgio Fumera (UNICA), Fabio Roli (UNICA)
V1.0	02.05.2018	Final version ready to be submitted to EC	Giorgio Fumera (UNICA), Fabio Roli (UNICA)
V1.1	17.08.2018	Revised version with changes required by the Expert Review Report (26 July 2018)	Giorgio Fumera (UNICA), Fabio Roli (UNICA)

This project has received funding from the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement № 740466.

More information available at https://letscrowd.eu

Copyright Statement

Index

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 PURPOSE OF THE DOCUMENT

This document is a progress report on task T5.4, aimed at developing a human-centred computer vision tool to support Law Enforcement Agency (LEA) operators in monitoring and investigation tasks related to the security of mass gathering events. The human-centred computer vision tool will provide three functionalities as distinct software tools: (1) person re-identification (retrieving images of an individual of interest based on clothing appearance similarity with a query image of that individual), (2) people search (retrieving images of individuals whose clothing appearance matches a given textual description), and (3) crowd monitoring (including crowd density estimation, detection of patterns of crowd movement, and detection of anomalous/suspicious crowd behaviours). This document describes the work that has already been carried out toward the development of the human-centred computer vision tool (analysis of background research, definition and analysis of requirements, design of the person re-identification and people search tools), and the ongoing activities (design of the crowd monitoring tool, implementation of the three tools, and integration with the crowd modelling and Dynamic Risk Assessment tools). The second version (deliverable D5.8) will include a user guide for LEAs as an Appendix.

1.2 SCOPE OF THE DOCUMENT

Task T5.4 is part of work package WP5, whose overall objective is to develop a human-centred supporting toolkit for LEAs made up of different, integrated techniques and software modules. In particular, it is planned in the Description of Work that the human-centred computer vision tool is integrated with the crowd modelling tool developed in task T5.1. A higher-level integration is also planned between the human-centred supporting toolkit of WP5 (including the human-centred computer vision tool) and the Dynamic Risk Assessment tool of work package WP3. Accordingly, the design of the human-centred computer vision tool is being conducted with a strong focus on the integration with the outcomes of T5.1 and of WP3. This document will be the basis for deliverable D5.8.

1.3 STRUCTURE OF THE DOCUMENT

This document is structured into four main sections. Background research on the three considered computer vision tasks is reported in Sect. 2, including a literature review, an analysis of related commercial products and an analysis of publicly available data sets. User and system requirements, and the use case for the human-centred computer vision tool, previously defined in work package WP2, are summarized in Sect. 3. Design of the human-centred computer vision tool is described in Sect. 4, including the links to other tasks and work packages. Finally, implementation details are reported in Sect. 5. The user guide will be reported as an annex of deliverable D5.8.

2 BACKGROUND RESEARCH

The computer vision (CV) research community is devoting a considerable effort to topics related to video surveillance (VS) since over twenty years; typical examples are object detection, tracking and recognition (e.g., cars and pedestrians), event recognition (e.g., understanding the behaviour of individuals or of a crowd) and, as a more recent example, person re-identification from a network of cameras. The main goal is to develop intelligent VS tools capable of supporting human operators in monitoring and forensic investigation tasks, partially automatizing them. Since the 2000s some CV tools became to be deployed in commercial products by a few video analytics companies (usually academic spin-offs) and solution providers. The interest in these tools has rapidly grown in the past few years due to the increasing demand of security, and of the pervasive deployment of VS systems both in private (e.g., banks) and public places (e.g., streets and stadiums). However, except for specific tasks (e.g., face or license plate recognition) and under controlled conditions, CV research outcomes are not mature yet for a large-scale commercial deployment, and many open issues remain to be solved before their performance meets the requirements of practical applications in more complex and unconstrained real-world tasks like crowd behaviour understanding and person re-identification.

This section gives an overview of the research and applications of CV for intelligent VS, focusing on the functionalities of the human-centred computer vision (HCV) tool: appearance-based person re-identification, attribute-based people search, and crowd monitoring. First a review of the scientific literature is given; then the state of the art is described in terms of the available software implementations and commercial products; the developments required to achieve the goals of the HCV tools are then pointed out; finally, the publicly available data sets that can be used to develop, validate and demonstrate the HCV tool are described.

2.1 LITERATURE REVIEW

In the following a concise literature review of existing approaches and methods for person re-identification, people search, and crowd monitoring is given. Given the considerable amount of research papers on person re-identification and crowd monitoring, recent literature surveys of these fields published in top international journals are used as reference.

2.1.1 Person re-identification

Person re-identification is the task of retrieving images of an individual of interest in video frames acquired by different, possibly non-overlapping cameras of a VS network, using an available image of that individual as a *query* (or *probe*) (1). A possible application is to support human operators in forensic investigations, to search for a suspect individual over a large amount of videos recorded by a VS system. Person re-identification focuses on typically unconstrained video surveillance settings, which are characterized by relatively low camera resolution, pose and lighting variations, occlusions, and differences between cameras. Such features make recognition technology based on strong biometrics, like face, unfeasible. The most widely used cue in the literature is therefore clothing appearance, although it is weaker than face biometrics, and is valid only for a relatively short time. This research topic has been introduced in 2006 (2), and rapidly gained popularity since 2013 (3); the number of papers currently published in international journals and conferences exceeds 600.[1]

The core functionality of a person re-identification system (see Figure 1) is to match a query image of an individual of interest to a given set of images of individuals, called *template gallery*, and to sort the template gallery for decreasing visual similarity to the query (in terms of clothing appearance), evaluated by a suitable *similarity measure*. The query image has to be selected by an operator from a video frame acquired by one of the cameras of a VS system. The template gallery is instead made up of images of individuals automatically detected and extracted from footage acquired by the same VS system. Ideally, the

[1] Source: dblp computer science bibliography, https://dblp.uni-trier.de/, search query: "person re-identification", visited on March 23, 2018.

gallery images (if any) of the query identity will appear on the top of the returned ranked list, allowing the operator to quickly find them without having to manually scan all the underlying videos. As a variant, query and templates can be consist of *tracks*, i.e., multiple images extracted from a sequence of video frames, to increase robustness to pose and lighting variations and to occlusions.

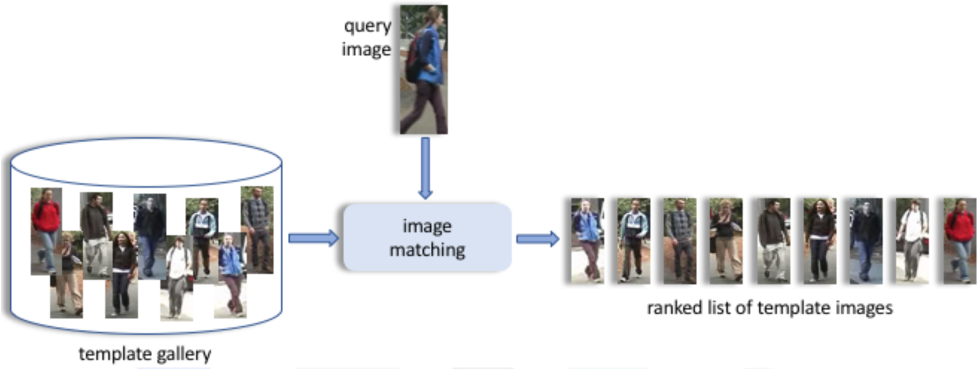


Figure 1 – Core functionality of a person re-identification system: matching a probe image to a set of template gallery images, and sorting the templates by decreasing similarity of clothing appearance to the query; ideally, all the templates having the same identity as the query (if any) should appear at the top of the list.

Two main automatic preprocessing steps are required to build the template gallery: pedestrians have to be automatically detected from raw videos, and their images have to be extracted, possibly in terms of tight bounding boxes to minimize interference with the background, objects or other people (ideally, only the silhouette should be extracted). Tracking algorithms can also be used either to extract tracks instead of single images, or to improve detection accuracy. In the person re-identification literature these are usually considered as independent steps and are not directly addressed, although they strongly affect the re-identification performance (4): this issue is discussed below.

The two main components of a person re-identification method are a *descriptor* of clothing appearance, to be computed for each template and probe image, and a similarity measure between any pair of descriptors, which is used to match probe and template images; the similarity value is commonly called *matching score*. Descriptors should ideally be robust to pose and lighting variations, occlusions, and differences between cameras. In early work ad hoc descriptors were proposed, together with a suitable similarity measure (see, e.g., (5)). Usually they are based on colour (using different colour spaces like RGB and HSV) and texture, possibly computed from different body parts, e.g., simply defined as horizontal strips of predefined size of a pedestrian image, or obtained from more sophisticated body part detectors. More complex approaches use machine learning techniques to model the colour transformation among different cameras, or to create a more discriminative descriptor for a specific query; other approaches also exploit spatial information about the camera setting (position and field of view) (1).

A different approach from the definition of ad hoc descriptors and similarity measures is to use existing descriptors (features) previously proposed for object recognition tasks (pedestrians, faces, textures, etc.), like histogram of oriented gradients (HOG) and local binary patterns (LBP), and then automatically defining a similarity measure through metric learning techniques, using pairs of images of the same or different identities as training data (6). Since 2014 deep learning techniques, and in particular convolutional neural networks (CNN) have become mainstream also in person re-identification, in the wake of their success in the CV field. Their main potential advantage is that they do not require the explicit definition of image descriptors or features, which are automatically learnt from the raw image pixels during the training process, although this requires relatively large training data. Deep learning is being used in person re-identification in two main ways (4): either to automatically define image descriptors, by training a CNN to recognize a predefined set of individuals (different from the queries, which are usually known only during operation), and then using standard metrics like the Euclidean or cosine distance as similarity/dissimilarity measure; or to directly match a pair of input images of individuals (query and template), using a training set made up of pairs of images of the same and of different identities (also in this case, different from the query identities). The latter approach performs both descriptor and metric learning. In the case several images of the query and the template identities are available, obtained by a tracking tool from a sequence of frames, CNN architectures that take into account temporal information have also been defined (4).

In addition to low-level descriptors, some authors proposed to use high-level, semantic attributes related to clothing appearance (e.g., the colours of upper and lower body clothing, and the presence of bags), automatically detected by machine learning techniques. Attribute-based re-identification techniques are described in Sect. 2.1.2, since they are closely related to the people search task.

As in many, challenging computer vision tasks, the effectiveness of person re-identification methods is still far from meeting the requirements of real-world applications, despite increasing performance is being reported on benchmark data sets (3). One of the solutions that have recently been proposed in the CV field to address this issue, especially when machine learning (thus, data-driven) techniques are used, is the so-called human-in-the-loop (HITL) approach (see, e.g., (7)). Its rationale is to leverage complementary strengths of humans and machines in vision tasks, by exploiting some form of human feedback on the outcome of CV algorithms to improve the performance of such algorithms. The HITL approach is particularly suited to person re-identification systems, where a low-effort feedback can be obtained from the operator, e.g., by indicating whether or not the identity of a given template image corresponds to the query's identity. However, in the currently large body of literature on person re-identification only a few authors have proposed HITL methods. In the following the existing methods are described in more detail, as the HITL approach will be one of the main features of the HCV tool.

HITL methods proposed so far are characterized by two main features: the kind of feedback requested to the operator, and how such a feedback is exploited to update the re-identification system. In the method of (8) the operator is shown the sorted list of templates, and is asked to select some templates similar to the query and some dissimilar ones; the similarity measure is then updated through a metric learning technique to push the selected dissimilar templates to the bottom ranks and the similar ones to the top ranks; the template gallery is then re-ranked; several iterations can be carried out, e.g., until the operator finds a template image having the same identity as the query. In (9) a two-stage method was proposed: in the first stage a generic descriptor is used; the operator is shown the top 50 matches and is asked whether the query identity is present or not among them; if not, a classifier is trained to discriminate the query from the bottom-ranked template images (assumed to be the most diverse to the query), and the template gallery is re-ranked according to the similarity measure evaluated by such a classifier. Similarly, in (10) a generic descriptor is used first, and the top 50 matches are shown to the operator; if the query identity is not present among them, the operator is asked to select a "strong negative" template (a different identity than the query, exhibiting a different clothing appearance) and optionally a few "weak negatives" (different identities than the query, with similar clothing appearance); this information is used to update the matching scores of the template gallery with respect to the query, with the aim of pushing the strong negative toward bottom ranks and the weak negatives (if any) toward the top ranks; several such iterations can be carried out. In (11) the re-identification task with operator's feedback is viewed as a particular case of the more general task of image retrieval with relevance feedback, and the setting where queries and templates are represented by tracks instead of single images is considered; their method is aimed at retrieving all the existing tracks of the query identity in the template gallery (in image retrieval terms, to maximize the *recall* performance metric). A generic feature set is used as a descriptor, and a weighted Euclidean distance with different weights for each feature is used as the similarity measure. At each retrieval iteration the operator is asked to select template tracks corresponding to the query identity (if any), and tracks corresponding to different identities; relevance feedback techniques are then used to update the feature weights and to re-rank the template gallery. In (12) a pre-trained classifier is used for each query identity, which however does not fit the application scenario where the query identities are known only during operation. The method of (13) is based on the observation that similarity in clothing appearance is often due to local rather than global body regions, especially in large template galleries; accordingly, starting from a ranked list of templates obtained using any descriptor and similarity measure, the operator is asked to select pairs of regions from the query and from some template images (chosen from predefined horizontal strips), labelling them as 'similar' or 'dissimilar'; the matching scores are then updated and the template gallery is re-ranked accordingly, in one or more iterations. Finally, in (14) a generic feature vector is used together with the negative Mahalanobis distance as the similarity measure; the operator is asked at each iteration to provide a feedback about a *single* template, with three possible labels: 'match' (if the template identity is the same as the query), 'similar' (different identities, similar clothing appearance), or 'dissimilar' (different identities, different clothing appearance); the similarity measure is then updated using an online metric learning approach to push the selected template to the bottom ranks (if the feedback label is 'dissimilar') or to the top ranks (if the feedback label is 'match' or 'similar'), and the template gallery is re-ranked. The main distinguishing feature is that the method of (14) is not query-specific, contrary to all the other ones: this means that the considered similarity measure is *incrementally* updated after each feedback (starting from the Euclidean distance), instead of starting from scratch at each new query. As a final, general remark on existing HITL methods, almost all works consider a simple kind of feedback suited to the HCV tool, except for the more complex kind of feedback of (13).

2.1.2 People search

This task is similar to person re-identification, except for the nature of the query: instead of an image (or track) of an individual of interest, only a textual description of its appearance is available. This is a common scenario in forensic investigations, e.g., when a witness of a crime provides a description of the perpetrator. The goal of a people search system is therefore to rank the template gallery for decreasing similarity to the textual description of the individual of interest (see Figure 2). Since in unconstrained VS footage automatic face recognition may not be feasible, clothing appearance should be considered as the main cue instead of face appearance, similarly to person re-identification.

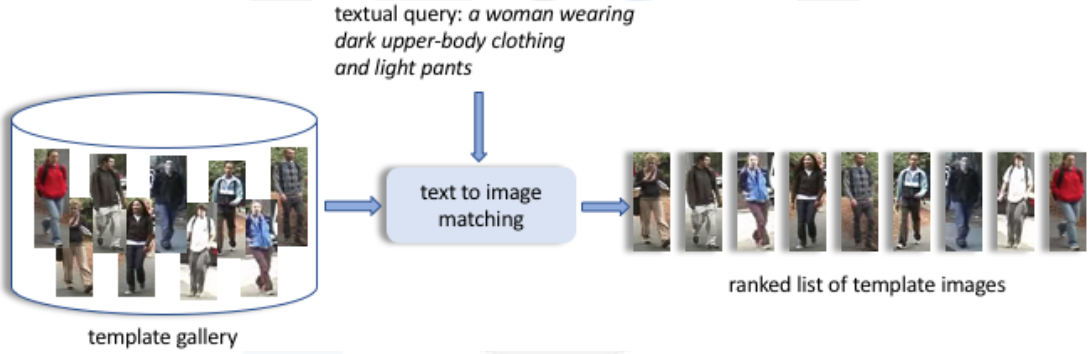

textual query: *a woman wearing dark upper-body clothing and light pants*
text to image matching
ranked list of template images
template gallery

Figure 2 – Core functionality of a people search system: ranking a given template gallery for decreasing similarity to a given textual description of clothing appearance and other possible attributes; ideally, all the templates whose appearance exactly matches the input description (if any) should appear at the top of the list.

The people search task has been introduced in 2011, more recently than person re-identification, and has been independently proposed in (15), (16).[2] The main, common feature of both works is that the description of the appearance of an individual is defined in terms of a predefined set of *attributes*. In (15) the following attributes were considered: gender; head colours (hair or hat); colours of upper body and lower body clothing (up to three colours can be specified); number, colour and type of bags, where type can be backpack, hand-carried bag, or rolled luggage. A probabilistic model is then defined to encode the relationship between an image of an individual and such attributes, whose parameters are set by a machine

[2] A similar task had been previously proposed in (70), based however only on face appearance.

learning technique based on a set of images of individuals labelled according to the presence or absence of each attribute. In (16) the following attributes were considered: colour of upper body and lower body clothing, short sleeves, short trousers/skirt. An automatic classifier was then trained to detect each attribute, on a labelled data set analogous to the one of (15).

