



Title:	Document Version:
D5.8 Human-centred computer vision for crowd protection	0.3

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*:
M24 (April 2019)	M24 (April 2019)	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):
Giorgio Fumera (UNICA)

Abstract:
<p>This document describes the Human-centred Computer Vision tool developed in the context of work package WP5. It consists of a prototype software tool with three integrated functionalities to support operators and officers of Law Enforcement Agencies in the use of video surveillance systems to search for individuals of interest, such as suspect individuals, in large amounts of video data (appearance-based person re-identification and attribute-based people search) and to monitor in real time the size of a crowd (crowd density estimation). This document reports a comprehensive literature survey, an analysis of the state of the art, definition of requirements, the design of the HCV tool modules, their implementation and validation.</p>

Keywords:
<p>Computer vision, video surveillance, appearance-based person re-identification, attribute-based people search, crowd monitoring, crowd density estimation, detection of anomalous crowd behaviours, human-in-the-loop computer vision</p>

## Revision History

Revision	Date	Description	Author (Organisation)
V0.1	25.04.2019	First draft ready for peer review	Giorgio Fumera (UNICA)
V0.2	07.05.2019	Comments from ETRA, peer review comments by BayHfoeD and ZB	Jordi Arias Martí (ETRA), Gorka Sanz Monllor (ETRA), Sebastian Allertseder (BayHfoeD), Carlo Dambra (ZB)
V0.3	08.05.2019	Revised version, ready to be submitted to the EC	Giorgio Fumera (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

The work described in this document has been conducted within the LETSCROWD project. This document reflects only the LETSCROWD Consortium view and the European Union is not responsible for any use that may be made of the information it contains.

This document and its content are the property of the LETSCROWD Consortium. All rights relevant to this document are determined by the applicable laws. Access to this document does not grant any right or license on the document or its contents. This document or its contents are not to be used or treated in any manner inconsistent with the rights or interests of the LETSCROWD Consortium or the Partners detriment and are not to be disclosed externally without prior written consent from the LETSCROWD Partners.

Each LETSCROWD Partner may use this document in conformity with the LETSCROWD Consortium Grant Agreement provisions.

**Video surveillance systems** are nowadays pervasively deployed in public places such as streets, stations, airports and stadiums. They are widely used by Law Enforcement Agencies (LEAs), and have a crucial role in mass gathering events both for monitoring crowds during event execution and for post-event forensic investigations of crimes and incidents. However the number of video cameras of current CCTV systems and the corresponding amount of videos recorded during mass gathering events make it very challenging and time-consuming for LEA operators and forensic investigators to exploit them to effectively carry out their monitoring and investigation tasks. Accordingly, there is an increasing interest in **video analytics** solutions derived from **computer vision** research outcomes that can support human operators in such tasks.

### INTELLIGENT VIDEO SURVEILLANCE

Since over twenty years the **intelligent video surveillance** computer vision field deals with the development of algorithms based on pattern recognition and machine learning techniques for various tasks such as pedestrian detection and tracking, people counting, person re-identification, crowd behaviour understanding, and detection of anomalous events and crowd behaviours. Video analytics functionalities based on research outcomes of this field have been deployed since the 2000s by companies (often academic spin-offs) and solution providers. However, whereas reliable tools are already available for specific tasks under controlled conditions (e.g., tracking or counting a few pedestrians in absence of overlapping and occlusions), the performance of state-of-the-art computer vision algorithms is still unsatisfactory for complex and unconstrained real-world application scenarios such as the ones related to crowd monitoring in mass gathering events.

### THE HUMAN-CENTRED COMPUTER VISION TOOL

This deliverable describes the **human-centred computer vision** (HCV) software tool which has been developed in the context of work package WP5 of the LETSCROWD project. The HCV tool provides a set of functionalities aimed at supporting LEA operators and forensic investigators in two common kinds of monitoring and investigation tasks related to the security of mass gathering events, which are still challenging for current computer vision algorithms: (i) estimating in real time the size of a crowd ("crowd density estimation") and detecting related anomalous events such as overcrowding, and (ii) searching for individuals of interest in recorded videos, during forensic investigations, starting either from an image of the individual of interest ("appearance-based person re-identification") or from a description provided by an eyewitness ("attribute-based people search").

- **Appearance-based person re-identification.** Scanning a large amount of recorded videos to search for a suspect individual who has been observed in one of the available videos is a common task in forensic investigations. Whereas face recognition technology is effective under controlled conditions, in unconstrained video surveillance scenarios typical of mass gatherings in large outdoor or indoor venues the face of individuals may be not visible due to pose, occlusion or distance to the camera. In this case *clothing appearance* is the main available cue. Starting from a query image of an individual of interest manually selected by the user from a video frame, the appearance-based person re-identification module of the HCV tool automatically retrieves and ranks images of pedestrians detected in recorded videos, based on clothing appearance to the query image. The user can then access the video from which each retrieved image was extracted, to analyse the behaviour of the corresponding individual. This functionality can allow users to find individuals of interest in a much lower time than the one required by manually scanning the available videos.
- **Attribute-based people search.** A similar forensic investigation scenario is the one when a description of a suspect individual is available, typically provided by eyewitnesses, instead of an image. The description may involve clothing appearance, as well as attributes such as gender and carried items (e.g., bags or backpacks). The attribute-based people search module of the HCV tool allows the user to input the *attribute profile* of an individual of interest in terms of a predefined set of attributes, such as upper and

lower body clothing colours, gender, and carried items; it then automatically retrieves images of pedestrians detected in recorded videos, and ranks them based on the similarity to the target attribute profile.

- **Crowd density estimation and anomalous behaviours detection.** Estimating the size of a crowd from a video, during a mass gathering event, is a common but time-consuming task for LEA operators. The crowd density estimation module of the HCV tool provides a real-time, frame-by-frame estimate of the number of people in a scene, either in the whole video frame or in a user-defined region of interest. It is focused on crowded scenes with significant overlapping between people and partial occlusions by other objects, where an exact count is unfeasible. This module can also raise alerts when the estimated number of people exceeds a user-defined threshold, and when anomalous and potentially dangerous behaviours are automatically detected, such as a sudden increase or decrease of the estimated number of people.

A distinctive feature of the HCV tool is the use of two kinds of **solutions for improving the accuracy** of the underlying computer vision algorithms in the real-world, challenging application scenarios considered in the LETSCROWD project:

- In the appearance-based person re-identification and people search modules the user can provide a **feedback** about the actual similarity of any subset of retrieved images of pedestrians, respectively to the individual of interest in the query image and to the target attribute profile. Both modules can then exploit such a feedback to re-rank all the retrieved images, with the aim of moving toward the top ranks the ones more similar to the query image or attribute profile.
- The crowd density estimation module includes a machine learning algorithm which has to be trained off-line on a set of images of crowds *manually* annotated with the exact number of people they contain. The accuracy of this module relies on the presence in the training set of images representative of the ones that will be later processed during operation, in terms of camera view, number and location of people in the scene, and background. This is a very strong requirement which is nearly impossible to satisfy in real-world applications. Nevertheless, in the context of the human-centred toolkit developed in the work package WP5 this requirement can become feasible by exploiting **synthetic videos** generated by the Crowd Modelling and Planning tool: all the above mentioned conditions can be controlled in synthetic videos, including the exact number of people in each video frame.