So far no other work has considered the people search task. However, clothing appearance attributes have recently been used in several person re-identification methods (based on image queries), which were inspired by previous work on attribute detection in related CV tasks; in particular, attribute detection had been proposed for object recognition tasks, to complement low-level visual features with high-level, semantic information (17). Most of the existing attribute-based person re-identification methods combine low-level descriptors with the output of attribute detectors (either 'present'/'absent' Boolean values, or probabilistic values), which are usually built as in (16), as automatic classifiers using machine learning techniques (18), (19), (20), (21), (22). The only exception is (23), where only attribute information (obtained by CNNs) is used to match the query with the template gallery. The main difference between these methods is the classifier used (mainly support vector machines - SVMs - and CNNs), and the use of information about attribute co-occurrence to improve detection accuracy. Even if clothing appearance attributes are used in these works to perform image-based person re-identification, the underlying attribute detectors can also be used to perform text-based people search (although this task has not been considered in the mentioned works). Finally, also in (24) clothing appearance attributes were used to perform image-based re-identification, but attributes are inferred jointly with low-level features during the matching phase, which requires a query image beside a template: this means that attribute detection is obtained as a by-product of the matching process between a query and a template image, and therefore the underlying technique cannot be exploited for text-based people search (where no query image is available).

2.1.3 Crowd monitoring

The application of CV techniques to crowd monitoring encompasses a variety of related but distinct tasks such as people counting and crowd density estimation, tracking of individuals in a crowd, detecting patterns of crowd movement, and detecting anomalous crowd behaviours. Early work in this field dates back to more than 20 years ago. As stated in (25), the general goal is to **assist** humans in the difficult task of video monitoring, increasing the efficiency of crowd surveillance, and making it **proactive**; this definition points out the **semi-automatic**, not fully autonomous nature of tools developed in this field, due to the fact that the involved CV tasks are still very challenging.

In this section an overview of the main existing approaches is given, based on the main literature surveys published during the last decade (26), (27), (28), (29), (30), (31), (25). In particular, in (28) a detailed analysis is given of selected methods proposed up to 2013, focused on practitioners; a larger analysis including more recent works is reported in (30), (31), as well as an empirical comparison of some techniques on publicly available data sets.

A useful categorization of the existing crowd monitoring approaches can be made according to the specific task addressed (27):

- people counting and crowd density estimation;
- tracking individuals or groups in a crowd;
- understanding crowd behaviour, which can be further subdivided into:
    - detecting patterns of movement;
    - detecting anomalous behaviours;
    - detecting specific events or behaviours (e.g., panic, fighting, loitering, abandoned objects).

A second dimension refers to the scale to which a crowd is analysed. The main distinction is between the "microscopic" approach, based on detecting individuals in videos or frames/images (which is typically feasible only in low-density crowds); and the "macroscopic" (or "holistic") one, which considers a crowd as a whole (typically used for high-density crowds, when strong occlusions or far views do not allow reliable detection of individuals) (27). The following overview is structured around the above mentioned tasks.

People counting and crowd density estimation. This is one of the oldest tasks addressed in the literature. The specific survey of (29) identifies three approaches: counting by detection, by clustering and by regression. Counting by detection ("object-based" approach in (26), (27)) is based on the detection of the individuals in a crowd, e.g., through head contour detection, edge detection, or body models (27); it allows in principle accurate counting, and therefore also accurate density estimation. It is however feasible only for low density crowds with limited occlusions (27), (29), (25). Counting by clustering assumes that a crowd is made up of individual entities, each one characterized by unique yet coherent motion patterns; the number of people can therefore be estimated by detecting and clustering motion patterns. This approach is therefore based on people tracking, and thus it is feasible, as well as counting by detection, only for low density crowds with limited occlusion. Counting by regression is based instead on using machine learning techniques to infer a direct mapping from low-level image features to crowd density, avoiding explicit segmentation and tracking of individuals. It is specifically suited to dense crowds with severe occlusions. This approach does not allow accurate people counting, and is therefore suitable only for density estimation. In (27) it is named appearance-based approach, and it is further subdivided into pixel-based and texture-based methods. Pixel-based methods rely on local features (even extracted from individual pixels), obtained through background subtraction or edge detection. Texture-based methods use higher-level features computed from image patches (blocks of pixels). In (29) it is pointed out that regression-based methods have to deal with perspective distortion, which causes farther objects to appear smaller than the ones closest to the camera. This problem is addressed through geometric correction or perspective normalisation. The usual processing steps in regression-based methods are the following (29): (i) defining a region of interest in the whole image, to avoid processing image regions that cannot contain people; (ii) finding the perspective normalisation map of the scene; (iii) extracting holistic features; (iv) training a regression algorithm using the perspective normalised features.

Counting by detection and by regression methods can be applied to single frames or images, and are usually based on static appearance features; counting by clustering requires tracking of individuals, instead, and therefore requires video sequences and dynamic, motion features (30), (25). In (29) a useful discussion is given about the choice of features (including the combination of different kinds of features) and of the algorithms most suitable to the specific application scenario (e.g., in terms of the degree of illumination change in the scene), together with some guidelines; the issue of *local* density estimation (i.e., in different image regions) is also addressed.

Tracking of individuals or groups. Pedestrian detection and tracking is a well studied problem in CV, as a particular case of the more general problem of object detection and tracking. Detected trajectories can be exploited in related tasks like people counting (see above), identification of the main flows of a crowd, and detection of abnormal behaviours (27). Two of the main challenges are the presence of occlusions, which requires to solve the data association problem to recognize the same identity over different frames, and multi-target tracking, which is typical of crowded scenes; due to such challenges, the results of tracking algorithms are usually reliable only for low-density crowds. In particular, multi-target techniques can track people either assuming independence of their motion, or taking into account interactions between different people (26).

A widely used approach for tracking is the one based on the Particle Filter framework, an appearance-based technique proposed in (32) for single target tracking (based only on colour information), and then extended to multiple targets and to address crowded events in which people can exhibit similar clothing appearance, like in sports matches and celebrations (28). Several works exploited crowd-level cues to improve tracking, like high-level contextual information (e.g., background information on common direction of flow, in a structured crowd), and social interaction models that take into account that the reciprocal influence of the behaviour of nearby individuals (28). Most of the existing work focuses on tracking in low density crowds: only recently the problem of tracking individuals in dense crowds has being addressed, but the reliability of existing techniques is still unsatisfactory.

Group detection is another challenging problem, and requires accurate detection and tracking of individuals as a prerequisite. Currently it is less studied than other crowd analysis tasks (33), and only recently some promising results have been achieved (34). The main existing methods focus on low-density crowd scenarios, and exploit sociological models of human collective behaviour (33), (34). One open issue is that no agreed performance metrics exist yet (34).

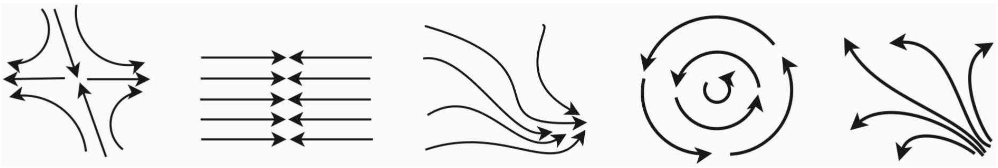
Understanding crowd behaviour. This task encompasses three main, distinct sub-tasks: detecting patterns of crowd movement (e.g., main directions and velocities of a crowd), detecting anomalous behaviours (e.g., unusual motions), and detecting specific events or behaviours (27). Approaches to each sub-task can be categorized also in this case into microscopic (object-based, or bottom-up) and macroscopic (holistic, or top-down) (27), (28), (30), (31). The former requires segmentation or detection of individuals, and is thus feasible only for non-dense crowds; the latter has been developed to deal with dense crowds, typically for outdoor scenes with wide field of view and low resolution for each target, where detection and tracking of individual targets is difficult or impossible (30).

Object-based methods have three main goals: detecting the dominant motion patterns; identifying groups of people, possibly using sociological models; modelling activities and interactions in crowded and complex scenes (27). Usually such methods are based on the analysis of trajectories of individual targets (28).

Holistic methods aim instead at detecting the global patterns of the crowd flow (27). Some of the existing methods exploit models taken from sociology or psychology, like the social force model of (35). Many works include the detection of abnormal behaviours at a macroscopic level (not related to individuals), and use cues like density estimation, e.g., to detect density changes (overcrowded scenes and excessive emptiness). On of the issues pointed out in (28) is that in unstructured crowds (made up of people relatively free to move in different directions) holistic methods usually fail to identify abnormal events related to the behaviour of a single individual (e.g., a running person in a walking crowd).

Detection of patterns of crowd movement. The main cue used in holistic approaches is optical flow, a dense field of instantaneous velocities computed between two consecutive frames, which is commonly used to extract motion features; spatio-temporal gradients are also used to model the regular movement of a crowd (28). A finer categorization is given in (30) into three kinds of crowd motion features: flow-based features (optical and particle flow), suitable for outdoor scenes with structured crowd; local spatio-temporal features, suitable to limited field of view with high crowd density; and higher-level trajectory ("tracklet"), based on tracking information, suitable for small- to medium-level crowd density and high resolution for single targets. In particular, some works exploit physics- and hydrodynamics-based models (28), (30).

One example of holistic (top-down) method is the one proposed in (36), which is aimed at detecting five specific, medium-level patterns of crowd movement (in contrast to high-level, semantic patterns like 'panic'), named bottlenecks, fountainheads, lanes, arches, and blocking patterns (see Figure 3, top); an interesting feature is that this method does not require object detection or tracking, nor it requires training. Similar patterns of movement are detected by the method by (37): lane, clockwise arch, counter-clockwise arch, bottleneck and fountainhead patterns (see Figure 3, bottom).



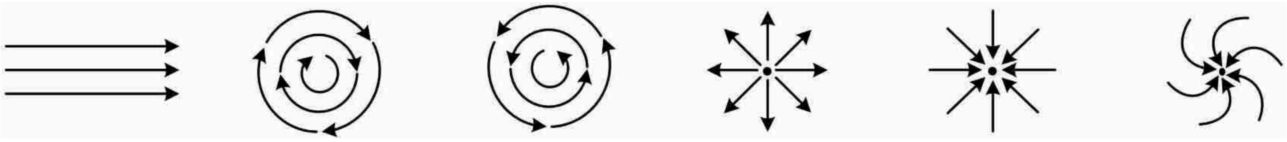


Figure 3 – Top: patterns of crowd movement detected by the method of (36) (from left to right: blocking, lane, bottleneck, ring/arch, fountainhead). Right: patterns detected by (37) (from left to right: lane, clockwise arch, counterclockwise arch, fountainhead, and two examples of bottleneck). Figures taken from the cited works.

Detecting anomalous behaviours. This is a very challenging task due to the high variability of the concept of "anomaly", which is often subjective and strongly application-dependent (38), (28), (30), (31). A useful reference for this task is the survey of (38), which is focused on anomaly detection in crowd behaviour and on real-time implementations.

Also for this task the existing methods can be subdivided into object-based (when anomalies are defined in terms of the behaviour of single individuals) and holistic (when anomalies refer to the global crowd motion), which often use optical flow. In (38) three usual assumptions underlying existing methods are identified: (i) anomalous events are much less frequent than normal events; (ii) they exhibit significantly different characteristics from normal events; (iii) anomalous events have a specific meaning. Assumption (i) implies that representative information of anomalous events may be lacking or missing. If either of the first two assumptions is not satisfied, false positive detections can occur (e.g., rare normal events misclassified as anomalies). Assumption (iii) corresponds to the fact that "anomalies" usually refer to specific events, related to the particular application scenario (e.g., the kind of mass gathering event).

Assumptions (i) and (ii) above motivate the anomaly detection approach, which is widely used in different applications like intrusion detection in computer networks; it consists of building a model of the "normal" behaviour, and of detecting any deviation from this model as a potential anomaly. The model of normal behaviour is usually built using machine learning (data-driven) techniques, either using only data representative of the normal behaviour, or also data representative of anomalous behaviours (if any). If the available data is not labelled as 'normal' and 'anomalous', an *unsupervised* learning approach can also be used: data are clustered assuming to be mostly representative of normal behaviour, and then outliers are considered as anomalies. In (30) existing methods are also categorized into global and local anomaly detection, depending on whether they detect only the occurrence of some anomaly in a scene (global) or also where the anomaly is taking place (local). A further categorization of local methods is between the ones based only on CV techniques (in particular, on visual features), and methods inspired by physics, which employ physical models for crowd dynamics representation.

An example of holistic, or top-down method for abnormal behaviour detection (defined as deviation from a normal model) is (39); in addition to optical flow it uses the social force model of pedestrian dynamics by (35), which has then been used in several subsequent works.

Detecting specific events or behaviours. Some authors proposed methods for detecting specific behaviours such as panic, fighting, and abandoned objects. These are however challenging tasks due to the difficulty of collecting sufficiently representative examples (videos) to train machine learning techniques, and the need of accurate tracking results. In particular, tasks like abandoned object detection also require a complex analysis of the interactions between different targets. For instance, the method of (40) was developed in the context of the project SUBITO[3] (Surveillance of Unattended Baggage and the Identification and Tracking of the Owner), funded by the European Commission under FP7-Security, and spanning from 2009 to 2011; performances reported in (40) were considered by the authors not yet acceptable for a deployed threat assessment system.

[3] https://cordis.europa.eu/project/rcn/89391_en.html

Interactions between computer vision-based crowd monitoring and crowd modelling. One of the main issues common to CV techniques for several crowd monitoring tasks is the collection of representative (and labelled) data for training machine learning techniques, and for validation (27). In particular, the lack of representative data is critical for anomalous behaviour detection, as explained above. An interesting solution, in the context of LETSCROWD, is the exploitation of data generated by computer graphics using **crowd modelling (simulation) techniques**, which has been proposed by several authors (26), (27), (30). One of the key advantages of crowd simulation algorithms is that they allow to simulate *controlled* scenarios, avoiding the tiresome, time-consuming and error-prone task of manual labelling images or video sequences; the simulation of unpredictable behaviours remains however an open issue. On the other hand, information obtained from CV algorithms during operation (even in real time) can in turn be exploited to improve the realism of crowd simulation algorithms; for instance, the detected crowd spatial distribution (local density) could be used to initialize a crowd simulator, and the detected people trajectories or main flow directions could be used to guide the motion of virtual agents (27).

2.2 EXISTING STATE OF THE ART

Overall, the current technology readiness level (TRL) of existing methods for the considered CV tasks is mostly TRL3 (experimental proof of concept), although for some specific crowd monitoring tasks commercial solutions (mentioned below) already exist, mainly for people counting, crowd density estimation, and pedestrian tracking; they seem however focused on sparse crowds. The state of the art of the three CV tasks involved in the HCV tools is analysed in the following subsections, including the performance reported in the literature, the available data sets, and the commercial solutions.

2.2.1 Person re-identification and people search

For the person re-identification and people search tasks, as pointed out in the literature review (Sect. 2.1) existing work focused on descriptor and similarity measure definition, disregarding the issue of template gallery population. In almost all works (especially early ones) experimental evaluations are made on publicly available, benchmark data sets of pedestrian images obtained by *manually extracting* bounding boxes. However, whereas the bounding box of the query image (in the case of person re-identification) has to be manually selected by the operator using a suitable GUI (41), in a real-world "end-to-end" system all the template images have to be *automatically* extracted by a pedestrian detector or tracking tool from video sequences or frames; the corresponding bounding boxes should be automatically generated as well. Inaccurate detection or bounding box extraction can negatively affect re-identification/search performance. Only recently a few data sets including automatically extracted bounding boxes have been made publicly available (see Sect. 2.4), and a few works have evaluated the performance of person re-identification systems by automatically extracting template images from raw videos or frames (42), (41).

The pedestrian detection and tracking tasks have been addressed separately in the CV literature, as a particular case of the more general object detection and tracking tasks, and a number of approaches and algorithms have been proposed so far (see, e.g., (43)). Several open source implementations of pedestrian detection and tracking algorithms are also available, which can be directly exploited in the context of the HCV tool (e.g., Faster R-CNN[4] (44)), as well as some commercial solutions. However, also for these tasks existing algorithms and solutions achieve satisfactory results only under relatively simple application scenarios (e.g., sparse crowds and limited overlapping). For instance, a very recent work published in a top journal in the CV and pattern recognition field (45) provided empirical evidence that top pedestrian detectors at the state of the art still exhibit a **ten-fold gap** with respect to human performance in real-world, challenging application scenarios.