## Index

<b>LIST OF FIGURES.....</b>	<b>7</b>
<b>LIST OF TABLES.....</b>	<b>8</b>
<b>1 INTRODUCTION.....</b>	<b>9</b>
1.1 PURPOSE OF THE DOCUMENT .....	9
1.2 SCOPE OF THE DOCUMENT .....	9
1.3 STRUCTURE OF THE DOCUMENT .....	9
<b>2 BACKGROUND RESEARCH.....</b>	<b>10</b>
2.1 LITERATURE REVIEW.....	10
2.1.1 APPEARANCE-BASED PERSON RE-IDENTIFICATION .....	10
2.1.2 ATTRIBUTE-BASED PEOPLE SEARCH .....	14
2.1.3 CROWD MONITORING .....	15
2.2 EXISTING STATE OF THE ART .....	19
2.2.1 PERSON RE-IDENTIFICATION AND PEOPLE SEARCH.....	19
2.2.2 CROWD MONITORING .....	20
2.2.3 PERFORMANCE OF STATE-OF-THE-ART COMPUTER VISION ALGORITHMS.....	21
2.2.4 AVAILABLE SOFTWARE .....	23
2.2.5 COMMERCIAL SOLUTIONS .....	24
2.3 DEVELOPMENT OVER THE STATE OF THE ART .....	25
2.3.1 HOW THE HCV TOOL ADVANCES THE CURRENT STATE OF THE ART.....	25
2.3.2 WHAT HAS BEEN ACHIEVED .....	26
2.4 PUBLICLY AVAILABLE DATA SETS .....	26
<b>3 REQUIREMENTS .....</b>	<b>31</b>
3.1 USE CASE .....	31
3.2 USER REQUIREMENTS.....	33
3.3 SYSTEM REQUIREMENTS .....	35
<b>4 DESIGN .....</b>	<b>37</b>
4.1 ARCHITECTURE.....	38
4.1.1 PERSON RE-IDENTIFICATION MODULE.....	38
4.1.2 PEOPLE SEARCH MODULE .....	39
4.1.3 CROWD MONITORING MODULE.....	40
4.2 LINKS/INTERFACES BETWEEN TASKS/WORK PACKAGES .....	41
4.3 ALGORITHM DESCRIPTION .....	42
4.3.1 PERSON RE-IDENTIFICATION MODULE.....	42
4.3.2 PEOPLE SEARCH MODULE .....	43
4.3.3 CROWD MONITORING MODULE.....	45
4.4 USER INTERFACE .....	46

4.4.1	PERSON RE-IDENTIFICATION MODULE .....	46
4.4.2	PEOPLE SEARCH MODULE .....	50
4.4.3	CROWD MONITORING MODULE.....	53
<b>5</b>	<b><u>IMPLEMENTATION .....</u></b>	<b>55</b>
<b>5.1</b>	<b>PERSON RE-IDENTIFICATION MODULE .....</b>	<b>55</b>
5.1.1	IMPLEMENTATION DETAILS .....	55
5.1.2	VALIDATION.....	56
<b>5.2</b>	<b>PEOPLE SEARCH MODULE.....</b>	<b>58</b>
5.2.1	IMPLEMENTATION DETAILS .....	58
5.2.2	VALIDATION.....	58
<b>5.3</b>	<b>CROWD MONITORING MODULE.....</b>	<b>59</b>
5.3.1	IMPLEMENTATION DETAILS .....	59
5.3.2	VALIDATION.....	59
<b>6</b>	<b><u>CONCLUSIONS.....</u></b>	<b>60</b>
<b>7</b>	<b><u>REFERENCES AND ACRONYMS .....</u></b>	<b>61</b>
7.1	REFERENCES .....	61
7.2	ACRONYMS.....	65
<b>8</b>	<b><u>ANNEX A – HCV TOOL USER GUIDE .....</u></b>	<b>66</b>
<b>8.1</b>	<b>PERSON RE-IDENTIFICATION FUNCTIONALITY .....</b>	<b>66</b>
<b>8.2</b>	<b>PEOPLE SEARCH MODULE.....</b>	<b>71</b>
<b>8.3</b>	<b>CROWD MONITORING MODULE.....</b>	<b>74</b>



## LIST OF FIGURES

Figure 1 – Core functionality of an appearance-based person re-identification system.....	11
Figure 2 – Core functionality of an attribute-based people search system. ....	14
Figure 3 – Top: patterns of crowd movement detected by the method of (46). Right: patterns detected by (47).....	18
Figure 4 – Example of images from the VIPeR person re-identification data set: pairs of images (128×48 pixels, BMP format) of ten different individuals taken from two different cameras.....	27
Figure 5 – Functional architecture of the person re-identification module.....	39
Figure 6 – Functional architecture of the people search module. ....	40
Figure 7 – Functional architecture of the crowd density estimation module.....	41
Figure 8 – Initial window (Web page) of the HCV tool GUI. ....	46
Figure 9 – GUI of the person re-identification module: main window. ....	47
Figure 10 – Selection of a bounding box containing the image of an individual in the main window of the person re-identification GUI.....	47
Figure 11 – Selection of the query image and start of the retrieval process from the main window of the person re-identification GUI.....	48
Figure 12 – Results window of the person re-identification GUI: query image (right), and ranked list of the retrieved template images together with context information and user feedback controls. ....	49
Figure 13 – Template detail window of the person re-identification GUI. ....	50
Figure 14 – Main window of the people search GUI. ....	51
Figure 15 – Results window of the people search GUI, corresponding to the attribute profile shown in Figure 14.....	52
Figure 16 – Template detail window of the people search GUI. ....	53
Figure 17 – GUI of the crowd density estimation module. ....	54
Figure 18 – GUI of the crowd density estimation module: selection of a region of interest (red rectangle on the left). ....	54
Figure 19 – CMC curves obtained on the VIPeR (left) and Market-1501 (right) data sets.....	56
Figure 20 – Starting window (Web page) of the HCV tool. ....	66
Figure 21 – Main window of the person re-identification functionality. ....	67
Figure 22 – Selection of a bounding box containing the image of an individual of interest.....	68
Figure 23 – Choice of the <i>query</i> image of the individual of interest, and start of the retrieval operation. ....	69
Figure 24 – Results window of the person re-identification module.....	70
Figure 25 – Template Detail window of the person re-identification module.....	71
Figure 26 – Main window of the people search module.....	72
Figure 27 – Results window of the people search module. ....	73
Figure 28 – Template Detail window of the people search module. ....	73
Figure 29 – indow of the crowd density estimation module. ....	74

Figure 30 – Selection of a region of interest in the video processed by the crowd density estimation module. ....75

## LIST OF TABLES

Table 1 – Publicly available data sets for person re-identification, suitable to the person re-identification module of the HCV tool. ....28

Table 2 – Publicly available data sets with annotated pedestrian attributes, suitable to the people search module of the HCV tool. ....29

Table 3 – Main publicly available data sets related to crowd monitoring tasks. ....31

TABLE 4 – Acronyms .....65





## 1 INTRODUCTION

### 1.1 PURPOSE OF THE DOCUMENT

This document is the final report on task T5.4, which is aimed at developing a human-centred computer vision (HCV) tool to support Law Enforcement Agency (LEA) operators in the use of video surveillance systems for monitoring and investigation tasks related to the security of mass gathering events. The HCV tool provides three functionalities as distinct software modules accessible under the same user interface:

- (1) **appearance-based person re-identification**: retrieving images of an individual of interest starting from a query image of that individual, based on clothing appearance similarity;
- (2) **attribute-based people search**: retrieving images of individuals whose appearance matches a given attribute profile defined in terms of a predefined set of clothing appearance characteristics and of attributes like gender and carried items;
- (3) **crowd monitoring**: estimating the number of people in a scene ("crowd density estimation"), and detecting related anomalous behaviours such as overcrowding.

This document is an update of deliverable D5.4, where the status of the HCV tool at M12 was reported, and describes the work that has lead to the development and implementation of the HCV tool: analysis of background research; definition and analysis of requirements; design, implementation and validation of the three modules listed above, and integration with other tools developed in the LETSCROWD project (Crowd Modelling and Planning tool, Dynamic Risk Assessment tool and LETSCROWD server). A user guide for LEAs is included in this document 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. As planned in the Description of Work, the HCV tool is integrated with the crowd modelling and planning tool developed in task T5.1 and with the dynamic risk assessment tool of work package WP3, via the LETSCROWD server developed in the work package WP4.

### 1.3 STRUCTURE OF THE DOCUMENT

This document is structured into four main sections. Background research on the three computer vision tasks involved in the HCV tool (appearance-based person re-identification, attribute-based people search and crowd monitoring) is reported in Sect. 2, including a literature review, a survey of related commercial products and an analysis of publicly available data sets. User and system requirements, and the use case for the HCV tool, previously defined in work package WP2, are summarized in Sect. 3, which includes also a discussion about privacy issues. The design of the HCV tool is described in Sect. 4, including the links to other tasks and work packages. Finally, implementation details are reported in Sect. 5. A user guide for LEAs is reported in Annex A.

## 2 BACKGROUND RESEARCH

The computer vision research community is devoting a considerable effort to topics related to video surveillance 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 video surveillance tools capable of supporting human operators in monitoring and forensic investigation tasks, partially automatizing them. Since the 2000s a few video analytics companies (usually academic spin-offs) and solution providers started to deploy some computer vision tools in commercial products. 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 video surveillance 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, computer vision 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 appearance-based person re-identification.

This section gives an overview of the research and applications of computer vision for intelligent video surveillance, focusing on the functionalities of the 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 tool are then pointed out; finally, the main publicly available data sets pertinent to the functionalities of the HCV tool are described, including the ones that have been used to develop, validate and demonstrate this tool, further considered in section 4.

### 2.1 LITERATURE REVIEW

In this subsection 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 a reference.

#### 2.1.1 Appearance-based person re-identification

Appearance-based 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 video surveillance network, using an available image of that individual as a *query* (or *probe*), on the basis of clothing appearance (e.g., clothing colour and texture) (1). A typical application is to support forensic investigators to search for a suspect individual over the videos recorded by a video surveillance system. The main goal of an appearance-based person re-identification system is to **reduce the time** that LEA operators and forensic investigators would have to spend by watching all the available videos when searching for an individual of interest, taking into account that the number and duration of videos acquired by existing video surveillance camera networks can be very high.

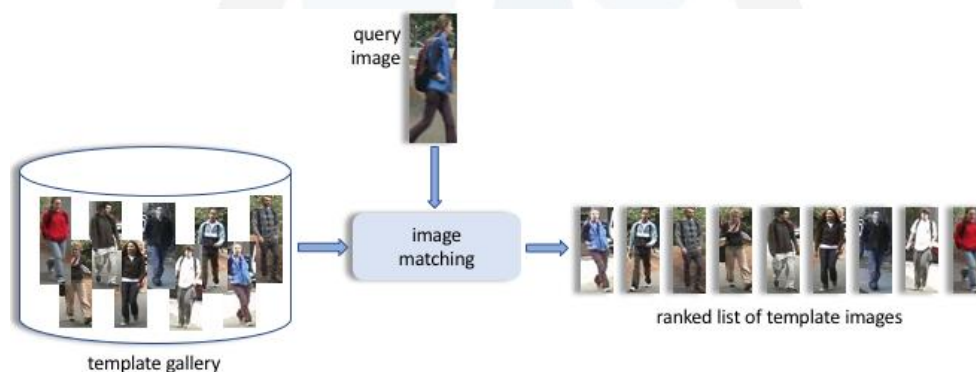
Appearance-based person re-identification focuses on typically unconstrained video surveillance settings, which are characterized by relatively low camera resolution, pose and lighting variations, occlusions, and differences between cameras. Such features make recognition technology based on "strong" 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 span. 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 900.<sup>1</sup> It is worth noting that an analogous

---

<sup>1</sup> Source: dblp computer science bibliography, <https://dblp.uni-trier.de/>, search query: "person re-identification", visited on April 13, 2019. It is worth noting that in March 2018, when the first version of this deliverable (D5.4) was written, about 600 publications of the same types mentioned in the text were listed in the dblp database:

computer vision task is the one of *face* re-identification: as its name suggests, it uses face recognition technology, and is thus applicable to more constrained settings where the face of individuals is relatively well visible (e.g., a security checkpoint in an airport). The HCV tool focuses only on the most challenging, unconstrained video surveillance settings, and thus only on the appearance-based person re-identification task. The two tasks can nevertheless be considered as complementary, since face re-identification (when feasible) can be strengthened by appearance-based cues. For the sake of brevity from now on the term "appearance-based" will be omitted when referring to "appearance-based person re-identification", if no ambiguity can arise.

The core functionality of a person re-identification system, summarized in 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, using 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 video surveillance system. The template gallery is instead made up of images of individuals automatically detected and extracted from videos acquired by the same video surveillance system. Ideally, the gallery images of the query identity (if any) will appear at the top of the returned ranked list. As a variant, the query and the template individuals can be represented as *tracks* extracted by a pedestrian tracking software tool (i.e., multiple images of the same individual detected in a video sequence); this may allow achieving a higher robustness to pose and lighting variations and to occlusions. In a deployed person re-identification system contextual information should be provided to the operator about the retrieved template images, such as the camera ID and position in the monitored area, the timestamp, and an access to the video from which an image was extracted; this would allow an operator to further analyse the behaviour of the query individual, if images of that individual are found among the ones retrieved by the person re-identification system.