Another critical issue is the template gallery size. Up to a few years ago, benchmark data sets were limited to hundreds or a few thousands pedestrian images, corresponding to tens or a few hundreds identities. This is not representative of real-world VS scenarios, and in particular of mass gathering events. Currently, a few

[4] https://github.com/rbgirshick/py-faster-rcnn

data sets containing tens of thousands images, corresponding to hundreds or a few thousand identities are also available (see Sect. 2.4). Dealing with large template galleries exhibits however two challenging issues. One is the processing cost required to build the descriptor of each template image, and to match each template to the query image, which has not been addressed by almost any work in the literature. Existing work focused indeed on achieving high re-identification accuracy, disregarding processing time; many of the proposed methods are however rather complex, and are likely to be not suitable for large template galleries or when real-time performance is required. A second issue is that the template gallery size strongly affects re-identification performance: for instance, in (14) it is shown that a ten-fold increase in gallery size can lead to a ten-fold decrease in the rank-1 accuracy of a given method, i.e., on the capability of placing the template image of the query identity (if any) in the top position of the ranked list of templates. This means that re-identification performances reported in the literature on relatively small data sets are likely to be far too optimistic for real-world application scenarios involving much larger template galleries.

To our knowledge, the only work presenting an end-to-end, prototype person re-identification system for a real world application is (46), where the authors describe a tag-and-track surveillance system developed for the medium-sized Cleveland Hopkins International Airport (Ohio, USA). The system was designed to assist the transportation security administration officers monitoring the airport using their existing surveillance camera network. It consisted of three cameras, one located just after a security checkpoint in which the subject of interest is tagged, and two cameras located at the entrances to different concourses, in one of which the subject will reappear. The entire system had to operate in real time using the airport's network infrastructure. In particular, differently from the cameras used in benchmark re-identification data sets, the available cameras (as in many airports) were oriented at sharp angles to the floor of about 45 degrees; moreover, they were analogue cameras, whose feeds were converted into the H.264 standard with a 704x408 resolution at about 30 frames per second (fps). Potentially useful suggestions are reported in (46) on the design, development and deployment of person re-identification systems for real-world application scenarios.

2.2.2 Crowd monitoring

The **crowd monitoring** task still exhibits many open issues. One is that pedestrian tracking, which is used in the object-based or microscopic approach for several tasks (like people counting by clustering, and anomaly detection based on trajectory analysis) is still a challenging problem in CV, especially in crowded scenes, due inter-object occlusion. Reliable analysis of crowd behaviour in this scenario is therefore achievable only at the macroscopic level, in terms of global motion patterns; however, this task is in turn difficult in the case of unstructured crowds, whose motion appears to be random (28).

Another issue is pointed out by (30): usually, existing methods which require tracking, learning and behaviour detection/recognition carry out these steps by simple integration of the corresponding functional modules, disregarding their interactions. Simultaneously carrying out these tasks could allow to fully exploit the underlying, hierarchical contextual information.

On the other hand, similarly to person re-identification, most of the existing methods focus on accurate scene understanding; they are however rather complex and do not guarantee real-time crowd analysis, which remains an open issue (31), (30).

An open issue specific to anomaly detection is the lack of a sound testing methodology to compare different algorithms and approaches, which is consistently pointed out in the existing survey papers (38), (30), (31), (25). Many authors used different, and sometimes non-publicly available datasets for evaluation, which makes the reported results not comparable. Under a practical viewpoint, this also makes it difficult to choose a suitable method for a given application scenario, based on information reported in published papers. Moreover, reported evaluations are often made in a single environment or in controlled conditions, and therefore do not provide information about the capability of the considered methods to generalize to new environments and to different scales, as well as their robustness to unexpected events, which would be critical for the deployment in real-world scenarios (38).

As a last remark, no evidence of the use of the HITL approach in crowd monitoring work has been found.

2.2.3 Performance of state-of-the-art computer vision algorithms

As mentioned in Sect. 2.2.1 and 2.2.2, publicly available data sets used so far by the research community to evaluate the performance of CV algorithms for the considered tasks exhibit several limitations that often make them not representative of real application scenarios. These limitations are described in more detail in Sect. 2.4. Moreover, in fields like person re-identification the authors tended to focus over the years on improving performance on the same, few benchmark data sets, typically devising more and more complex methods, but this does not guarantee similar improvements on different application scenarios (see (47) for the CV field, and (48) for a more general discussion of this issue in the machine learning field). As a consequence, there is no guarantee that performances reported in the literature for specific CV tasks can be attained in different real-world settings where the same algorithms can be deployed. A further issue is that the performance of different methods is often difficult to compare, since authors often evaluate their own methods on different (although not disjoint) subsets of the available data sets, sometimes using different experimental settings.

Bearing the above premise in mind, in this section examples of the performances reported in the literature are given for each of the considered CV tasks, with the caveat that they cannot be considered indicative of the performance that can be attained by the HCV tool in real application scenarios.

Person re-identification. A useful source is the very recent work of (3), where the largest empirical evaluation of person re-identification algorithms was carried out: several hundreds algorithms were evaluated, including different combinations of features, similarity measures and metric learning algorithms. As an example, we consider the two benchmark data sets VIPeR and Airport (see Sect. 2.4). VIPeR is one of the oldest, but still widely used and challenging benchmarks, despite its size is relatively small (it is made up of a pair of images of 632 different individuals acquired by two cameras) and despite its images consist of manually extracted bounding boxes. Airport is instead one of the most recent benchmarks, is larger than VIPeR (it contains about 40,000 images of 1,382 different individuals), and contain only bounding boxes automatically extracted by pedestrian detection algorithms. Results reported in (3) show that for both data sets the **first-rank recognition rate** (i.e., the estimated probability that a template image of the same identity as the query image is returned in the first position of the ranked list of templates) ranges **from less than 0.05** for the worst performing algorithms (out of 276 evaluated algorithms) **to more than 0.40**. For other data sets the range of the performance is even higher. Moreover, the performance of a given algorithm can change considerably depending on the data set; for instance, the best performing algorithm found in (3) was different for VIPeR and Airport (as well as for other data sets). To conclude, it is worth quoting here a sentence from (3) summarizing the observed results: "*the 'spread' in the performance of the algorithms for each dataset is huge, indicating the progress made by the re-id community over the past decade. However, on most datasets, the performance is still far from the point where we would consider re-id to be a solved problem.*"

People search. Much fewer empirical evidences are available for this task, which has been addressed so far by only a few authors. In particular, for people search there are no well-defined performance measures, since the degree to which the template images are "correctly" sorted by decreasing similarity to the query individual, in terms of clothing appearance and other attributes, cannot be objectively assessed. A reasonable proxy is the performance of attribute detectors, which can be quantitatively evaluated as the fraction of pedestrian images for which the presence or absence of a given attribute is correctly detected. Detection accuracy depends on the specific algorithm used to implement attribute detectors, and on the data set used to train and evaluate them. It also depends on the number of images in the data set in which the considered attribute appears: the larger it is, the higher the detection accuracy. For instance, in the two data sets used in the attribute-based re-identification method by (18) (the same data set VIPeR mentioned above, and the PRID data set) the detection accuracy on 21 attributes evaluated on VIPeR ranges from 54.5 for "has handbag carrier bag" to 84.0 for "dark shirt"; for PRID it ranges from 31.3 for "no coats" to 81.6 for "light shirt". Moreover, for some attributes a very different detection accuracy was reported on the two data sets, e.g., 0.81 in VIPeR and 0.41 in PRID for "red shirt"; 0.60 in VIPeR and 0.75 in PRID for "dark hair". In the larger PETA data set (including VIPeR and PRID, see Sect. 2.4) used by (49), an average accuracy of 0.70 was reported over 35 attributes, with a minimum of 0.50 for "sandals" and a maximum of 0.88 for "muffler".

Crowd monitoring. We separately consider the different tasks discussed in Sect. 2.1.3. For ***crowd counting / density estimation*** we consider the literature survey of (29), where experiments on three data sets (UCSD Pedestrian Traffic, QMUL Mall and PETS 2009, see Sect. 2.4) and six different algorithms are reported. For benchmark data sets in which the exact number of people present in each image or video frame is available (evaluated by a human), one of the performance measures is the mean deviation error (MDE) over the considered images:

$$MDE = \frac{1}{N}\sum_{n=1}^{N}\frac{|y_n - \hat{y}_n|}{y_n},$$

where N is the number of considered images, y_n the actual number of people in the n-th image and $\hat{y}_n$ the estimated count; in other words, MDE evaluates the relative error with respect to the actual count. In experiments carried out on training and testing images exhibiting a similar crowd density (either sparse or dense), MDE ranged from 0.075 to 0.175 in sparse crowd scenes, and from 0.055 to 0.2 in dense crowd scenes. To test generalisation capability to unseen density (a typical scenario in real applications), other experiments were carried out by training on sparse and testing on dense scenes, and vice versa; the corresponding MDE ranged from 0.086 to 0.264, and the best algorithm achieved an MDE of 0.217. This means that in real application scenarios estimates provided by state-of-the-art algorithms can exhibit a relative error over 0.20.

For the ***detection of patterns of movement*** we consider the methods of (36), (37) mentioned in Sect. 2.1.3, where experiments on the UCF Crowd Segmentation, PETS 2009, and CUHK Crowd data sets (see Sect. 2.4) were carried out, including real video sequences with high crowd density. On the considered patterns of movement (see Figure 3), a detection rate (or true positive rate, TP) of 0.8 was reported by the authors of these papers, with a false positive (FP) rate from 0.05 to 0.2, depending on the specific pattern.

For ***individual or group tracking*** we consider the recent method of (50), focused on tracking individuals in dense crowds. The evaluation was carried out on eight real video sequences showing commuters walking outdoors (hundreds of people), marathon runners (hundreds of people), and railway stations (tens of people). Tracking accuracy was evaluated as the fraction of pixels in all the detected tracks that lie within a given number T of pixels to the actual (manually annotated) track. For $T = 15$ pixels, the algorithm proposed by (50) attained an accuracy of 0.67 to 1.0 on the different sequences.

For ***group detection*** we consider the method of (34). Five data sets made up of real videos representing different kinds of scenes were considered: low crowded scenes outside a university and at a bus stop; medium density crowd outside a university; medium density crowd in a shopping arcade; and heterogeneous scenes of varying crowd densities showing people walking in a mall, crossing the street or participating at events. Performance was evaluated in terms of the complementary precision (Pr) and recall (Re) metrics, defined respectively as the fraction of correct group detections among the detected groups, and the fraction of correctly detected groups among all the actual groups. In experiments carried out using an automatic pedestrian detector and tracker, observed performance ranged from $Pr = 0.75$, $Re = 0.71$, to $Pr = 0.81$, $Re = 0.80$.

For the task of ***anomalous behaviour detection*** it is difficult to define benchmarks, due to the high variability of the concept of "anomalous behaviour" which depends on the application scenario, and can also exhibit subjective aspects. Here we consider the performance reported in some works on the few benchmark data sets available, which was evaluated in terms of TP and FP rates over manually defined anomalous behaviours; unless otherwise stated, we report the TP rate achieved for a reference FP rate equal to 0.1. In (39) experiments on the on UMN Crowd and UCF Web data sets were carried out (see Sect.

2.4): the accuracy varied depending on crowd density; e.g., on videos with low density crowd a TP rate above 0.8 was reported, whereas for high density crowds the same TP rate was achieved only for a FP rate around 0.4. In the more recent work of (51), experiments were carried out on a *different* data set, UCSD Anomaly Detection (see Sect. 2.4), which contains *different* kinds of anomalies with respect to the data sets used by (39); the reported TP rate was 0.6. In (52) the experiments were carried out on four data sets, including UMN Crowd and PETS 2009 (described in Sect. 2.4), and a data set of synthetic scenes; the reported TP rate was 0.87 on UMN Crowd, and 0.95 on PETS 2009. In (53) a TP rate of 0.50 was reported on the UMN Crowd data set, and a TP rate of 0.7 on UCSD Anomaly Detection.

Finally, for the ***detection of specific events or behaviours*** we consider as an example the "abandoned object" event. This kind of event was addressed by the EC-funded project SUBITO mentioned in Sect. 2.1.3. Some of its results have been published in (40), where experiments on two data sets (one of which specifically produced during the project) have been carried out, and the performance has been evaluated using the Pr and Re metrics. Under three different definitions of the event of interest, the reported performance ranged from $Pr = 0.46$, $Re = 0.15$ to $Pr = 0.59$, $Re = 0.36$. This means that among the events detected as "abandoned object" by the proposed algorithm, only a fraction of the detections from 0.46 to 0.59 were actually correct; moreover, among all such events present in the videos, only a fraction from 0.15 to 0.36 were detected. These results were judged in (40) as being "below acceptable performance for a deployed threat assessment system."

2.2.4 Available software

For person re-identification and pedestrian detection and tracking, a few authors made a software implementation of their methods available, either on their personal web pages or in software sharing platforms like GitHub.[5] Most of this software is written in Matlab and in Python, except for pedestrian detection where C++ is also often used (probably due to efficiency reasons, and to the fact that pedestrian detection is an older research topic than re-identification). A few such implementations are in the form of software libraries. However the main purpose of the available software is to allow other researchers to carry out experimental evaluations of the corresponding methods, and often the structure of the code itself is tailored to this purpose. The available code is therefore not directly usable for developing demos and prototypes, and requires substantial reworking to this purpose, as well as the integration between different components (e.g., pedestrian detection and re-identification modules). For the HCV tool, the only useful resources turned out to be related to the pedestrian detection and tracking (e.g., the Faster R-CNN object detection algorithm of (44) mentioned above). No software implementations are available for crowd monitoring tasks, instead.

2.2.5 Commercial solutions

Some companies offer video analytics solutions for VS systems targeted to security-related applications, which include crowd monitoring functionality:

- Vision Semantics Ltd[6] is a spin-out company of Queen Mary University of London established in 2000. It develops self-configuring video analysis and dynamic scene understanding tools and applications, with two main functionalities: multi-camera tracking, with a forensic tool for re-identifying and back-tracking individuals; and analysis of crowd dynamics and behaviours in a public space, including people counting and crowd density estimation, crowd profiling based on distribution of crowd over space and time, and crowd event detection and tracking. Expertise is mentioned on the detection of abnormal behaviours both with automatic learning of visual context to update the normal behaviour model, and with incorporation of some human feedback to enhance behaviour model learning. No demos or videos of their products are available.

[5] https://github.com

[6] http://www.visionsemantics.com

- iOmniscient[7] is an Australian video analytics company founded in 2001. It offers software solutions for VS tailored to several scenarios such as banks, airports, schools, roads and traffic monitoring. These include specific solutions for police (iQ-Police) and for crowded events (iQ-Event). Their solutions are collections of tools which implement functionalities like: people counting, both for sparse and dense crowds, and detection of sudden crowd gathering; pedestrian tracking in sparse crowds; detection of some events ("suspicious behaviours") such as slip and fall, man-down, loitering, and running; left object detection; and forensic capabilities (not further specified) to process large archived video. A few demos (videos) are available, only for abandoned object detection, and for people counting in a marathon scenario and in a train station.
- Ipsotek[8] is a U.K. company established in 2001. It develops scenario-based video analytics solutions tailored to individual clients, including investigation, forensics, and crowd management functionality. Crowd management functionality consists of people counting and crowd density estimation. Investigation and forensic functionality includes tracking of 'tagged' people (i.e., chosen by an operator from a video frame), and real-time content based video retrieval, where the search is based on colour, shape, location, speed and other behavioural features (not further specified) of targets. A few videos showing people tracking and crowd density estimation functionality are available, in small crowd scenarios (up to about twenty people).
- CrowdVision[9] is another U.K. company, founded in 2007, offering solutions targeted to management of airports, retail, convention centres and transport hubs. Although its solutions are not related to security, they include tools for people counting, people tracking, and detection of people flow. No demos or videos of their products are available.

Several other companies provide VS solutions, including functionality like face recognition, object tracking, object retrieval, etc., but not targeted to crowded scenes.

2.3 DEVELOPMENT REQUIRED

2.3.1 How we will go beyond the current state of the art

The advancement over the state of the art in the considered CV task will not be pursued in terms of improving the performance (accuracy) with respect to the one reported in the literature. The main reasons are that most of the existing work aim only at improving accuracy on benchmark data sets, disregarding other issues like complexity of implementation (including the need of tuning many parameters) and processing time; moreover, publicly available, benchmark data sets are not representative of real-word application scenarios (as discussed in Sect. 2.2), which make the reported accuracy to be significantly higher than that achievable in such scenarios.

The advancement over the state of the art will be pursued in the following aspects, instead:

- The HCV tool will work end-to-end: it will take as input raw videos, and will produce outputs in terms of easy to interpret, actionable information to LEA operators. Inputs will be acquired directly from video cameras or from pre-recorded videos, depending also on the tool and application scenario (e.g., pre-recorded videos will be used for the post-event, forensics use case). Outputs will be shown through suitable graphical user interfaces (GUIs), described in Sect. 4. In particular, for the person re-identification and people search tools, developing end-to-end systems will require the integration of different components usually considered separately in the literature, like pedestrian detection and bounding box extraction to populate the template gallery.

[7] http://iomniscient.com

[8] https://www.ipsotek.com

[9] https://www.crowdvision.com

- The CV techniques for the considered functionality of the HCV tool will be chosen to guarantee a processing cost suitable to real-time implementation, taking also into account the need of managing large template galleries in the person re-identification and people search tools. This may require a trade-off with the accuracy of the tools.
- The HITL feature will be implemented in the person re-identification and people search tools, and possibly in the crowd monitoring tool, to allow them to adapt to the characteristics of each specific application scenario by exploiting the operator's feedback. This feature has been considered so far only by a few authors, and only in the person re-identification task.
- A distinctive feature of the HCV tool, in the context of LETSCROWD, will be its integration with the CM tool developed in task T5.1 (deliverable D5.1), and with the DRA tool developed in work package WP3 (deliverables D3.2 and D3.4). In particular, a two-way integration with the CM tool will be pursued: currently it is envisaged that some outputs of the crowd monitoring tool will be used as input by the CM tool (as well as the DRA tool), and that videos of crowd simulations produced by the latter can be used for the development and validation of the former. The integration is described in Sect. 4.2.