**Figure 1 – Core functionality of an appearance-based person re-identification system: matching a query image of an individual, manually selected by an operator, to a set of images of individuals automatically extracted by a pedestrian detection or tracking tool (template gallery), and sorting such images 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: (i) pedestrian detection and (ii) pedestrian bounding box extraction. They consist in automatically detecting pedestrians in raw videos, and then extracting their images from each video frame, in terms of *bounding boxes*, i.e., rectangular image regions. To minimize interference with the background, nearby objects or other people, bounding boxes should be as tight as possible; ideally, only the silhouette of an individual should be extracted. Tracking algorithms can also be used either to extract tracks instead of single images (as mentioned above), or to

---

this means that the number of publications in the past twelve months accounts for half of the publications in the past twelve years.

improve detection accuracy. It is important to point out that pedestrian detection and tracking are relevant and independent research topics in the computer vision field (as specific instances of the more general and older *object* detection and tracking problems), and that they have been addressed as distinct tasks from person re-identification. Nevertheless, it is known that the effectiveness of the pedestrian detection and bounding box extraction steps can strongly affect the re-identification performance (4): this issue is discussed later in this section.

The two main components of a person re-identification algorithm are a *descriptor* of clothing appearance, to be computed for each template and query image, and a similarity measure between any pair of descriptors, which is used to match query and template images; the resulting 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 features, which can be extracted from different colour spaces like RGB and HSV, and texture features. To improve matching accuracy, such features can be computed from different body parts, e.g., defined simply as horizontal strips of predefined size from 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 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 computer vision field. Their main potential advantage is that they do not require the explicit definition of image descriptors or features, which are instead automatically learnt from the raw image pixels during the training process. The downside is that they require large amounts of training data, e.g., pairs of pedestrian images manually annotated as belonging to the same identity or not. 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 query individuals, 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 in a one-shot fashion both descriptor and metric learning. CNN architectures that take into account temporal information have also been defined for settings where a pedestrian tracking tool is used to obtain several images of the query and the template identities (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 most recent solutions that have been proposed in the computer vision 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 computer vision algorithms to improve their performance. The HITL approach is particularly suited to person re-identification systems, where a low-effort feedback can be obtained from the



operator; for instance, operators can indicate 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 so far. 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 the feedback is exploited to update the re-identification system.

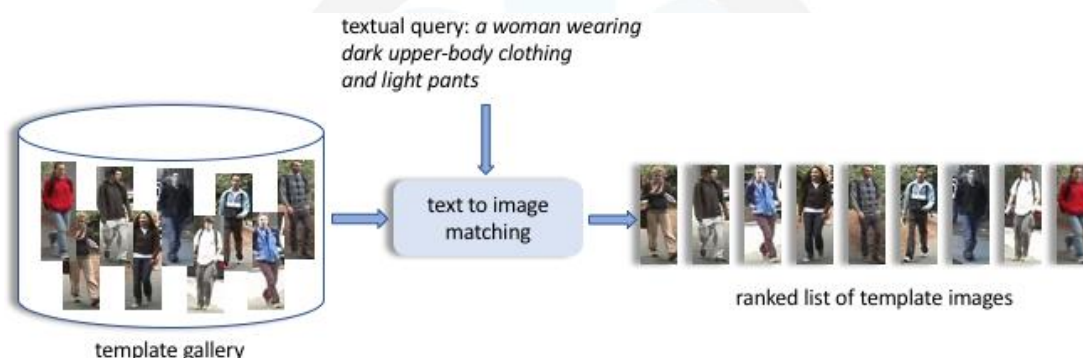
- In the method of (8) the operator is shown the sorted list of templates, and is asked to select some templates similar to the query and some dissimilar ones. The similarity measure is then updated through a metric learning technique to push the selected dissimilar templates to the bottom ranks and the similar ones to the top ranks. The template gallery is finally 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, and 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 a 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. This method is aimed at retrieving all the existing tracks of the query identity in the template gallery; in image retrieval terms, this amounts 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), and to label 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'; the template gallery is then re-ranked. The main distinguishing feature of this

method is that it is not query-specific, contrary to all the previous 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 Attribute-based 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 his/her appearance is available, for instance in terms of clothing colour and of attributes such as gender and accessories (e.g., carrying bags or backpacks). This is a common scenario in forensic investigations, e.g., when an eyewitness of a crime provides a description of the perpetrator. Similarly to appearance-based person re-identification, the attribute-based people search task focuses on unconstrained video surveillance footage where automatic face recognition is not feasible, and/or no information on the face appearance of the individual of interest is available (e.g., because the eyewitness could not see the face of that individual). For the sake of brevity, from now on the term "people search" will be used to refer to "attribute-based people search". As shown in Figure 2, the goal of a people search system is to rank the template gallery for decreasing similarity to the description of the individual of interest, such that all the templates whose appearance exactly matches the input description (if any) should appear at the top of the list. Similarly to person re-identification, the goal of a people search system is to reduce the time that LEA operators and forensic investigators have to spend to find individuals of interest, saving them from watching all the available videos.



**Figure 2 – Core functionality of an attribute-based people search system: ranking the images of a given template gallery for decreasing similarity to an input description of an individual of interest given in terms of predefined attributes related to clothing appearance and to other characteristics such as accessories (e.g., backpacks) and gender ("attribute profile"). Ideally, the templates whose appearance matches the input attribute profile (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).<sup>2</sup> 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 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

<sup>2</sup> A similar task had been previously considered in (83), and then in (85) and (84), based however only on face appearance.

trousers/skirt. An automatic classifier was then trained to detect each attribute, on a labelled data set analogous to the one of (15).

The people search task was then addressed by other authors, who proposed variants of the approaches described above (17) (18) (19) (20).

Clothing appearance attributes have recently been used also in several person re-identification methods (based on image queries), which were inspired by previous work on attribute detection in related computer vision tasks; in particular, attribute detection had been proposed for object recognition tasks, to complement low-level visual features with high-level, semantic information (21). 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 (22), (23), (24), (25), (26). The only exception is (27), 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 (SVM) 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 attribute-based people search (although this task has not been considered in the mentioned works). Finally, also in (28) 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 attribute-based people search, where no query image is available.

### 2.1.3 Crowd monitoring

The application of computer vision 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 (29), 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 computer vision 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 (30), (31), (32), (33), (34), (35), (29). In particular, in (32) a detailed analysis is given of selected methods proposed up to 2013, focused on practitioners; a wider-ranging analysis is reported in (34), (35), including a survey of more recent works and an empirical comparison of selected techniques on publicly available data sets.

A useful categorization of the existing crowd monitoring approaches can be made according to the specific task addressed (31):

- 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, which is based on detecting individuals in videos or frames/images, and is typically feasible only for low-density crowds; and the "macroscopic" (or "holistic") one, which considers a crowd as a



whole, and is typically used for high-density crowds when strong occlusions or far views do not allow reliable detection of individuals (31). 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 (33) identifies three approaches: counting by detection, by clustering and by regression.

**Counting by detection** ("object-based" approach in (30), (31)) is based on detecting the individuals in a crowd, e.g., through head contour detection, edge detection, or body models (31); it allows in principle accurate counting, and therefore also accurate density estimation. It is however feasible only for low density crowds with limited occlusions (31), (33), (29). **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 relies on people tracking as well, and thus it is feasible only for low density crowds with limited occlusion, similarly to counting by detection. **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. This approach is named "appearance-based" in (31), where it is further subdivided into pixel-based and texture-based methods. Pixel-based methods rely on local features extracted through background subtraction or edge detection, even from individual pixels. Texture-based methods use higher-level features computed from image patches (blocks of pixels). In (33) 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 (33): (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.

Methods based on counting by detection and by regression 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 it needs video sequences and is based on dynamic, motion features (34), (29). In (33) useful discussions and guidelines are given on the choice of features, including the combination of different kinds of features, and of the algorithms most suitable to specific application scenarios, e.g., in terms of the degree of illumination change in the scene; the issue of *local* density estimation (i.e., in different image regions) is also addressed.

Recent work in this field is focusing on deep learning-based approaches (36) (37), based on the use of CNN architectures, although "traditional" approaches based on manually defined features are still being investigated (38) (39).

**Tracking individuals or groups.** Pedestrian detection and tracking is a well studied problem in computer vision, and is a particular instance 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 (31). 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 (30).

A widely used approach for tracking is the one based on the Particle Filter framework, an appearance-based technique proposed in (40) for single target tracking (based only on colour information), and then extended to multiple targets and to crowded events in which people can exhibit similar clothing appearance, such as sports matches and celebrations (32). 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 the reciprocal influence of the behaviour of

nearby individuals (32). Most of the existing work focuses on tracking in low density crowds: only recently the problem of tracking individuals in dense crowds has been 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 (41), and only recently some promising results have been achieved (42). The main existing methods focus on low-density crowd scenarios, and exploit sociological models of human collective behaviour (41), (42). One open issue is that no agreed performance metrics exist yet (42).

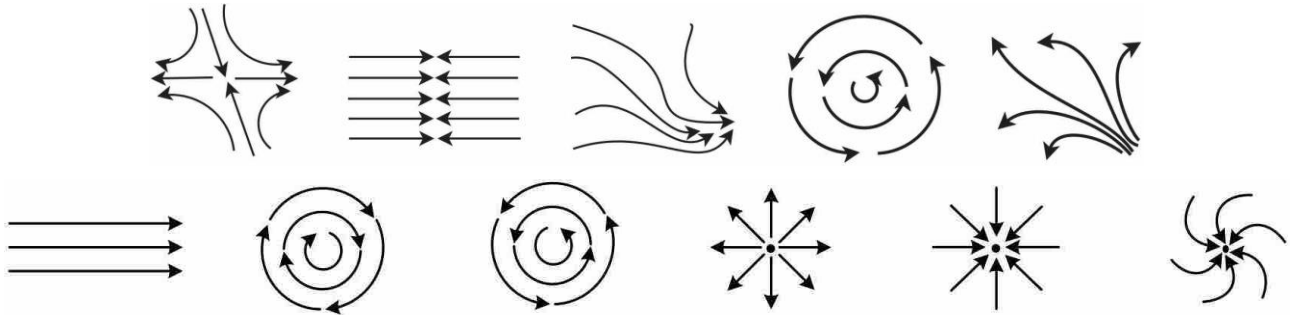
**Understanding crowd behaviour.** This task encompasses three main, distinct sub-tasks: detecting patterns of crowd movement (e.g., main directions and velocities of a crowd), detecting anomalous behaviours (e.g., unusual motions), and detecting specific events or behaviours (31). Approaches to each sub-task can be categorized also in this case into microscopic (object-based, or bottom-up) and macroscopic (holistic, or top-down) (31), (32), (34), (35). 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 (34).

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 (31). Usually such methods are based on the analysis of trajectories of individual targets (32).

Holistic methods aim instead at detecting the global patterns of the crowd flow (31). Some of the existing methods exploit models taken from sociology or psychology, like the social force model of (43). 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). One of the issues pointed out in (32) is that in unstructured crowds, which are 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, such as 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 (32). A finer categorization into three kinds of crowd motion features is given in (34): 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"), which exploit tracking information and are suitable for small- to medium-level crowd density and high resolution of single targets. In particular, some works exploit physics- and hydrodynamics-based models (32), (34).

One example of holistic (top-down) method is the one proposed in (44), 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 of (45): 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 (44) – from left to right: blocking, lane, bottleneck, ring/arch, fountainhead. Right: patterns detected by (45) – 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 (46), (32), (34), (35). A useful reference for this task is the survey of (46), 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 (46) three usual assumptions about anomalous events, underlying existing methods, are identified: (i) they are much less frequent than normal events; (ii) they exhibit significantly different characteristics from normal events; (iii) they 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, i.e., 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., to the kind of mass gathering event.

Assumptions (i) and (ii) above motivate the *anomaly detection* approach, which is widely used in different applications fields such as 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 data-driven machine learning techniques, either using only data representative of the normal behaviour, or also data representative of anomalous behaviours (if any). If the available data are not annotated as 'normal' and 'anomalous', an *unsupervised* learning approach can also be used: data are clustered assuming they are mostly representative of normal behaviour, and then outliers are considered as anomalies. In (34) 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 computer vision techniques (in particular, on visual features), and methods inspired by physics, which employ physical models to represent crowd dynamics.

An example of holistic, or top-down method for abnormal behaviour detection, which is defined as any deviation from a normal model, is the one of (47); in addition to optical flow it uses the social force model of pedestrian dynamics by (43), which has 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 to the need of accurate tracking results. In particular, tasks like abandoned object detection require a complex analysis of the interactions between different targets. For instance, the method of (48) was developed in the context

of the project SUBITO<sup>3</sup> (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 (48) were considered by the authors not yet acceptable for a deployed threat assessment system.

**Interactions between computer vision-based crowd monitoring and crowd modelling.** One of the main issues common to computer vision techniques for several crowd monitoring tasks is the collection of representative and labelled data for training machine learning techniques, and for validating them (31). In particular, the lack of representative data is critical for anomalous behaviour detection, as discussed above. An interesting solution to be investigated in the context of LETSCROWD is the exploitation of **simulated crowd videos** data generated by computer graphics using **crowd modelling techniques**, which has been proposed by several authors (30), (31), (34), (49), (50). One of the key advantages of this solution is that it allows to simulate *controlled* scenarios, avoiding the tiresome, time-consuming and error-prone task of manual labelling images or video sequences, although the simulation of unpredictable behaviours remains an open issue. On the other hand, information obtained from computer vision algorithms during operation, even in real time, can 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 (31).