2.3.2 What will be achieved

The HCV tool will consist of three integrated tools aimed at supporting LEA operators both during event execution and in post-event forensic investigations:

- **a person re-identification tool**, to search for an individual of interest, based on clothing appearance, among videos acquired by the cameras of a video surveillance network, using a *query image* of that individual;
- **a people search tool**, to search for an individual of interest among videos acquired by the cameras of a video surveillance network, using a *textual description* of the clothing appearance and other attributes of that individual;
- **a crowd monitoring tool**, made up in turn of sub-tools that will implement three main functionalities:
    - density estimation of a crowd;
    - detection of patterns of movement of a crowd, in terms of main directions and velocities;
    - detection of anomalous crowd behaviours in terms of the two kinds of information above, i.e., density (static or dynamic) and patterns of movement;

The possibility to develop additional functionality (sub-tools) in the crowd monitoring tool will be also evaluated, including:

- detection of groups of people inside a crowd (group formation/breaking up);
- detecting of anomalous or suspicious behaviours of groups (e.g., sudden group formation, groups of people wearing similar attire).

As mentioned in Sect. 2.3.1, the crowd monitoring tool will be integrated with the CM tool developed in task T5.1, and with the DRA tool developed in work package WP3.

According to the DoW, these tools will move from TRL3 of existing methods to TRL5 (technology validated in relevant environment).

2.4 PUBLICLY AVAILABLE DATA SETS

Several image and video data sets have been made publicly available over the years by the research community for the three CV tasks of person re-identification, people search and crowd monitoring. In default of data provided by LEA partners of LETSCROWD, some of these data sets will be selected for developing and validating the HCV tool.

Person re-identification. More than 20 data sets are currently available, released since 2007. An updated and detailed list is maintained by the Robust Systems Lab at the Northeastern University (Boston, USA).[10] These data sets were acquired in different scenarios, such as streets, campuses, stations, shopping malls, and airport halls. All data sets provide images of single pedestrians in the form of bounding boxes extracted from video frames, and manually labelled according to their identity; more precisely, inside each data set a unique ID is associated to every identity, and each image is labelled according to the corresponding ID. For most data sets bounding boxes are manually extracted; only in eight data sets they are automatically obtained from pedestrian detectors. These data sets exhibit different characteristics in terms of:

- number of pedestrian images: from hundreds to tens of thousands, with the exception of one data set (MARS) containing over one million images;
- number of identities: from tens to thousands;
- number of cameras: from one to sixteen;
- image (pedestrian bounding box) size: variable in the majority of cases, and equal to 128x64 for some data sets;
- beside pedestrian bounding box images, in some cases full video frames (for about half of the data sets), or tracking sequences of individuals (for a few data sets), or video sequences (for about ten data sets) are also available.

An example taken from the well-known VIPeR[11] data set is shown in Figure 4.

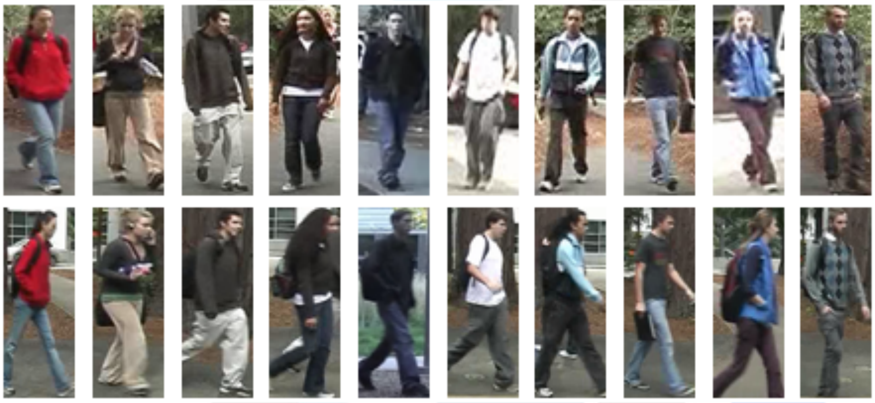


Figure 4 – Example of images from the VIPeR person re-identification data set: pairs of images (128x48 pixels, bmp format) of ten different individuals taken from two different cameras.

The available data sets exhibit several limitations:

- For most data sets only manually extracted pedestrian bounding boxes are available (as pointed out in Sect. 2.2), which does not reflect real-world application scenarios when an automatic pedestrian detector should be used to populate the template gallery. Only in a few data sets automatically extracted bounding boxes (with different pedestrian detectors) are provided, including 'false positives' (bounding boxes which do not contain a pedestrian) to mimic real scenarios.

[10] http://robustsystems.coe.neu.edu/sites/robustsystems.coe.neu.edu/files/systems/projectpages/reiddataset.html

[11] https://vision.soe.ucsc.edu/node/178

- Several data sets do not include the source video sequences, and thus cannot be used to test re-identification systems that process videos as inputs, as in typical application scenarios. For a few data sets this limitation is partially overcome by the availability of full video frames.
- Most of the data sets contain a relatively small number (tens or a few hundreds) of different individuals (identities), which is not suited to application scenarios like the one relevant to LETSCROWD, i.e., mass gathering events.

Among the available data sets, only the four ones listed in Table 1 can be considered suitable to the development and evaluation of the HCV person re-identification tool, in terms of the number of different identities they contain, and of the availability of full video frames or video sequences.

Data set	Description
PRW[12] (Person Re-identification in the Wild) (41)	Acquired by 6 cameras on a University campus. Identities: 932; video frames are available; bounding boxes: manually extracted.
MARS (Motion Analysis and Re-identification Set) (54)	Acquired by 6 cameras in a University campus (same as PRW). Identities: 1,261; bounding boxes: automatically extracted.
DukeMTMC-4ReID[13] (55)	Acquired by 8 cameras on a University campus. Identities: 1,413; videos are available; bounding boxes: automatically extracted.
Airport[14] (46)	Acquired by 6 cameras installed post a central security checkpoint at an active commercial airport within the United States, by researchers at Northeastern University and Rensselaer Polytechnic Institute, affiliated to ALERT (Awareness and Localization of Explosives-Related Threats), a multi-university Department of Homeland Security Center of Excellence. Identities: 9,651; videos are available; bounding boxes: automatically extracted.

Table 1 – Publicly available data sets for person re-identification suitable to the HCV tool.

People search. Four data sets are available for attribute-based person re-identification with *image* queries, made up of pedestrian images (bounding boxes) acquired in VS settings and manually annotated according to the presence or absence of a predefined set of attributes. Their main features are reported in Table 2.

Data set	Description
VIPeR[15] (18)	1,264 outdoor images acquired by 2 cameras; image resolution 128x48; 21 binary attributes.
APiS[16] (Attributed Pedestrians in Surveillance) (56)	3,661 outdoor images; image resolution 128x48; 11 binary attributes.
PETA[17] (PEdesTrian Attribute) (49)	19,000 indoor and outdoor images collected from ten re-identification data sets, with varying camera angle, view point, illumination, and resolution (from

[12] http://www.liangzheng.org/Project/project_prw.html

[13] http://vision.cs.duke.edu/DukeMTMC/

[14] http://www.northeastern.edu/alert/transitioning-technology/alert-datasets/alert-airport-re-identification-dataset/

[15] http://www.eecs.qmul.ac.uk/~rlayne/

[16] http://www.cbsr.ia.ac.cn/english/APiS-1.0-Database.html

[17] http://mmlab.ie.cuhk.edu.hk/projects/PETA.html

Data set	Description
	39x17 to 365x169); 61 binary attributes plus 4 attributes with 11 values each.
RAP[18] (Richly Annotated Pedestrian) (57)	41,585 indoor images from 26 cameras; image resolution from 92x36 to 554x344, 69 binary attributes.

Table 2 – Publicly available data sets with annotated pedestrian attributes, suitable to the HCV people search tool.

These data sets can be directly exploited also for training attribute detectors of the HCV people search tool, where the query consists of a description of clothing appearance in terms of an attribute profile, i.e., a suitable combination of the available attributes. Similarly to person re-identification data sets (with no annotated attributes), the main limitation of these data sets is that pedestrian bounding boxes are manually extracted. Among the four data sets of Table 2, the most suitable to the people search tool is PETA, as it is the largest one that contains bot indoor and outdoor images (RAP is larger, but contains only indoor images), and exhibits a large variability in camera angle, view point, illumination, and resolution. PETA also includes 15 attributes (12 are appearance-based, and 3 are soft-biometrics) chosen in (18) on the basis of **operational procedures of human experts**:[19] *shorts*, *skirt*, *sandals*, *backpack*, *jeans*, *logo*, *v-neck*, *open outerwear*, *stripes*, *sunglasses*, *headphones*, *long hair*, *short hair*, *gender*, *carrying object*.

Crowd monitoring. Several data sets related to different crowd analysis applications are currently available. The ones potentially useful for developing and validating the HCV tool are summarized in Table 3, including the crowd monitoring task(s) for which they can be used (i.e., for which manual annotations are provided). Most of these data sets have been surveyed in (30), (31), (25). In particular, we point out the **synthetic** data set Agoraset, developed according to the approach discussed in Sect. 2.1.3 to overcome the drawbacks of real videos; it however exhibits several limitations pointed out by the authors: very simple environment depicted with a uniform shaded colour, mostly flat scenes with no objects beside people, limited variability in the geometric appearances and textures of people, very simple crowd motion model.

Data set	Description
UCSD Pedestrian Traffic[20] (58)	Tasks: **crowd counting / density estimation**, and **pedestrian tracking**. Made up of videos of pedestrians on walkways at the University of California, San Diego (UCSD), taken from a stationary camera, with 8-bit grayscale, 238x158 pixels, 10 fps; region-of-interest (ROI) and perspective map are included.
QMUL Mall[21] (29)	Tasks: **crowd counting / density estimation**. Collected from a publicly accessible webcam in a mall by researchers of the Queen Mary University of London (QMUL). Made up of 2000 jpeg frames of 640x480 pixels, extracted from videos at about 1 fps. Frames are exhaustively annotated by labelling the head position of every pedestrian (over 60,000 annotations) and the number of pedestrians in all frames; the perspective map is included.
UCF Crowd Counting[22] (59)	Tasks: **crowd counting / density estimation**. Made up of 50 publicly available web images (mainly from Flickr) collected by researchers of the University of

[18] http://rap.idealtest.org/

[19] The following source was cited by [Layne-2014], but has not been found online at the time of writing this document: T. Nortcliffe, People Analysis CCTV Investigator Handbook, Home Office Centre of Applied Science and Technology, UK Home Office (2011).

[20] http://www.svcl.ucsd.edu/projects/peoplecnt/

[21] http://personal.ie.cuhk.edu.hk/~ccloy/

[22] http://crcv.ucf.edu/data/crowd_counting.php

Data set	Description
	Central Florida (UCF). Images are related to different kinds of events (concerts, protests, stadiums, marathons, and pilgrimages), with a number of people between 94 and 4543, with an average of 1280 individuals per image.
WorldExpo'10 Crowd Counting[23] (60)	Tasks: **crowd counting / density estimation**. Made up of videos collected at the Shanghai 2010 WorldExpo: 1132 annotated video sequences captured by 108 surveillance cameras, mostly with disjoint bird views, covering a large variety of scenes. A subset of video frames is annotated with regions of interest, positions of pedestrian heads, and perspective map.
UCF Crowd Segmentation[24] (61)	Tasks: **detection of patterns of crowd movement**, and of **anomalies**. Made up of videos collected by UCF researchers of the mainly from the BBC Motion Gallery and Getty Images websites. Video frames are segmented according to dominant crowd flows, and to detected abnormalities in such flows. Beside crowds, other high density moving objects are present.
UCF Web[25] (39)	Tasks: **anomalous behaviour detection in crowds**. Made up of high-quality videos collected by UCF researchers from sites like Getty Images and ThoughtEquity.com in different urban scenes: 12 sequences of normal crowd scenes (pedestrian walking, marathon running, etc.) and 8 scenes of escape panics, protesters clashing, and crowd fighting as abnormal scenes. Abnormal scenes are taken from old b/w movies or documentaries, and from the fiesta of San Firmin in Pamplona. All the frames are resized to 480 pixels width.
UMN Crowd[26]	Tasks: **anomalous behaviour detection in crowds**. Collected by researchers of the University of Minnesota (UMN), and made up of three simulated scenes (one indoor and two outdoor) with low-density crowd, starting with normal behaviour and ending with anomalous behaviours (escape panics).
UCSD Anomaly Detection[27]	Tasks: **anomalous behaviour detection in crowds**. Videos of two real scenes collected by UCSD researchers using a stationary camera mounted at an elevation, overlooking pedestrian walkways, with low to high crowd density. Abnormal events are due to either the circulation of non pedestrian entities in the walkways, or to anomalous pedestrian motion patterns (mainly due to bikers, skaters, small carts, and people walking across a walkway or in the grass that surrounds it). All frames are annotated as normal as anomalous, and a subset of clips are provided with manually generated pixel-level binary masks which identify the regions containing anomalies.
Violence-Flows[28] (62)	Tasks: **anomalous behaviour detection in crowds** (violence outbreak). Made up of 246 videos downloaded from YouTube by researchers of the Open University of Israel, showing real-world scenes of crowd violence to test both violent/non-violent classification and violence outbreak detections.
UCF Tracking in High Density Crowds[29] (63)	Tasks: **tracking individuals in a crowd**. Four videos collected by UCF researchers: three marathon sequences, and a busy train station sequence; trajectories of hundreds of individuals are manually annotated.

[23] http://www.ee.cuhk.edu.hk/~xgwang/expo.html

[24] http://crcv.ucf.edu/data/crowd.php

[25] http://crcv.ucf.edu/projects/Abnormal_Crowd/

[26] http://mha.cs.umn.edu/proj_events.shtml#crowd

[27] http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

[28] http://www.openu.ac.il/home/hassner/data/violentflows/index.html

[29] http://crcv.ucf.edu/data/tracking.php

Data set	Description
Train Station[30] (64)	Tasks: **tracking individuals in a crowd**. It consists of a video sequence collected at the New York Grand Central Station, and made available by researchers of the Chinese University of Hong Kong (CUHK). The video is 33m 20s long, with 24 fps and a resolution of 480x720 pixels. Tens of thousand trajectories of individuals are manually annotated.
PETS 2009[31]	Tasks: **crowd counting / density estimation**, **tracking of individual(s) within a crowd**, **detection of crowd flow** and of **crowd events**. Made up of video sequences with actions simulated by about 40 actors, with different views (4 to 8), collected by the organizers of PETS 2009 workshop for the three crowd surveillance tasks mentioned above. The following **crowd events** are present: walking, running, evacuation (rapid dispersion), local dispersion, crowd formation and splitting.
CUHK Crowd[32] (65)	Tasks: **group detection in crowds**. Made up of 474 video clips from 215 crowded scenes collected from Getty Images and Pond5 by CUHK researchers.
Agoraset[33] (66)	A synthetic crowd video dataset that can be used for evaluation of different low-level video crowd analysis methods, like tracking and segmentation. See text for further details.

Table 3 – Publicly available data sets related to crowd monitoring tasks, potentially useful for the development and evaluation of the HCV crowd monitoring tool.

3 REQUIREMENTS

During the activities of work package WP2, a use case has been defined for the HCV tool (reported in deliverable D2.2) starting from the goal and functionality proposed in the DoW. The use case is focused on the person re-identification and people search tools, and is summarized in Sect. 3.3. User (LEA) requirements have then been collected: they are reported in Sect. 3.1. System requirements have been finally defined (see deliverable D2.1), which are reported in Sect. 3.2.

3.1 USER REQUIREMENTS

User requirements have been collected through a questionnaire filled out by LEAs, and through discussions during project meetings. For the HCV tool the goal of the questionnaire was to understand the legal, operational and technical context of the use of VS systems for monitoring mass gathering events by LEAs of different countries, as well as the potential needs by LEAs. This was necessary to guarantee the compliance of the HCV tool with existing regulations and with LEAs' operational procedures, and to check its potential usefulness for supporting LEAs during event execution and in post-event forensics investigations. Questions were subdivided according to the two considered event phases, and are reported below.

- Event execution phase
    - Do you use VS systems to monitor crowded events? If so:
    - What are the laws (privacy, etc.) that rule the use of VS systems by LEAs in your country (possibly depending on the kind of event, like a sport event, a demonstration, etc.)?
    - What kinds of vs systems do you use/can be used in your country? (e.g., fixed cameras, cameras worn by officers, airborne cameras, etc.)