## 2.2 EXISTING STATE OF THE ART

Overall, the current technology readiness level (TRL) of existing methods for the considered computer vision 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 and pedestrian tracking; they are however focused on sparse crowds, which is not the scenario of interest for the HCV tool. The state of the art of the three computer vision tasks 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

As pointed out in the literature review (Sect. 2.1), existing work on the person re-identification and people search tasks focused on the definition of descriptors and similarity measures, disregarding the issue of how the template gallery is populated. 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 graphical user interface (GUI) (51), in real-world, operational settings the bounding boxes of all the template images have to be *automatically* extracted by a pedestrian detector or tracking tool from video sequences or frames. 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 (52), (51).

Pedestrian detection and tracking have been addressed separately in the computer vision literature, as particular cases of the more general object detection and tracking problems, and a number of approaches and algorithms have been proposed so far (see, e.g., (53)). 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<sup>4</sup> (54)), as well as some commercial solutions. However, also for these tasks existing algorithms and solutions achieve satisfactory results only under relatively simple application scenarios, such as sparse crowds with limited occlusion. For instance, a very recent work published in a top journal for the computer vision and pattern recognition field (55) provided empirical evidence that state-of-

<sup>3</sup> [https://cordis.europa.eu/project/rcn/89391\\_en.html](https://cordis.europa.eu/project/rcn/89391_en.html)

<sup>4</sup> <https://github.com/rbgirshick/py-faster-rcnn>



the-art pedestrian detectors 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 video surveillance scenarios, and in particular of mass gathering events. Currently, a few 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 ranking the template images of the same identity as the query image (if any) near the top of the list of templates. This means that re-identification performance reported in the literature on relatively small data sets is 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 (56), where the authors describe a tag-and-track surveillance system developed for the medium-sized Cleveland Hopkins International Airport in Ohio, USA. The system was designed to assist the transportation security administration officers in monitoring the airport using the existing surveillance camera network. It consisted of three cameras, one located just after a security checkpoint in which a subject of interest can be 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 were oriented at sharp angles to the floor of about 45 degrees, as in many airports; moreover, they were analogue cameras, whose feeds were converted into the H.264 standard with a 704×408 resolution at about 30 frames per second (fps). Potentially useful suggestions are reported in (56) 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 related to pedestrian tracking, which is used in the object-based or microscopic approach for several tasks such as people counting by clustering, and anomaly detection based on trajectory analysis. Pedestrian tracking is still a challenging computer vision problem, 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 (32).

Another issue pointed out by (34) is that most of the existing methods which require tracking, learning and behaviour detection or recognition, carry out these steps by simple integration of the corresponding functional modules, disregarding their interactions for the sake of simplicity. 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 (35), (34).

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 (46), (34), (35),

(29). Many authors used different, and sometimes non-publicly available datasets for evaluation, which makes the reported results not comparable. Under a practical viewpoint, this 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 (46).

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. Interesting discussions on this issue can be found in (57) for the computer vision field, and in (58) for the machine learning field. As a consequence, there is no guarantee that the same performances reported in the literature for specific computer vision tasks can be achieved also in real-world settings. 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, this subsection reports examples of the performance reported in the literature for each computer vision task considered above, with the caveat that these examples 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: 276 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: it is larger than VIPeR (it contains about 40,000 images of 1,382 different individuals), and all its bounding boxes are 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 to more than 0.40** across the considered algorithms. For other data sets the performance range 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). Quoting from (3): *"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."*

A further issue that needs to be pointed out is that the evaluation of person re-identification methods based on data-driven, machine learning techniques, including recent deep learning techniques, is typically carried out using both training and testing data from the same data set. This setting does not reflect real-world deployment scenarios, where the environmental conditions (e.g., lighting) and the characteristics and set-up of the cameras (e.g., image resolution and quality, and camera tilt angle) can be significantly different from the ones of training data. As a consequence, reported performances are likely to be optimistically biased. This issue could be addressed using HITL techniques, which is one of the solutions considered in the HCV tool.

**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, the attribute-based re-identification method of (22) has been evaluated on two data sets used: VIPeR (mentioned above), and PRID; the detection accuracy reported on VIPeR for 21 attributes 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 (59), 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** and **density estimation** we consider the literature survey of (33), where experiments on three data sets (UCSD Pedestrian Traffic, QMUL Mall and PETS 2009, see Sect. 2.4) are reported to compare six different algorithms based on machine learning techniques. 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:

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

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 evaluate the generalisation capability to unseen density (a typical scenario in real applications), other experiments were carried out by training on sparse scenes and testing on dense ones 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. However, the same issue pointed out above for person re-identification arises also for crowd density estimation: all the results mentioned above were obtained using training and testing data coming from the same data set, which is likely to produce optimistically biased performance estimates. Cross-data set evaluations are lacking, despite they are more representative of real-world deployment scenarios in which data (videos) processed during system operation can be very different from the ones used as training set during design. This issue was confirmed by experiments carried out during the development of the HCV tool, both on publicly available data sets and on videos collected during the practical demonstrations of the LETSCROWD project: the results have clearly shown that the accuracy of the considered crowd density estimation techniques drastically drops when testing data exhibit different characteristics than training data in terms of factors such as background, perspective and crowd size. For crowd density estimation, using crowd simulations obtained through crowd modelling and computer graphics tools (mentioned in Sect. 2.1.3) is a potentially interesting solution to investigate, also in the context of the LETSCROWD project.

For the **detection of patterns of movement** we consider the methods of (44), (45) 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 (60), focused on tracking individuals in dense crowds. The evaluation was carried out on eight real video sequences showing commuters walking outdoor (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 (60) attained an accuracy of 0.67 to 1.0 on the different sequences.

For **group detection** we consider the method of (42). 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 (47) experiments on the on UMN Crowd and UCF Web data sets (see Sect. 2.4) were carried out; 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 (61), 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 (47): the reported TP rate was 0.6. In (62) 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 (63) 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 (48), 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 (48) 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.<sup>5</sup> Most of this software is written in Matlab and in Python, except for pedestrian detection where C++ is also often used.<sup>6</sup> A few such implementations are in the form of software libraries.

<sup>5</sup> <https://github.com>

<sup>6</sup> This is probably due to efficiency reasons, and to the fact that pedestrian detection is an older research topic than re-identification.

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 such as pedestrian detection and re-identification modules. The only useful resources to the HCV tool turned out to be related to pedestrian detection and tracking tasks, e.g., the Faster R-CNN object detection algorithm of (54) mentioned above. No software implementations were found for crowd monitoring tasks, instead.

### 2.2.5 Commercial solutions

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

- Vision Semantics Ltd<sup>7</sup> 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.
- iOmniscient<sup>8</sup> is an Australian video analytics company founded in 2001. It offers software solutions for video surveillance 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 such as 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 video demos are available for abandoned object detection and for people counting, in a marathon scenario and in a train station.
- Ipsotek<sup>9</sup> 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 of up to about twenty people.
- CrowdVision<sup>10</sup> is another U.K. company, founded in 2007, offering solutions targeted to the 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.