[30] https://www.ee.cuhk.edu.hk/~xgwang/grandcentral.html

[31] http://www.cvg.reading.ac.uk/PETS2009/a.html

[32] http://www.ee.cuhk.edu.hk/~jshao/projects/CUHKcrowd_files/cuhk_crowd_dataset.htm

[33] https://www.sites.univ-rennes2.fr/costel/corpetti/agoraset/Site/AGORASET.html

- How is the setting of the camera network decided (number of cameras, type, positioning, field of view, etc.), and who decides it?
- Do/can your operators watch videos in real time during event execution to monitor the crowd, or are videos just stored for offline analysis (if needed)? In the former case: What is the procedure followed by operators? Is there any software tool already in use to support them in the analysis of videos during event execution, possibly in real time? (e.g., the kind of analysis involved in task T5.4: searching for a specific person observed in some video frames, or described by a witness to some event of interest.)

- Post-event phase. If you use VS systems to monitor crowded events:
    - Are recorded videos used (or can they be used) in the post-event phase, for forensic analysis? (e.g., the kind of analysis involved in task T5.4: searching for a specific person observed in some video frames, or described by a witness to some event of interest.)
    - Is there any software tool already in use to support your operators in this task?

Answers from six LEAs of six different countries where obtained: Fachhochschule fur Offentliche Verwaltung und Rechtspflege in Bayern (**BayFHVR**, Germany), Ertzaintza (**ERT**, Basque country), Ministerul Afacerilor Interne (**MAI**, Romania), Ministero dell'Interno (**INTERNO**, Italy), Lokale Politie Voorkempen (**LPV**, Belgium) and Ministerio da Administracao Interna (**PSP**, Portugal). They can be summarized as follows:

- VS systems are used by all the six LEAs.
- Fixed, existing VS systems are used by all LEAs. Depending on the country regulations (e.g., about the type of event and the crowd size), other kinds of VS systems are used by some LEAs: temporary VS systems (LPV), mobile cameras worn by LEA officers (BayFHVR, MAI, ERT, PSP), by special video teams (LPV), or by forensic operators on field (INTERNO); aerial views from cameras mounted either on helicopters (ERT, INTERNO, LPV and PSP), or on remotely piloted aircraft systems (RPASs) managed by LEAs (LPV, INTERNO and PSP).
- LEA operators from all six countries can watch VS videos in real time (e.g., in a situation room for INTERNO), for monitoring crowds to guarantee public security and to react in case of incidents.
- Videos can also be recorded, subject to constraints on the purpose and on storing time by country-specific regulations. Allowed purposes can be to collect evidence of incidents (LPV), or if there is a concrete threat for the public security or an indication for a criminal behaviour (BayFHVR). Recorded videos can be stored only for a limited time (from three weeks to one month) unless they are useful for investigations, in which case they may be used as evidence against the perpetrators.
- All six LEAs point out that currently no software tools are available for supporting them to monitor and analyse videos. Notably, a working group of BayFHVR is defining the requirements for improving the practical use of VS technologies, and one of their goals is to find a software system capable to support LEA officers in forensic and preventive work, by automatic analysis of video material.
- Finally, all six LEAs agree that it is difficult, although it may be possible, to collect recorded videos for the design and demonstration of the HCV tool, also taking into account the limited storing time.

From the questionnaire outcomes summarized above **two main conclusions** can be drawn: (1) the envisaged HCV tool would be compliant with existing regulations on the use of VS systems by LEAs of the six involved countries; (2) its three main functionalities (person re-identification, people search and crowd monitoring) are not provided by existing tools, and would support monitoring and investigation/forensic tasks that are currently carried out by LEA operators and investigators.

On the other hand, the difficulty of collecting recorded videos from real crowd gathering events for developing and validating the HCV tool may be a limitation, which can be partly overcome using publicly available data sets collected by the CV research community (see Sect. 2.4) as well as further suitable videos available on the Internet; synthetic videos obtained from crowd simulations can be also used for the CM tool of task T5.1.

3.2 SYSTEM REQUIREMENTS

Fourteen requirements have been defined for the HCV tool in deliverable D2.1. They are listed below according to their type (the requirement ID used in deliverable D2.1 is also reported for ease of reference):

- Legal requirements:
  - HCV_002: The tool shall be compliant with EU privacy regulations and with any other regulations of the use of VS systems by the LEAs, including related LEAs internal procedures. The rationale is that the use of VS data affects privacy, and related regulations vary from country to country; moreover, compliance with LEAs internal procedures on the use of VS systems is necessary to make the tool usable.
  - HCV_008: The tool should respect the principle of non-discrimination. In particular, the collection and processing of images of people will have to exclude discriminatory criteria that are not criminological based.
- Look and feel requirement (HCV_006): The tool must be user-friendly: it should provide a simple and intuitive GUI and should be easy and fast to use. The rationale is that the HCV tool must provide a useful support to LEA operators during their crowd monitoring and investigation tasks; especially during event execution it should take as little time as possible to use.
- Usability and humanity requirements (HCV_009): The tool will be first in English, and then translated into other languages after it is reliable.
- Operational requirements:
  - HCV_003: The tool will be designed and evaluated using data sets (images and videos) collected for research purposes and publicly available. The rationale of using image and video data in the design phase is that CV techniques to be used in this tool are data-driven, and thus data as much as possible representative of their operation phase is required during design. The motivation of using publicly available data sets is the difficulty (discussed in Sect. 3.1) of using real VS data provided by LEAs.
  - HCV_005: The tool shall exploit feedback from LEA operators to improve its effectiveness over time. The rationale is that existing CV techniques do not achieve human-like performance in challenging scene interpretation and recognition tasks like the ones considered in the HCV tool, and human feedback can help improving their performance.
- Functional and data requirements:
  - HCV_010: The tool shall provide a crowd monitoring functionality, including anomaly detection in crowd behaviour, crowd density estimation and group detection.
  - HCV_007: The crowd monitoring tool will process videos acquired by standard, fixed or PTZ, VS colour cameras. Tilt angle with horizontal plane: about 45 degrees or more; height: about 5 m or more. The rationale is that near-top view reduces/minimizes occlusions in crowded scenes, which is one of the main difficulties for CV techniques.
  - HCV_011: The tool should provide a person re-identification functionality: given a query image of an individual of interest, it will return a list of images of individuals exhibiting a similar clothing appearance, sorted for decreasing similarity to the query.
  - HCV_012: The tool should provide a people search functionality: given a description of clothing appearance, it will return a list of images of individuals matching that description, sorted for decreasing degree of matching.

- HCV_013: The person re-identification and people search tools will process videos from standard, fixed/PTZ/mobile (managed by stewards/agents), VS colour cameras. Tilt angle with horizontal plane: less than 45 degrees; height: about 3 m or less. The rationale is that for these two functionalities the whole body of individuals must be visible.
- HCV_014: The tool should be designed using videos acquired by VS systems during relevant, real or simulated mass gathering events.
- HCV_015: The tool shall provide a web-based graphical interface for each functionality. The reason is that a web-based graphical interface does not require any software to be installed and configured by LEAs for validation, and can be a feasible interface for (future) real tools.
- HCV_016: The crowd monitoring tool may process videos acquired by RPASs, if allowed by EU regulation on this matter currently in progress. The rationale is that aerial views are the most suitable ones for analysing the behaviour of a large crowd, to minimize the impact of occlusions (e.g., for people counting/density estimation) and of perspective distortion.

3.3 USE CASES

A use case was developed in work package WP2 (reported in deliverable D2.2 as UC-005) to illustrate the functionality of the person re-identification and people search tools. It focuses on two usage scenarios of VS data: supporting LEAs reaction to some incident during event execution, and supporting forensic investigation in the post-event phase. Two main facts are pointed out by this use case: the aim of CV tools like the ones envisaged in the HCV tool is not to work autonomously in place of human operators, but only to support them; moreover, their effectiveness can be improved using operator's feedback.

According to the common structure of deliverable D2.2, this use case is structured into inputs requested, pre conditions, trigger event, basic path, exception paths, actors involved, and a narrative description.

Two **inputs** are requested to implement the use case. The first is information from LEAs about the use of VS systems during mass gathering events, collected through questionnaires (see Sect. 3.1), to guarantee the compliance with existing regulations on the use of VS systems and the potential usefulness of the proposed tools for LEA operations. The second input is a set of videos acquired by VS systems, satisfying system requirements (see Sect. 3.2); such videos are necessary both to develop the tools using data-driven CV techniques, and for their validation by LEAs, as well as for the final demonstrations. Ideally, these should be real videos available to LEAs and recorded during past relevant mass gathering events. However, taking into account the difficulty of retrieving and using such data in LETSCROWD (as pointed out by the LEA questionnaires), alternative solutions can be used, like videos acquired by simulations of mass gathering events, and publicly available video data sets used in the CV research community (see Sect. 2.4).

Three **pre-conditions** are planned: (i) The mass gathering location is monitored by a VS system, including a recording system capable of recording the acquired videos; videos acquired by LEA operators through wearable cameras, or by participants to the event through their own smartphones could also be used for post-event investigations (depending on the quality of the video and the legislation in EU Member States). (ii) LEA operators are allowed to watch the videos acquired by the CCTV system during event execution to monitor the crowd and to support their colleagues in the field. (iii) LEA operators are allowed to analyse the acquired videos to carry out investigations, with suitable permissions by the competent authority (if needed). These pre-conditions are all satisfied, according to LEA questionnaires summarized in Sect. 3.1.

The use case can be **triggered** by three kinds of events: (i) A LEA operator who is watching the CCTV videos on several screens in real time during event execution, or is analysing the recorded videos during post-event investigation of an incident, sees a suspect individual, and would like to reconstruct his/her previous movements and actions in the monitored areas. (ii) An incident occurred during event execution. During the reaction phase, or during investigation in the post-event phase, some witnesses describe to LEAs the clothing appearance of suspect individuals. LEA operators would like to analyse the acquired videos to find images of such individuals. (iii) LEAs have been informed of the presence of one or more suspect individuals in the mass gathering location, and a description of their clothing appearance is available. As above, LEA operators would like to analyse the acquired videos to find images of such individuals.

The **actors involved** are mainly Police officers, and possibly Public Authorities.

What follows is the narrative **description** of the use case.

(i) During event execution LEA operators are watching in real time, on several screens, the videos acquired by a CCTV system. Two things can happen:

- They see a suspect individual in one of the videos (possibly after some incident), and would like to reconstruct the previous movements and actions of that individual in the monitored areas. An operator selects such an individual from the observed video sequence (e.g., using a playback facility and a point-and-click graphical interface), and runs the person re-identification tool. This tool automatically searches for individuals wearing similar dresses on all the videos acquired so far, and returns the operator the retrieved frames/tracks (including the timestamp, the camera location and other relevant information), sorted for decreasing degree of similarity to the suspect individual. This can allow the operator to find other images of the suspect individual (if any) and to analyse the corresponding video tracks, in a much shorter time (possibly in real time) than watching all the available unsorted videos, which can be infeasible for long videos and/or many cameras.
- An incident occurred, and the responsible(s) is (are) not visible on the acquired videos, but in the reaction phase some witnesses describe to LEA operators the appearance (including clothing appearance) of one or more suspect individuals. Alternatively, LEAs have been informed of the presence of one or more suspect individuals in the mass gathering location, and a description of their clothing appearance is available. In both cases, LEA operators would like to immediately (during event execution) retrieve video frames/tracks of individual(s) matching the available description, if any. To this aim they run the people search tool. They input the description of clothing appearance using a predefined set of attributes (e.g., the specific colour and/or texture of the upper-body and lower-body garment, the presence of short or long sleeves), then the tool automatically searches for individuals exhibiting a similar clothing appearance, and returns the retrieved frames/tracks (including contextual information like the camera location and the timestamp), sorted for decreasing degree of matching to the input description. Similarly, to the previous case, this allows LEA operators to find the images of the suspect individuals (if any) in a much shorter time than watching all the available unsorted videos.

(ii) During the post-event phase LEAs are carrying out an investigation on an incident occurred during event execution. As part of the investigation, some LEA operators are analysing the videos acquired by a CCTV system to find images of the responsible(s) and reconstruct their movements and actions in the monitored areas. Two similar cases as the ones described above can happen:

- LEA operators see a suspect individual in one of the videos, and would like to know his/her movements and actions in the monitored areas: to this aim they run the person re-identification tool (see above).
- The responsible(s) is (are) not visible on the recorded videos, but a description of their appearance is available (for instance, provided by a witness): in this case LEA operators run the people search tool (see above).

In both cases, regardless of whether or not the suspect individual(s) appears in the video frames returned by the chosen tool, the person re-identification (people search) tool can ask the LEA operator to select a few similar individuals (exhibiting the described appearance) near to bottom of the list, and/or a few dissimilar ones (exhibiting a different appearance) near the top of the list, if any. This feedback allows the tools to update and improve the algorithm they use to evaluate the similarity between images of individuals (person re-identification), and to match a given description of clothing appearance to images of individuals (people search). The LEA operator can then run another retrieval step to get a refined list of results, which can contain novel video frames/tracks containing the suspect individual.

In the **basic path** one or more video tracks showing the suspect individual(s) are found by LEA operators in the video frames returned by the chosen tool, and in the case of the people search tool they are recognized by the witnesses. This can trigger further actions, e.g., informing colleagues in the field of the presence of that individual(s) during event execution, or enabling the prosecution of a post-event investigation.

The **exception path** is that the suspect individual(s) does not appear in the video frames returned by the tools: in this case LEA operations proceed as they would currently do.

4 DESIGN

This section describes the architecture of the three components of the HCV tool, their link to tools developed in other tasks and work packages, a description of their algorithms, and their user interface.

4.1 ARCHITECTURE

The architecture of the three components of the HCV tool is described in distinct sub-sections.

4.1.1 Person re-identification tool

This tool will process two kinds of inputs: one or more videos either coming in real time from video cameras (during event execution) or pre-recorded (in the post-event phase), and inputs from the operator (through a GUI). The tool will consist of the following components (see Figure 5), which provide the person re-identification functionality of Figure 1 in a **end-to-end system**, and with a **HITL approach**:

- **pedestrian detector**: it processes the input videos to extract images (bounding boxes) of pedestrians;
- **feature extractor**: it takes as input a pedestrian image (bounding box) coming from the pedestrian detector and computes a descriptor of that image;
- **template gallery**: it contains all the images produced by the pedestrian detector, their descriptors and contextual information (including the source video and the timestamp); the template gallery is incrementally updated as soon as new pedestrian images arrive from the pedestrian detector;
- **matching module**: it takes as input the descriptor of a query image (chosen by the operator), matches it with all the descriptors of template gallery images, and returns a ranked list of template images sorted by decreasing similarity to the query;
- **HITL module**: it receives **feedback** information from the operator on a ranked list of templates, for a given query image, and updates the matching module accordingly;
- **GUI**: it will show the operator the input videos, and will allow him/her to stop any of the videos, to select form a frame a bounding box containing an individual of interest as the query image, to run the matching module, to scroll the ranked list of template images, to select any template image to see contextual information, to play the corresponding video back, and to provide feedback about any template image.

The architecture of this tool will be **client-server**: the GUI will be at the client side, all the other modules will operate at the server side. A **web-based** GUI will be implemented to maximize flexibility for the sake of validation and demonstration.

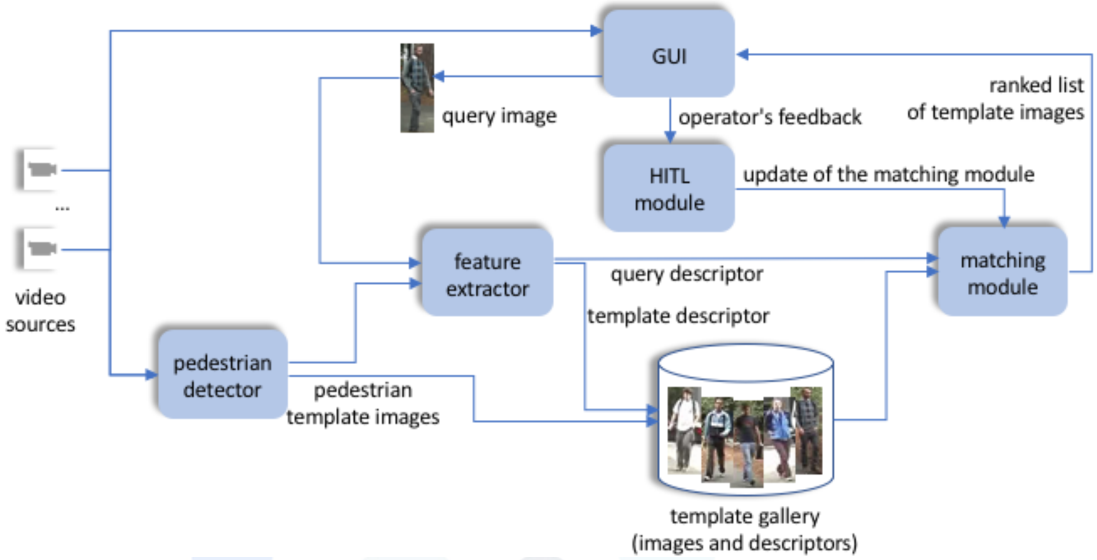


Figure 5 – Functional architecture of the person re-identification tool.

A related issue is the setting of the cameras (position, resolution, etc.). This issue is not discussed in the literature. As emerged from the user requirements (Sect. 3.1) not all cameras are controlled by LEAs, and some details of the camera setting are regulated by country-specific rules. Accordingly, system requirements (Sect. 3.2) include only general indications: tilt angle with horizontal plane less than 45 degrees, height of about 3 m or less. With regard to resolution, several benchmark data set contain images of 128x64 pixels or similar size; accordingly, this can be translated into a constraint on the camera distance, in terms of (approximately) the minimum size of a pedestrian bounding box useful for person re-identification. For mobile cameras the presence of blur should be taken into account, as it can limit the image quality. At this stage of the development of the HCV tool it is not possible to give more precise indications and guidelines for LEAs on camera setting; such information can be obtained from practical demonstrations of work package WP6 (currently under definition), and will be reported in deliverable D5.8.

4.1.2 People search tool

The people search tool will process two kinds of inputs: one or more videos either coming in real time from video cameras (during event execution) or pre-recorded (in the post-event phase), and inputs from the operator (through a GUI). The tool will consist of the following components (see Figure 6), which provide the people search functionality of Figure 2 in a **end-to-end system**; note that the pedestrian detector is shared with the person re-identification tool, whereas the feature extractor may be different.

- **Pedestrian detector**: it processes the input videos to extract images (bounding boxes) of pedestrians;
- **feature extractor**: it takes as input a pedestrian image (bounding box) coming from the pedestrian detector and computes a descriptor to be used by the attribute detectors (see below);
- a set of **attribute detectors**: each one is specific to one of a predefined set of attributes; each detector takes a pedestrian image descriptor as the input, and outputs either a Boolean value denoting whether the corresponding attribute is present or not in the input image, or a continuous score in a predefined range representing the probability that the the corresponding attribute is present; all the detectors will be pre-trained, and can be automatically updated (re-trained) based on the operator's feedback;
- **template gallery**: it contains all the images produced by the pedestrian detector, together with their descriptors, contextual information (including the source video and the timestamp), and the attribute profile (i.e., the output of attribute detectors); the template gallery is incrementally updated as soon as new pedestrian images arrive from the pedestrian detector;

- **retrieval module**: it takes as input a query consisting of a description of clothing appearance given by the operator in terms of an attribute profile (a combination of the predefined attributes), matches it with the attribute profile of each template gallery image, and returns a ranked list of template images sorted by decreasing similarity to the query description;
- **HITL module**: it receives **feedback** information from the operator on the ranked list of templates obtained for a given query description, and updates the attribute detectors accordingly;
- **GUI**: it allows the operator to input a query in terms of an attribute profile, to run the retrieval module, to scroll the ranked list of template images, to select any template image to see contextual information, to play the corresponding video back, and to provide feedback about any template image.

Similarly to the person re-identification tool, the architecture of this tool will be **client-server** (the GUI will be at the client side, the other modules at the server side), and a **web-based GUI** will be implemented.

With regard to the camera setting, the same considerations made for the person re-identification tool (see Sect. 4.1.1) apply also to the people search tool.

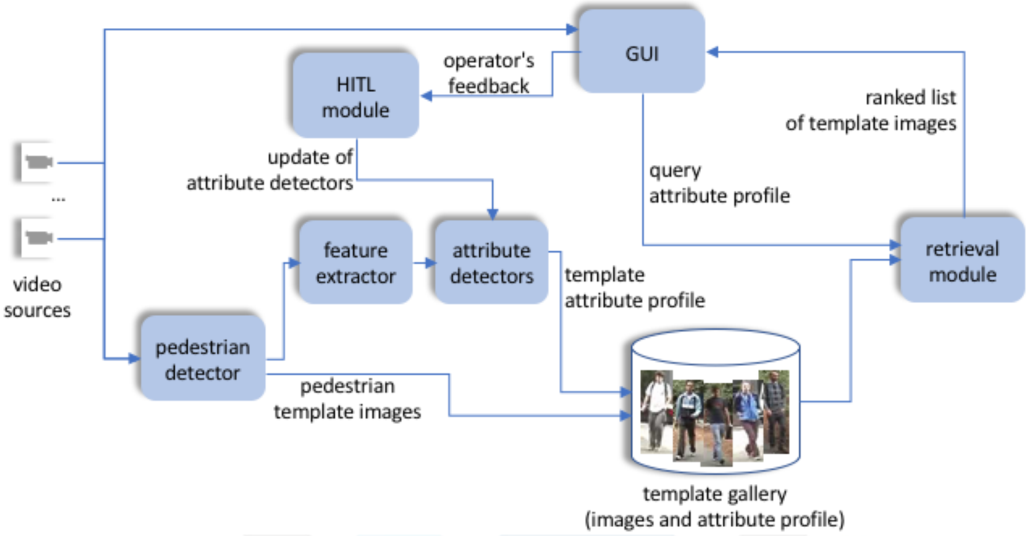

GUI
operator's feedback
HITL module
ranked list of template images
update of attribute detectors
query attribute profile
...
feature extractor
attribute detectors
retrieval module
video sources
template attribute profile
pedestrian detector
pedestrian template images
template gallery (images and attribute profile)

Figure 6 – Functional architecture of the people search tool.

4.1.3 Crowd monitoring tool

This will be a collection of modules that will process in real time input videos coming from one or more video cameras, and will show the operator the output of the analysis for each of the implemented monitoring task; some of the outputs (e.g., estimated crowd density, or anomalous variations of crowd density) will be sent to the DRA tool and to the CM tool, as further discussed in Sect. 4.2. In the first version of the crowd monitoring tool the following functionality will be implemented through distinct modules, related to the analysis of a crowd at a **macroscopic** level; the architectural details of these modules are currently under definition.

- **Crowd density estimation**: this module will provide an estimate of the number of people (or of the density of people) that occupies one or more given regions of the camera view. Video simulations of a crowd obtained from the CM tool will be used as additional training data, beside real videos, to train the machine learning-based algorithms used in this module. The output of this module will be shown in a GUI, and will be sent to the CM and DRA tools.
- **Detection of crowd patterns of movement**: this module will detect the main directions of crowd flow and their velocities, possibly in terms of predefined patterns depending on the chosen algorithm. Similarly to the previous module, this information will be graphically shown in a GUI, and will be sent to the crowd modelling tool.

- **Anomaly detection in crowd behaviour**: this module will take as input the outputs of the two previous modules and possibly a model of normal crowd behaviour obtained from the CM tool, and will detect the following kinds of anomalies:
    - anomalies in the global or local crowd density: either static or dynamic (e.g., overcrowding with respect to a predefined density, or a sudden increase or decrease in density);
    - anomalies in the crowd patterns of movement with respect to a model of normal behaviour, either predefined (e.g., in terms of the normal directions and velocities of movement) or learnt from data; the predefined model of normal behaviour can be obtained from the CM tool; additionally, synthetic videos of normal (and possibly anomalous) crowd flow obtained from the CM tool could be used to train anomaly detection algorithms.

  The output of this module will be shown to the operator in the form of alerts, both as a short text message and graphically.

Additional monitoring functionality related to the analysis of a crowd at the **microscopic** level could be implemented in the second, final version of the HCV tool, e.g.:

- a **pedestrian tracking** module to track individuals in each input video;
- a **group detection** module which processes the input videos and the output of the pedestrian tracking module, to detect groups of pedestrians; possibly, their clothing appearance will also be extracted for further analysis (see below) through a descriptor similar (or identical) to the one used in the person re-identification tool; groups of people will be graphically shown superimposed to the input video;
- an **event detection and suspicious behaviour detection** module: it will process the input videos and the output of the pedestrian tracking and group detection modules, and will detect specific events (to be defined) both at at the individual and at the group level, e.g.: running people, sudden group formation or breaking, groups made up of people with similar clothing, etc.; the output of this module will be shown to the operator in the form of alerts, both as a short text message and graphically, superimposed to the input video, and will be sent to the DRA tool.

With regard to camera setting, according to system requirements (Sect. 3.2) fixed and PTZ cameras are planned, with an angle with the horizontal plane of about 45 degrees or more, and height of about 5 m or more. For crowd analysis at the macroscopic level relatively far views are preferred. For analysis at the microscopic level closer view are needed. For reasons similar to the ones discussed in Sect. 4.1.1, more specific indications are not included in system requirements. We point out here that the use of PTZ cameras should deal with the need of a perspective map by algorithms like the ones used for crowd density estimation: in this case information of the current view by a PTZ camera is needed, e.g., to automatically update the perspective map, or to choose among a predefined set of such maps (in the latter case the possible camera views should be predefined and known in advance). The use of RPASs has been considered in the system requirements only as a potential source of videos, since its regulation at the EU level is currently in progress, although it is already allowed in some countries (as indicated by LEAs in the user requirements, see Sect. 3.1). More precise indications and guidelines to LEAs on camera setting will emerge from practical demonstrations of work package WP6, and will be reported in deliverable D5.8.

4.2 LINKS/INTERFACES BETWEEN TASKS/WORK PACKAGES

Links are currently being defined between the crowd monitoring functionality of the HCV tool and two other tools: the CM tool of task T5.1, and the DRA tool of work package WP3. These links have been mentioned in Sect. 4.1, and are summarized below.

The **link with the CM tool** will be two-way:

- the CM tool will use the output of the **crowd density estimation** module of the HCV tool, possibly **in real time**; it could also exploit **offline** the output of the of the detector of **patterns of crowd movement** (main directions and velocities), e.g., to calibrate and validate the simulation set up;
- video simulations of a crowd obtained from the CM tool will be used **offline** to train the machine learning-based algorithms used by the HCV **crowd density estimation** module;
- video simulations of normal (and possibly anomalous) crowd flow obtained from the CM tool will also be used **offline** to train the machine learning-based algorithms of the HCV **anomaly detection** module related to macroscopic crowd behaviour (main directions and velocities).

The **link with the DRA tool** will be one-way, from the HCV to the DRA tool:

- the DRA tool will use the output of the HCV **crowd density estimation** module **in real time** to update the risk level, according to the approach described in deliverable D3.4;
- outputs of the HCV **anomaly detection** module will be sent **in real time** to the DRA tool as **weak signals**;[34] these outputs can include detected anomalies at the crowd (macroscopic) level, i.e., related to crowd density and patterns of movement, as well as events and suspicious behaviours at the individual and group (microscopic) level that the HCV tool will be capable to detect (this functionality is currently under design), and that will be considered useful to the DRA tool.

According to the DRA tool architecture in deliverable D3.4, the outputs of the HCV tool will be sent to the Data Logger module using the Common Alerting Protocol data format; the outputs will include:

- a time signature (absolute time);
- the geolocation of the detected event;
- a signature, made up of the features related to what it has been detected by the sensor, expressed using a semantic to be defined (e.g., the estimated number of people or crowd density - individuals per square meter - in a given area, or keywords describing a specific event);
- a reliability measure in the range $[0,1]$, representing, e.g., the uncertainty in the estimated crowd density, or the uncertainty in a specific event detected;
- a snapshot or a link to the video source of the detection, to be used by the operator to confirm, discard (false alarm) or amend the detection.

4.3 ALGORITHM DESCRIPTION

In this section the algorithms used to implement the HCV tool are described in distinct sub-sections.

4.3.1 Person re-identification tool

The first version of this tool is being developed using the algorithm proposed in (14) (see Sect. 2.1.1), which exhibits three main advantages for the HCV tool: it uses the HITL approach; the implementation of the HITL approach requires a simple kind of feedback from the operator; any descriptor of pedestrian images, in the form of a fixed size feature vector, can be used, which allows to tune the trade-off between descriptor effectiveness and efficiency (processing cost) depending on the application at hand.

As most of the re-identification methods proposed in the literature, also the one of (14) defines only two of the modules described in Sect. 4.1.1: the image descriptor built by the feature extractor, and the similarity measure used by the matching module. These two components are summarized in the following, together with the algorithm used to implement the HITL approach.

[34] A *weak signal* is defined in deliverable D3.4 as "the minimum quantum of information managed by the DRA, i.e. the detection of each sensor involved."

- The pedestrian image **descriptor** is the one proposed in (67). The pedestrian image is first resized to 128 by 64 pixels, and then subdivided into eight horizontal strips of identical size. From each strip three weighted colour histograms are extracted from the RGB, HSV (only the hue and saturation channels) and Lab (CIELAB) spaces, where the weights are defined by a Gaussian kernel centred at the image centre to reduce the contribution of pixels more likely to belong to the background. Then, two image regions are considered, obtained by discarding the top and bottom strips (which are likely to contain less discriminative information, respectively the head and the feet, including a relatively large background region), and by combining the three top and three bottom remaining strips; from each of these regions the Local Binary Pattern (LBP) texture descriptor and the Histogram of Oriented Gradients (HOG) edge descriptor are extracted (both of them are histograms). These colour, texture and edge histograms are then concatenated into the final descriptor.
- The **similarity measure** between two descriptors (column vectors) $\mathbf{x}_1$ and $\mathbf{x}_2$ is computed as the negative Mahalanobis distance $-(\mathbf{x}_1 - \mathbf{x}_2)^{\mathrm{t}} M(\mathbf{x}_1 - \mathbf{x}_2)$, where M is a square, positive semi-definite matrix. The matrix M is defined through the HITL approach, and is initially defined as the identity matrix.
- The **HITL** algorithm consists in updating the matrix M each time the operator provides a feedback on one of the ranked template images returned by the matching module for a given query image. The feedback can take three values: "true-match" (if the identity of the template is the same as the query), "strong-negative" (if the identities are different, and the clothing appearance are dissimilar), and "weak-negative" (if the identities are different, but the clothing appearance are similar). In the case of "true-match" and "weak-negative", M is updated to push the template image chosen by the operator toward the top of the ranked list (the rationale is that in both cases the template looks similar to the query), whereas in the case of "strong-negative" M is updated to push the template image toward the bottom of the ranked list. After the update, the template gallery is re-ranked and the new ranked list is shown to the operator. It should be noted that M is incrementally updated: this allows the similarity measure to adapt to the characteristics of the template gallery at hand.

4.3.2 People search tool

As described in Sect. 2.1.2, existing methods for clothing appearance attribute detection in pedestrian images use two different approaches: one is to implement a binary classifier for each attribute, using low-level features extracted from the input pedestrian image; the other consists in using a more complex multi-label deep neural network which takes as input a pedestrian raw image, and acts both as a feature extractor and as a classifier/detector. In the first version of the people search tool the former, simplest approach will be used.

- The chosen **feature** descriptor is the Ensemble of Localised Features (ELF) used in (18). It consists of a concatenation of colour and texture feature vectors extracted from six horizontal strips of identical size. In particular, eight colour channels (RGB, HSV and YCbCr) and 21 texture filters (Gabor and Schmid) derived from the luminance channel are extracted from each strip.
- The chosen set of **attributes** is the one of the PETA data set (see Sect. 2.4), currently the largest available data set of pedestrian images taken both in outdoor and in indoor VS settings (made up of a collection of several benchmark person re-identification data sets). Each image is manually annotated according to 65 different attributes.
- For each attribute, a specific **detector** is implemented as a binary Support Vector Machine (SVM) classifier with histogram intersection kernel. Note that all the attributes of the PETA data set are binary except for four 11-valued attributes, which can be converted into 11 binary attributes. The attribute detectors will be pre-trained on the whole PETA data set.
- The **HITL** approach will be implemented by allowing the operator to provide a feedback on any subset of attributes for any subset of the retrieved pedestrian images, for a given query description. The feedback consists in indicating false positive or false negative errors made by the detectors, i.e., attributes labelled as present in a given pedestrian image whereas they are absent, and attributes labelled as absent

whereas they are present, respectively. This feedback will be used to re-train the corresponding SVM classifiers (detectors), after adding the template images chosen by the operator to their training sets, respectively as negative and positive examples. To reduce the processing cost of the re-training phase, two possible solutions will be considered: running the learning algorithm on a batch of images, i.e., only after a minimum predefined number of images has been collected for the corresponding attribute; and using incremental learning techniques. Anyway, classifier re-training can be carried out offline, and each detector can be updated only when the new training phase is completed.