<sup>7</sup> <http://www.visionsemantics.com/>

<sup>8</sup> <http://iomniscient.com>

<sup>9</sup> <https://www.ipsotek.com>

<sup>10</sup> <https://www.crowdvision.com>

- BriefCam<sup>11</sup> is a company founded in 2008 by a computer vision researcher of the Hebrew University of Jerusalem. It offers a video content analytics platform including functionalities analogous to people search (based on clothing colours and gender) and to person re-identification. Only a video demo of these functionalities are available.
- Avigilon<sup>12</sup> is a Motorola Solutions company founded in 2006, which designs, develops, and manufactures video analytics, network video management software and hardware, surveillance cameras, and access control solutions. Its video analytics solution includes an "appearance search" functionality analogous to person re-identification and people search. Also in this case only a video demo of this functionality is available.

It is worth noting that the case studies reported in the Web sites of the above companies about functionality analogous to person re-identification or people search do not refer to people tracking in crowd.

Several other companies provide video surveillance solutions, including functionality such as face recognition, object tracking, object retrieval, etc., which are however not targeted to crowded scenes.

## 2.3 DEVELOPMENT OVER THE STATE OF THE ART

### 2.3.1 How the HCV tool advances the current state of the art

The advancement achieved by the HCV tool over the state of the art in the considered computer vision tasks has not been pursued in terms of performance (accuracy) improvement over results reported in the literature. The main reasons are that most of the existing work aimed 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, as discussed in Sect. 2.2, publicly available benchmark data sets are not representative of real-word application scenarios, which makes the reported accuracy to be optimistically biased.

The advancement over the state of the art has been pursued in the following aspects, instead:

- The HCV tool works end-to-end: it takes as input raw videos and outputs easy to interpret, actionable information to LEA operators. In the current prototype implementation inputs consist of pre-recorded videos. This reflects post-event, forensic use cases for the person re-identification and people search tasks. Nevertheless, the crowd density estimation component works in real time, and is thus capable of processing streaming videos (see below). The outputs are shown through a suitable GUI, which is described in Sect. 4. In particular, developing end-to-end person re-identification and people search components required the integration of different processing steps separately considered in the literature, such as pedestrian detection and bounding box extraction to populate the template gallery.
- The computer vision techniques for the crowd monitoring module have been chosen taking also into account that processing time has to be suitable to real-time operation, which is the main (or even unique) application scenario of interest for this kind of functionality. Achieving a low processing time of the person re-identification and people search modules, suitable to their usage during event execution, was instead out of the scope of the HCV tool: in this case processing time strongly depends on the template gallery size, which in the last resort depends on the number and duration of videos to be processed, and on the number of pedestrians that appear in such videos; accordingly, in this case a low processing time can be achieved in a real operational scenario by suitably scaling the hardware facility.
- The HITL technique has been implemented in the person re-identification and people search components to adapt them to the characteristics of a specific application scenario by exploiting the operator's feedback.

---

<sup>11</sup> <https://www.briefcam.com>

<sup>12</sup> <http://avigilon.com>

As pointed out above, this feature has been considered so far only by a few works in the literature, and only for the person re-identification task.

- A distinctive feature of the HCV tool is its integration with the crowd modelling and planning (CMP) tool developed in task T5.1 (deliverable D5.1), and with the dynamic risk assessment (DRA) tool developed in work package WP3 (deliverables D3.2 and D3.4). In particular, crowd density estimated in real time by the HCV tool can be used as input by the CMP and DRA tools, and videos of crowd simulations produced by the CMP tool can be used off-line for training the crowd density estimation algorithm, to adapt it to a specific operational environment (e.g., specific camera views and crowd size and configuration). This integration is described in Sect. 4.2.

### 2.3.2 What has been achieved

The HCV tool consists of three integrated modules which provide functionalities aimed at supporting LEA operations both during event execution and in post-event forensic investigations:

- **person re-identification**, to search for an individual of interest based on clothing appearance, using a *query image* of that individual, on videos acquired by the cameras of a video surveillance network;
- **people search**, to search for an individual of interest starting from a *textual description* of clothing appearance and other attributes, on videos acquired by the cameras of a video surveillance network;
- **crowd density estimation**, which provides in real time a frame-by-frame estimate of the number of people in a video, and can raise alerts when anomalous behaviours are detected, in terms of sudden increase or decrease of crowd density, or crowd density exceeding a user-defined threshold.

As mentioned in Sect. 2.3.1, the crowd monitoring module can be integrated with the CMP tool developed in task T5.1, and with the DRA tool developed in work package WP3, through the LETSCROWD server.

With respect to the planned functionalities of the crowd monitoring module, the accuracy of state-of-the-art techniques for detecting patterns of crowd movement and related anomalous behaviours was found to be not satisfactory for real application scenarios of interest to the LETSCROWD project. For this reason, only crowd density estimation and the related anomaly detection functionality have been implemented.

At the time of submitting this deliverable, the HCV tools has been tested both on publicly available data sets, as well as in three practical demonstrations of the LETSCROWD project involving three different partner LEAs and including two real mass gathering events (see the 'Validation' subsections of Sect. 5); further practical demonstrations are planned during the last six months of the project.

## 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 computer vision tasks of person re-identification, people search and crowd monitoring. Due to the difficulty in collecting recorded videos from the video surveillance systems of the LEA partners of LETSCROWD, some of these data sets have been selected for developing and validating the HCV tool.

**Person re-identification.** More than 20 data sets were available during the development of the HCV tool, released since 2007. An updated and detailed list is maintained at a Web site by the Robust Systems Lab at the Northeastern University (Boston, USA).<sup>13</sup> 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;

<sup>13</sup>

<http://robustsystems.coe.neu.edu/sites/robustsystems.coe.neu.edu/files/systems/projectpages/reiddataset.html>



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 128×64 for some data sets;
- beside pedestrian bounding box images, in some cases full video frames (for about half of the data sets), tracking sequences of individuals (for a few data sets), and video sequences (for about ten data sets) are indicated as available in the Web site mentioned above, but in most cases they have not been found in the data set Web pages.

An example taken from the well-known VIPeR<sup>14</sup> data set is shown in Figure 4.



**Figure 4 – Example of images from the VIPeR person re-identification data set: pairs of images (128×48 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 (using different pedestrian detectors) are provided, including 'false positives', i.e., bounding boxes which do not contain a pedestrian, to mimic real scenarios.
- 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.
- 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, the five ones listed in Table 1 turned out to be suitable to the development and evaluation of the HCV person re-identification module, in terms of the number of different identities they contain, and of the availability of full video frames or video sequences.