4.3.3 Crowd monitoring tool

Currently only the first version of the algorithm of the crowd density estimation module has been defined. Based on the guidelines provided in the survey (29) for density estimation in crowded scenes (see Sect. 2.1.3), an algorithm based on the counting by regression approach will be used. It provides an estimate of the number of people in a single video frame (scene), using static features. Its main processing steps are:

- **Perspective normalization**: it will be implemented using the simplest approach described in (29) for fixed camera views. A quadrilateral is manually determined on a video frame, corresponding to a rectangle on the ground plane of the scene (in a region which can be occupied by pedestrians), such that two opposite sides are parallel to the image horizontal scan lines. The length of the sides, that correspond to the closest and farthest points of the quadrilateral to the camera, and the height of a reference pedestrian that passes the two sides are recorded, to be used for normalizing the image features extracted at any image coordinate. The normalization procedure assumes that the size of foreground segments changes at a quadratic rate with respect to the perspective, and that their edge pixels change linearly.
- Holistic **features** will be extracted to represent a whole scene, with no detection of single pedestrians. Different kinds of features considered in [Loy-2013] will be used, as it was shown in (29) that each kind of feature (among the ones proposed in the literature) is effective on different kinds of scenes:
    - foreground segment features obtained through background subtraction (e.g., number of pixels in the segment and in its perimeter, perimeter-area ratio, perimeter edge orientation histogram, number of connected components with area larger than a predefined threshold); either static or dynamic background subtraction methods can be used, depending on the amount of illumination changes that is expected in the scene;
    - edge features (e.g., number of edge pixels, histogram of edge orientations, Minkowski fractal dimension of the edges); edges are extracted using the well-known Canny detector;
    - texture and gradient features, e.g., grey-level co-occurrence matrix and LBP.
- A linear **regression model** will be used to map the perspective normalized feature vector extracted from a video frame to the estimate of the number of people in the scene. Whereas more complex, non-linear regression models were found to provide often better results on benchmark data sets, when the testing images exhibited similar characteristics as the training images, the simplest linear models were found to be more robust to the mismatch between training and testing images, which is likely to occur in real-world applications (29). In particular, two versions of the linear regression model will be used: the standard linear regression, whose output is defined as a linear combination of the input variables (features), and the partial least squares regression, which addresses the overfitting problem due to multicollinearity of the input variables, likely to occur in high-dimensional feature vectors.

4.4 USER INTERFACE

The user interfaces of the HCV tool are described in distinct sub-sections. Currently, only the user interface of the person re-identification and people search tools has been defined.

4.4.1 Person re-identification tool

The GUI will be made up of three different windows: the main window, the results window and the focus window. An example of each of them is given in Sect. 5.1.3.

- The **main window** (see Figure 8, top) will show the operator the input videos; depending on the number of videos, in a real application scenario it could be subdivided in several screens. When the operator sees an individual of interest in one video, he/she can stop the video, select from a frame a bounding box containing an individual of interest (which becomes the query image) through a point-and-click functionality, and run the search function of the matching module by clicking on a button.
- The outcome of the matching module is shown in a new, **results window** (see Figure 8, middle). This window shows the query image, and a subset of the ranked list of template images, starting with the top-ranked ones. The operator can scroll the ranked list, and can select any template image for further analysis, which triggers the opening of the focus window (see below). Additionally, three radio buttons are shown for each template image, corresponding to the feedback information that the operator can provide: "true-match", "strong-negative" and "weak-negative" (see Sect. 4.3.1). As soon as the operator provides a feedback on one template image, the HITL module updates the similarity measure and then the ranked list of templates, and the updated list is shown to the operator in the same result window.
- When a template image is selected by the operator, a new **focus window** opens (see Figure 8, bottom), which shows the query image, and the selected template image together with its contextual information: the camera from which that image was taken, geolocation, the whole video frame containing that image, and the time stamp. By clicking on the video frame, the corresponding video is played back to allow the operator to better analyse the appearance and the behaviour of the individual in the template image. Exploiting geolocation, the position of a given template image can also be shown on a map; if more template images of the individual of interest are found, this will allow the operator to show the positions of that individual in different times to analyse his/her movements.

According to system requirements (Sect. 3.2), the GUI will be web-based to maximize flexibility for the sake of validation and demonstration.

4.4.2 People search tool

The GUI will be made up of three different windows, which will be similar to the ones of the person re-identification tool (see Sect. 4.4.1 and Figure 8): main window, results window and focus window.

- The **main window** allows the operator to input the description of the clothing appearance of the target individual, in terms of any combination of the predefined attributes (attribute profile). This window will contain a list of such attributes grouped by category (e.g., attributes related to upper body colours). Each attribute is associated with three radio buttons corresponding to the values 'Yes' (if the attribute is present in the clothing appearance description of the target individual), 'No' (if the attribute is indicated as not present in the target's description) and 'Unspecified' (if the presence or absence of the attribute is not mentioned in the target's description). For all attributes the 'Unspecified' radio button is selected by default. If a subset of attributes is mutually exclusive (e.g., different plain colours of the upper body), selecting 'Yes' for any of them will force the automatic selection of 'No' for the others; changing a 'Yes' selection to 'No' will force the automatic selection of 'Unspecified' for all of them. After the operator completes the desired attribute profile, he/she can start the retrieval process by clicking a button.
- The outcome of the retrieval module is shown in a new **results window** (similar to the one of Figure 8, middle). This window will show a textual description of the selected attribute profile, and a subset of the ranked list of template images, starting with the top-ranked ones (i.e., the ones containing pedestrians whose clothing appearance is most similar to the selected attribute profile). The operator can scroll the ranked list, and can select any template image for further analysis, which triggers the opening of the focus window (see below). Additionally, for each template image the operator is allowed to provide a feedback for each attribute, in terms of the same values 'Yes' (the attribute is present in the template image), 'No' (the attribute is absent), or 'Unspecified' (the default choice for all attributes). To this aim, for each template image a list of the attributes indicated as present ('Yes') or not present ('No') in the selected attribute profile is shown, together with a pop-up containing all the attributes indicated as 'Unspecified'. If the operator selects 'Yes' or 'No' for a given attribute of a given template image, that

image is added to the training set of the corresponding classifier (attribute detector) respectively as a positive and negative example; the updated training sets will be used by the HITL module to re-train (offline) the corresponding classifiers, as described in Sect. 4.3.2.

- When a template image is selected by the operator, a new **focus window** opens (similar to the one of Figure 8, bottom), which shows the textual description of the selected attribute profile, and the selected template image together with its contextual information: the camera from which that image was taken (e.g., its location), the whole video frame containing that image, and the time stamp. By clicking on the video frame, the corresponding video is played back to allow the operator to better analyse the appearance and the behaviour of the individual in the template image. As in the person re-identification GUI, the position of selected templates can be shown also on a map.

According to system requirements (see Sect. 3.2), also the GUI of this tool will be web-based.

4.4.3 Crowd monitoring tool

The GUI of the crowd monitoring tool will be designed after the the detailed functionality of the corresponding module is defined. It is currently planned that the GUI will show in real time the videos coming from the different video cameras (possibly pre-recorded for demonstration purpose), and that the output of each module is superimposed to the original video, possibly in a distinct window to allow the operator to watch both the original video and the same video with superimposed output from each module. All the detected events will be associated to their geolocation and time stamp. Examples of the three modules that will be implemented in the first version of the crowd monitoring tool are given below.

- **Crowd density estimation**: the output can be shown as a heat map with a colour scale that shows the crowd density in different image regions (e.g., from blue to light red to represent lowest to highest densities). An example taken from (29) is shown in Figure 7.
- **Detection of crowd patterns of movement** (main directions and velocities): the output can be shown as arrows indicating the directions of flow, with length or thickness proportional to the corresponding velocity; the values of the velocity can be shown above each arrow.
- **Anomaly detection in crowd behaviour** (density and patterns of movement): the output can be shown as red graphics on the detected image region, e.g., a dark red in the image region where overcrowding has been detected, or a red arrow for anomalous directions or velocities; an alert can also be shown next to the image, as a short text message.

Additionally, thanks to geolocation, it will be possible to display all the detected information on a map of the event venue.

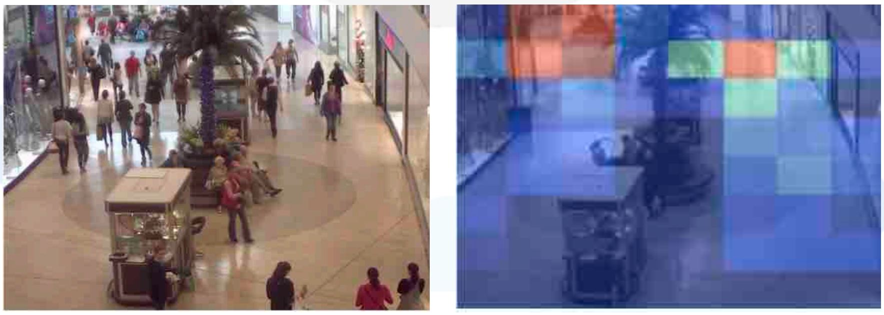


Figure 7 – Left: a video frame with a crowd scene; right: a heat map representing crowd density for the same camera view (for a different scene). Both images are taken from (29).

5 IMPLEMENTATION

This section reports details about the implementation of the person re-identification and people search tools, in distinct sub-sections. The crowd monitoring tool is currently under design.

5.1 PERSON RE-IDENTIFICATION TOOL

5.1.1 Percentage complete

The first version of this tool is 70% complete. With reference to the architecture described in Sect. 4.1.1:

- A publicly available **pedestrian detection** tool will be used (available tools are under evaluation).
- The **feature extractor** has been implemented.
- The **matching module** has been implemented.
- The **HITL module** has not been implemented yet.
- The **GUI** is under development (75% complete).

5.1.2 How implemented

Except for the pedestrian detection tool, all the software modules are being implemented using the following programming languages, libraries and frameworks:

- The Python[35] (v. 2.7) open-source language (it is developed under an OSI-approved open source license, which makes it freely usable and distributable).
- The following Python libraries:
    - matplotlib[36] (v. 2.1.2): a 2D plotting library;
    - numpy[37] (v. 1.14.1): a package for scientific computing;
    - skimage[38] (v. 0.13.1): an image processing toolbox.
- The OpenCV[39] (Open Source Computer Vision) library (v. 3.3), released under a BSD license (free for both academic and commercial use). This library has been designed for computational efficiency, with a strong focus on real-time applications. To this purpose it is written in optimized C/C++, and has C++, Python and Java interfaces. It supports Windows, Linux and Mac OS operating systems.
- The server side of the tool is being implemented using the high-level Python Web framework Djiango.[40] In particular, the communication with the client side is being implemented through JavaScript, using the Ajax pattern,[41] the JQuery library,[42] and the Json format to encode the server-to-client communication.
- The client side (GUI) is being implemented in HTML, using the Bootstrap open-source library.[43]

5.1.3 Current progress

A draft of the GUI is shown in Figure 8.

[35] https://www.python.org/

[36] https://matplotlib.org/2.1.2/

[37] http://www.numpy.org/

[38] http://scikit-image.org/

[39] https://opencv.org/

[40] https://www.djangoproject.com/

[41] https://www.w3schools.com/xml/ajax_intro.asp

[42] https://jquery.com/

[43] https://getbootstrap.com/

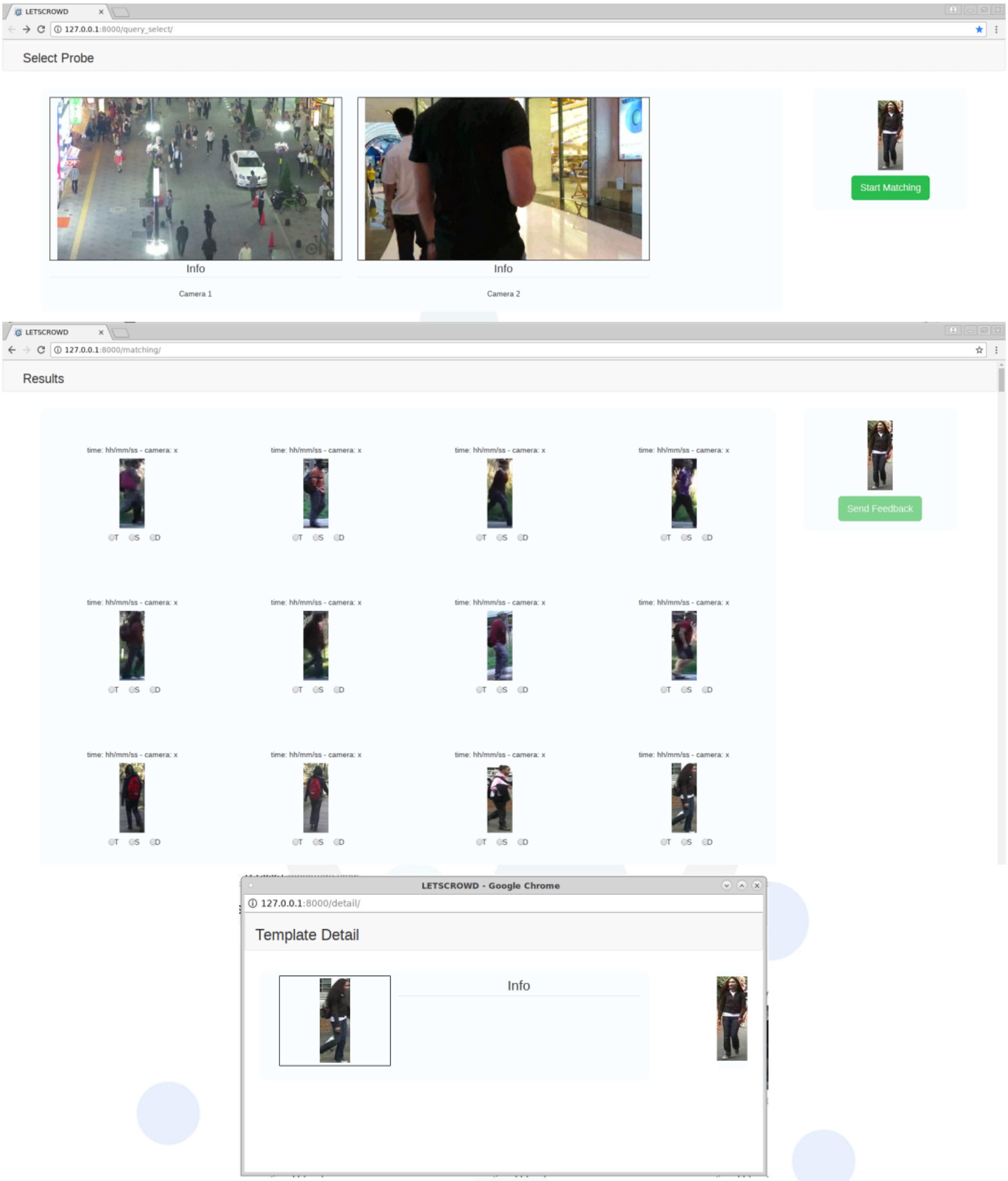


Figure 8 – Draft GUI of the person re-identification tool (see Sect. 4.4.1). From top: main window, results window, focus window.

5.1.4 Validation

A first validation of the person re-identification performance has been carried out on one of the most widely used publicly available benchmark data sets, VIPeR.[44] It consists of images (manually extracted

[44] https://vision.soe.ucsc.edu/node/178/

bounding boxes) of 632 pedestrians. Two images are available for each identity, taken by two different cameras with a significant amount of viewpoint and illumination variation, for a total of 1,264 images. All images are 128x48 pixels. This is one of the oldest data sets (dating back to 2007), and is relatively small with respect to more recent ones, but is still believed to be one of the most challenging data sets for person re-identification.

The usual experimental set-up is to use one randomly chosen image for each identity as a query, and the other as a template. This results in a template gallery of 632 images of distinct identities, and for each query identity there is exactly one template image in the gallery. Under this setting, re-identification performance is commonly evaluated as the cumulative matching characteristic (CMC) curve (1): it is defined as the estimated probability (over the considered query images) that the query identity is among the top-*r* positions of the ranked list, where *r* ranges from 1 to the number of templates. The CMC curve obtained using the chosen descriptor and similarity measure (see Sect. 4.3.1) is shown in Figure 9. For instance, this curve shows that for about 10% of the query images the template image of the corresponding identity was in the first position of the ranked list returned to the operator; for about 40% of the query images it was among the top 50 positions, and so on. This performance is similar to the one reported in the literature for the same kind of descriptors and similarity measures.

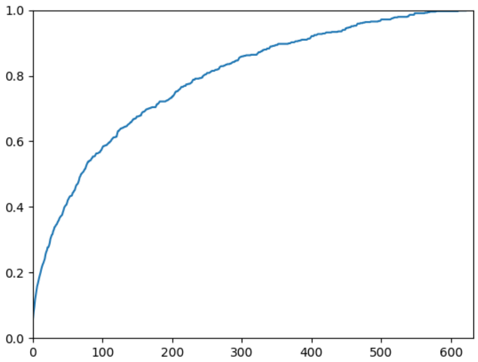


Figure 9 – CMC curve obtained on the VIPeR data set using the image descriptor and similarity measure described in Sect. 4.3.1.

5.1.5 Future work

Ongoing work is aimed at:

- completing the GUI: integration with video sources, development of the tool for selecting the query image from a still video frame;
- integrating a pedestrian detection software tool;
- implementing the HITL functionality.

5.2 PEOPLE SEARCH TOOL

5.2.1 Percentage complete

The first version of this tool is 70% complete. With reference to the architecture of Sect. 4.1.2:

- The same **pedestrian detection** software tool as the person re-identification module will be used.
- The **feature extractor** has already been implemented.
- The **attribute detectors** have already been trained.
- The **retrieval module** has not been implemented yet.
- The **HITL module** has not been implemented yet.

- The **GUI** is under development (75% complete: it shares several components with the GUI of the re-identification tool).

5.2.2 How implemented

The same programming languages, libraries and frameworks as the person re-identification tool are being used (see Sect. 5.1.2).

5.2.3 Validation

The attribute detectors have been trained and validated on the PETA data set. Their detection accuracy is similar to the one reported in the literature (49). Detection accuracy is relatively high, ranging from 0.6329 for the attribute 'lower body trousers' to 0.9976 for 'carrying umbrella'. However detection accuracy is not indicative of the actual performance in the cases when very few images contain the attribute of interest; in this case the Pr and Re measures are more suitable; for instance, for an attribute like 'accessory Headphone' which is present only in 31 out of 19,000 images, detection accuracy is 0.9887, but Pr and Re are respectively 0.0175 and 0.0323, which denotes a very low performance. Accordingly, it is planned that only a subset of the attributes will be used in this tool, selected among the ones which exhibit sufficiently high Pr and Re values.

5.2.4 Future work

Ongoing work is aimed at:

- implementing the retrieval module;
- completing the GUI: integration with video sources, selection of the query attribute profile;
- integrating a pedestrian detection software tool (the same as the person re-identification tool);
- implementing the HITL functionality.

5.3 CROWD MONITORING TOOL

This tool is currently under design, and implementation details will be reported in deliverable D5.8.

6 CONCLUSIONS

Task T5.4 aims at developing a human-centred computer vision tool to support LEAs operators in monitoring and investigation tasks related to the security of mass gathering events. The HCV tool will include three components which will provide functionalities related to the analysis of video surveillance footage: allowing LEAs operators and forensic investigators to retrieve images of an individual of interest, based either on an image of that individual (person re-identification) or a description of its appearance in terms of an attribute profile related to clothing appearance and to soft biometrics like gender (people search); providing a crowd monitoring functionality, including crowd density estimation, detection of patterns of crowd movement, and detection of anomalous/suspicious crowd behaviours. In particular, the crowd monitoring tool will be integrated with the Dynamic Risk Assessment tool of work package W3 (described in deliverables D3.2 and D3.4), and with the Crowd Modelling tool of task T5.1 (described in deliverable D5.1).

This deliverable described the preliminary steps of the development of the HCV tool (literature survey, definition and analysis of user and system requirements), the design of its three components (design of the crowd monitoring tool is under way), and the state of their implementation (currently only the person re-identification and people search tool are under implementation), including the planned integration with the DRA and CM tools.

The next steps will be the following:

- completing the design of the crowd monitoring tool, and the implementation of the first version of all three tools;

- as soon as a first version of each component of the HCV tool is available, LEAs will be asked to carry out a preliminary validation on publicly available video surveillance data sets;
- defining the details of the integration with the DRA and CM tools;
- validating the HCV tool in the first version of the practical demonstrations planned in work package WP6.

The final version of the HCV tool will be described in the second version of this deliverable, D5.8.

7 REFERENCES AND ACRONYMS

7.1 REFERENCES

1. **Vezzani, R., Baltieri, D. and Cucchiara, R.** People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys.* 2013, Vol. 46, 2, pp. 37, Article 29.

2. *Person re-identification using spatio-temporal appearance.* **Gheissari, N., Sebastian, T.B. and Hartley, R.** : IEEE, 2006. Proc. Int. Conf. on Computer Vision and Pattern Recognition. Vol. 2, pp. 1528-1535.

3. **Karanam, S., et al.** A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets. *IEEE Trans. on Pattern Analysis and Machine Intelligence.* In press.

4. **Zheng, L., Yang, Y. and Hauptmann, A.G.** Person Re-identification: Past, Present and Future. *CoRR.* 2016, Vol. abs/1610.02984.

5. *Person re-identification by symmetry-driven accumulation of local features.* **Farenzena, M., et al.** : IEEE, 2010. Proc. Int. Conf. on Computer Vision and Pattern Recognition. pp. 2360-2367.

6. *Large scale metric learning from equivalence constraints.* **Köstinger, M., et al.** : IEEE, 2012. Proc. Int. Conf. on Computer Vision and Pattern Recognition. pp. 2288-2295.

7. *Visual Recognition with Humans in the Loop.* **Branson, S. et al.** : Springer, 2010. Proc. European Conf. on Computer Vision.

8. **Ali, S., et al.** Interactive retrieval of targets for wide area surveillance. *ACM Multimedia.* 2010, pp. 895-898.

9. *Person Re-identification by Descriptive and Discriminative Classification.* **Hirzer, M., et al.** : Springer, 2011. Proc. Scandinavian Conf. on Image Analysis. pp. 91-102.

10. *POP: Person Re-identification Post-rank Optimisation.* **Liu, C., et al.** : IEEE, 2013. Proc. Int. Conf. on Computer Vision. pp. 441-448.

11. **Metternich, M.J. and Worring, M.** Track based relevance feedback for tracing persons in surveillance videos. *Computer Vision and Image Understanding.* 2013, Vol. 117, 3, pp. 229-237.

12. *Active image pair selection for continuous person re-identification.* **Das, A., Panda, R. and Roy-Chowdhury, A.** : IEEE, 2015. Proc. Int. Conf. on Image Processing. pp. 4263-4267.

13. *Region-Based Interactive Ranking Optimization for Person Re-identification.* **Wang, Z., et al.** s.l. : Springer, 2014. Proc. Pacific Rim Conference on Multimedia. pp. 1-10.

14. *Human-in-the-Loop Person Re-identification.* **Wang, H., et al.** : Springer, 2016. Proc. European Conf. on Computer Vision. Vol. IV, pp. 405-422.

15. *Person attribute search for large-area video surveillance.* **Thornton, J., et al.** : IEEE, 2011. Proc. Int. Conf. on Technologies for Homeland Security. pp. 55-61.

16. *A General Method for Appearance-Based People Search Based on Textual Queries.* **Satta, R., Fumera, G. and Roli, F.** : Springer, 2012. Proc. European Conf. on Computer Vision Workshops and Demonstrations.

17. *Describing objects by their attributes.* **Farhadi, A., et al.** : IEEE, 2009. Proc. Int. Conf. on

Computer Vision and Pattern Recognition. pp. 1778-1785.

18. **Layne, R., Hospedales, T.M. and Gong, S.** Attributes-Based Re-identification. [ed.] S. Gong, et al. *Person Re-Identification.* : Springer, 2014, pp. 93-117.

19. *Deep Attributes Driven Multi-camera Person Re-identification.* **Su, C., et al.** : Springer, 2016. Proc. European Conference on Computer Vision. Vol. 2, pp. 475-491.

20. *Attributes co-occurrence pattern mining for video-based person re-identification.* **Zhang, X., Pala, F. and Bhanu, B.** : IEEE, 2017. Proc. Int. Conf. on Advanced Video and Signal Based Surveillance. pp. 1-6.

21. *Person Re-identification by Deep Learning Attribute-Complementary Information.* **Schumann, A. and Stiefelhagen, R.** : IEEE, 2017. Proc. Int. Conf. on Computer Vision and Pattern Recognition Workshops. pp. 1435-1443.

22. **Su, C., et al.** Multi-Task Learning with Low Rank Attribute Embedding for Multi-Camera Person Re-identification. *IEEE Trans. on Pattern Analysis and Machine Intelligence.* 2018, Vol. 40, 5, pp. 1167-1181.

23. **Su, C., et al.** Multi-type attributes driven multi-camera person re-identification. *Pattern Recognition.* 2018, Vol. 75, pp. 77-89.

24. **Li, A., et al.** Clothing Attributes Assisted Person Reidentification. *IEEE Trans. on Circuits and Systems for Video Technology.* 2015, Vol. 25, 5, pp. 869-878.

25. **Kok, V.J., Lim, M.K. and Chan, C.S.** Crowd behavior analysis: A review where physics meets biology. *Neurocomputing.* 2016, Vol. 177, pp. 342-362.

26. **Zhan, B., Monekosso, D.N. and Remagnino, P. et al.** Crowd analysis: a survey. *Machine Vision and Applications.* 2008, Vol. 19, pp. 345-357.

27. **Silveira Jacques Jr, J.C., Musse, S.R. and Jung, C.R.** Crowd Analysis Using Computer Vision Techniques. *IEEE Signal Processing Magazine.* 2010, Vol. 27, 5, pp. 66-77.

28. **Thida, M., et al.** A Literature Review on Video Analytics of Crowded Scenes. [ed.] P. Atrey, M. Kankanhalli and A. Cavallaro. *Intelligent Multimedia Surveillance.* : Springer, 2013, pp. 17-36.

29. **Loy, C.C., et al.** Crowd Counting and Profiling: Methodology and Evaluation. [ed.] S. Ali, et al. *Modeling, Simulation and Visual Analysis of Crowds.* : Springer, 2013, pp. 347-382.

30. **Li, T., et al.** Crowded Scene Analysis: A Survey. *IEEE Trans. on Circuits and Systems for Video Technology.* 2015, Vol. 25, 3, pp. 367-386.

31. **Zitouni, M.S., et al.** Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques. *Neurocomputing.* 2016, Vol. 186, pp. 139-159.

32. **Isard, M. and Blake, A.** CONDENSATION conditional density propagation for visual tracking. *Int. Journal of Computer Vision.* 1998, Vol. 29, 1, pp. 5-28.

33. **Ge, W., Collins, R.T. and Ruback, R.B.** Vision-Based Analysis of Small Groups in Pedestrian Crowds. *IEEE Trans. on Pattern Analysis and Machine Intelligence.* 2012, Vol. 34, 5, pp. 1003-1016.

34. **Solera, F., Calderara, S. and Cucchiara, R.** Socially Constrained Structural Learning for Groups Detection in Crowd. *IEEE Trans. on Pattern Analysis and Machine Intelligence.* 2016, Vol. 38, 5, pp. 995-1008.

35. **Helbing, D. and Molnár, P.** Social force model for pedestrian dynamics. *Physics Review E.* 1995, Vol. 51, 5, pp. 4282-4286.

36. **Solmaz, B., Moore, B.E. and Shah, M.** Identifying Behaviors in Crowd Scenes Using Stability Analysis for Dynamical Systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence.* 2012, Vol. 34, 10, pp. 2064-2070.

37. **Wu, S., et al.** Crowd Behavior Analysis via Curl and Divergence of Motion Trajectories. *Int. J.*

Computer Vision. 2017, Vol. 123, pp. 499-519.

38. **Sodemann, A.A., Ross, M.P. and Borghetti, B.J.** A Review of Anomaly Detection in Automated Surveillance. *IEEE Trans. on Systems, Man, and Cybernetics, Part C (Applications and Reviews).* 2012, Vol. 42, 6, pp. 1257-1272.

39. *Abnormal crowd behavior detection using social force model.* **Mehran, R., Oyama, A. and Shah, M.** : IEEE, 2009. Proc. Int. Conf. on Computer Vision and Pattern Recognition. pp. 935-942.

40. **Ferryman, J., et al.** Robust abandoned object detection integrating wide area visual surveillance and social context. *Pattern Recognition Letters.* 2013, Vol. 34, pp. 789-798.

41. *Person Re-identification in the Wild.* **Zheng, L., et al.** : IEEE, 2017. Proc. Int. Conf. on Computer Vision and Pattern Recognition. pp. 3346-3355.

42. **Pham, T.T.T., et al.** Fully-automated person re-identification in multi-camera surveillance system with a robust kernel descriptor and effective shadow removal method. *Image and Vision Computing.* 2017, Vol. 59, pp. 44-62.

43. **Dollár, P., et al.** Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Trans. on Pattern Analysis and Machine Intelligence.* 2012, Vol. 34, 4, pp. 743-761.

44. **Ren, S., et al.** Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence.* 2017, Vol. 39, 6, pp. 1137-1149.

45. **Zhang, S., et al.** Towards Reaching Human Performance in Pedestrian Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence.* 2018, Vol. 40, 4, pp. 973-986.

46. **Camps, O., et al.** From the Lab to the Real World: Re-identification in an Airport Camera Network. *IEEE Trans. on Circuits and Systems for Video Technology.* 2017, Vol. 27, 3, pp. 540-553.

47. *Unbiased look at dataset bias.* **Torralba, A. and Efros, A.A.** : IEEE, 2011. Proc. Int. Conf. on Computer Vision and Pattern Recognition. pp. 1521-1528.

48. **Hand, D.J.** Classifier Technology and the Illusion of Progress. *Statistical Science.* 2006, Vol. 21, 1, pp. 1-14.

49. **Deng, Y., et al.** Pedestrian Attribute Recognition At Far Distance. *ACM Multimedia.* 2014, pp. 789-792.

50. **Idrees, H., Warner, N. and Shah, M.** Tracking in dense crowds using prominence and neighborhood motion concurrence. *Image and Vision Computing.* 2014, Vol. 32, pp. 14-26.

51. **Li, W., Mahadevan, V. and Vasconcelos, N.** Anomaly Detection and Localization in Crowded Scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence.* 2014, Vol. 36, 1, pp. 18-32.

52. **Pennisi, A., Bloisi, D.D. and Iocchi, L.** Online real-time crowd behavior detection in video sequences. *Computer Vision and Image Understanding.* 2016, Vol. 144, pp. 166-176.

53. **Wang, J. and Xu, Z.** Spatio-temporal texture modelling for real-time crowd anomaly detection. *Computer Vision and Image Understanding.* 2016, Vol. 144, pp. 177-187.

54. *Mars: A video benchmark for large-scale person re-identification.* **Zheng, L., et al.** : Springer, 2016. Proc. European Conf. on Computer Vision. pp. 868-884.

55. *DukeMTMC4ReID: A Large-Scale Multi-camera Person Re-identification Dataset.* **Gou, M., et al.** : IEEE, 2017. Proc. Int. Conf. Computer Vision and Pattern Recognition Workshops. pp. 1425-1434.

56. *Pedestrian Attribute Classification in Surveillance: Database and Evaluation.* **Zhu, J., et al.** : IEEE, 2013. Proc. Int. Conf. on Computer Vision Workshops. pp. 331-338.

57. **Li, D., et al.** A Richly Annotated Dataset for Pedestrian Attribute Recognition. *CoRR.* 2016, Vol. abs/1603.07054.

58. *Privacy preserving crowd monitoring: Counting people without people models or tracking.* **Chan, A. B., Liang, Z.-S.J. and Vasconcelos, N.** : IEEE, 2008. Proc. Int. Conf. on Computer Vision and Pattern Recognition. pp. 1-7.

59. *Multi-source Multi-scale Counting in Extremely Dense Crowd Images.* **Idrees, H., et al.** : IEEE, 2013. Proc. Int. Conf. on Computer Vision. pp. 2547-2554.

60. *Cross-scene Crowd Counting via Deep Convolutional Neural Networks.* **Zhang, C., et al.** : IEEE, 2015. Proc. Int. Conf. on Computer Vision and Pattern Recognition. pp. 833-841.

61. *A Lagrangian Particle Dynamics Approach for Crowd Flow Segmentation and Stability Analysis.* **Ali, S. and Shah, M.** : IEEE, 2007. Proc. Int. Conf. on Computer Vision and Pattern Recognition.

62. *Real-time detection of violent crowd behavior.* **Hassner, T., Itcher, Y. and Kliper-Gross, O.** : IEEE, 2012. Proc. Int. Conf. on Computer Vision and Pattern Recognition Workshops. pp. 1-6.

63. *Floor Fields for Tracking in High Density Crowd Scenes.* **Ali., S., Shah, M.** : Springer, 2008. Proc. European Conf. on Computer Vision. Vol. 2, pp. 1-14.

64. *Understanding collective crowd behaviors: Learning a Mixture model of Dynamic pedestrian-Agents.* **Zhou, B., Wang, X. and Tang, X.** : IEEE, 2012. Proc. Int. Conf. on Computer Vision and Pattern Recognition. pp. 2871-2878.

65. *Scene-Independent Group Profiling in Crowd.* **Shao, J., Loy, C.C. and Wang, X.** : IEEE, 2014. Proc. Int. Conf. on Computer Vision and Pattern Recognition. pp. 2227-2234.

66. **Courty, N., et al.** Using the AGORASET dataset: Assessing for the quality of crowd video analysis methods. *Pattern Recognition Letters.* 2014, Vol. 44, pp. 161-170.

67. *Matching people across camera views using kernel canonical correlation analysis.* **Lisanti, G., Masi, I. and Del Bimbo, A.** : ACM, 2014. ACM Int. Conf. on Distributed Smart Cameras.

68. **Zhu, X., et al.** Image to Video Person Re-Identification by Learning Heterogeneous Dictionary Pair With Feature Projection Matrix. *IEEE Trans. on Information Forensics and Security.* 2018, Vol. 13, 3, pp. 717-732.

69. **Xiao, T., et al.** End-to-End Deep Learning for Person Search. *CoRR.* 2016, Vol. abs/1604.01850.

70. *Attribute-based people search in surveillance environments.* **Vaquero, D.A., et al.** : IEEE, 2009. Workshop on Applications of Computer Vision. pp. 1-8.

7.2 ACRONYMS

Acronyms List	
CCTV	Closed-Circuit Television
CM	Crowd Modelling (tool)
CV	Computer Vision
DoW	Description of Work
DRA	Dynamic Risk Assessment
FP	False Positive
fps	frames per second
GUI	Graphical User Interface
HCV	Human-centred Computer Vision
HITL	Human-In-The-Loop
LEA	Law Enforcement Agency
Pr	Precision
Re	Recall
RPAS	Remotely Piloted Aircraft System
TP	True Positive
VS	Video Surveillance

TABLE 4 – Acronyms