<sup>14</sup> <https://vision.soe.ucsc.edu/node/178>

Data set	Description
PRW <sup>15</sup> (Person Re-identification in the Wild) (51)	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) (64)	Acquired by 6 cameras in a University campus (same as PRW). Identities: 1,261; bounding boxes: automatically extracted.
DukeMTMC-4ReID <sup>16</sup> (65)	Acquired by 8 cameras on a University campus. Identities: 1,413; videos are available; bounding boxes: automatically extracted.
Airport <sup>17</sup> (56)	Acquired by 6 cameras installed past 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.
Market-1501 <sup>18</sup> (66)	Acquired in front of a supermarket at Tsinghua University from five high-resolution cameras and one low-resolution camera with overlapping views. Identities: 1,501; bounding boxes: mix of automatic and manual extraction; a set of false detections ("distractors") are also included to mimic real application scenarios.

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

**People search.** Four data sets are available for *attribute-based person re-identification* with *image* queries. They are made up of pedestrian images (bounding boxes) acquired in video surveillance 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 <sup>19</sup> (22)	1,264 outdoor images acquired by 2 cameras; image resolution 128×48; 21 binary attributes.
APiS <sup>20</sup> (Attributed Pedestrians in Surveillance) (67)	3,661 outdoor images; image resolution 128×48; 11 binary attributes.
PETA <sup>21</sup> (PEdesTrian Attribute) (59)	19,000 indoor and outdoor images collected from ten re-identification data sets, with varying camera angle, view point, illumination, and resolution (from 39×17 to 365×169); 61 binary attributes plus 4 attributes with 11 values each.

<sup>15</sup> [http://www.liangzheng.org/Project/project\\_prw.html](http://www.liangzheng.org/Project/project_prw.html)

<sup>16</sup> <http://vision.cs.duke.edu/DukeMTMC/>

<sup>17</sup> <http://www.northeastern.edu/alert/transitioning-technology/alert-datasets/alert-airport-re-identification-dataset/>

<sup>18</sup> <https://jingdongwang2017.github.io/Projects/ReID/Datasets/Market-1501.html>

<sup>19</sup> <http://www.eecs.qmul.ac.uk/~rlayne/>

<sup>20</sup> <http://www.cbsr.ia.ac.cn/english/APIs-1.0-Database.html>

<sup>21</sup> <http://mmlab.ie.cuhk.edu.hk/projects/PETA.html>

Data set	Description
RAP <sup>22</sup> (Richly Annotated Pedestrian) (68)	41,585 indoor images from 26 cameras; image resolution from 92×36 to 554×344, 69 binary attributes.

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

These data sets can be directly exploited also for training attribute detectors in the HCV people search tool, where the query consists of a description of the appearance of an individual of interest in terms of an **attribute profile**, i.e., a suitable combination of predefined attributes related to clothing appearance, gender, etc. 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 one to the people search tool is PETA, as it is the largest one that contains both 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) that had been chosen in (22) on the basis of **operational procedures of human experts**:<sup>23</sup> *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., tasks for which manual annotations are provided). Most of these data sets have been surveyed in (34), (35), (29). Notably, Agoraset and CrowdFlow are **synthetic** data sets developed according to the approach discussed in Sect. 2.1.3 to overcome the scarcity and drawbacks of real videos. Agoraset exhibits in turn several limitations, as pointed out by its 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. CrowdFlow contains instead realistic backgrounds and more realistic flows of people, although the number of individuals in this data set, ranging from 371 to 1,451, is relatively low for mass gathering events of interest to the LETSCROWD project.

Data set	Description
UCSD Pedestrian Traffic <sup>24</sup> (69)	Tasks: <b>crowd counting / density estimation</b> , and <b>pedestrian tracking</b> . 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, 238×158 pixels, 10 fps; region-of-interest (ROI) and perspective map are included.
QMUL Mall <sup>25</sup> (33)	Tasks: <b>crowd counting / density estimation</b> . 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 640×480 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.

<sup>22</sup> <http://rap.idealtest.org/>

<sup>23</sup> The following source was cited in (22), 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).

<sup>24</sup> <http://www.svcl.ucsd.edu/projects/peoplecnt/>

<sup>25</sup> <http://personal.ie.cuhk.edu.hk/~ccloy/>



Data set	Description
UCF Crowd Counting <sup>26</sup> (70)	Tasks: <b>crowd counting / density estimation</b> . Made up of 50 publicly available web images (mainly from Flickr) collected by researchers of the University of 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 <sup>27</sup> (71)	Tasks: <b>crowd counting / density estimation</b> . 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 <sup>28</sup> (72)	Tasks: <b>detection of patterns of crowd movement</b> , and of <b>anomalies</b> . 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 <sup>29</sup> (47)	Tasks: <b>anomalous behaviour detection in crowds</b> . 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 <sup>30</sup>	Tasks: <b>anomalous behaviour detection in crowds</b> . 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 <sup>31</sup>	Tasks: <b>anomalous behaviour detection in crowds</b> . 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 <sup>32</sup> (73)	Tasks: <b>anomalous behaviour detection in crowds</b> (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.

<sup>26</sup> [http://crcv.ucf.edu/data/crowd\\_counting.php](http://crcv.ucf.edu/data/crowd_counting.php)

<sup>27</sup> <http://www.ee.cuhk.edu.hk/~xgwang/expo.html>

<sup>28</sup> <http://crcv.ucf.edu/data/crowd.php>

<sup>29</sup> [http://crcv.ucf.edu/projects/Abnormal\\_Crowd/](http://crcv.ucf.edu/projects/Abnormal_Crowd/)

<sup>30</sup> [http://mha.cs.umn.edu/proj\\_events.shtml#crowd](http://mha.cs.umn.edu/proj_events.shtml#crowd)

<sup>31</sup> <http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm>

<sup>32</sup> <http://www.openu.ac.il/home/hassner/data/violentflows/index.html>

Data set	Description
UCF Tracking in High Density Crowds <sup>33</sup> (74)	Tasks: <b>tracking individuals in a crowd</b> . Four videos collected by UCF researchers: three marathon sequences, and a busy train station sequence; trajectories of hundreds of individuals are manually annotated.
Train Station <sup>34</sup> (75)	Tasks: <b>tracking individuals in a crowd</b> . 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 480×720 pixels. Tens of thousand trajectories of individuals are manually annotated.
PETS 2009 <sup>35</sup>	Tasks: <b>crowd counting / density estimation, tracking of individual(s) within a crowd, detection of crowd flow</b> and of <b>crowd events</b> . 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 <b>crowd events</b> are present: walking, running, evacuation (rapid dispersion), local dispersion, crowd formation and splitting.
CUHK Crowd <sup>36</sup> (76)	Tasks: <b>group detection in crowds</b> . Made up of 474 video clips from 215 crowded scenes collected from <a href="#">Getty Images</a> and <a href="#">Pond5</a> by CUHK researchers.
Agoraset <sup>37</sup> (77)	A synthetic crowd video data set that can be used for evaluation of different low-level video crowd analysis methods, like tracking and segmentation. See text for further details.
CrowdFlow <sup>38</sup> (78)	A synthetic data set which simulates five kinds of crowd flows: few pedestrians walking against a main crowd flow; bottleneck dividing one major flow into three; two dense flows walking close past each other; spread of collective panic and subsequent escape; a marathon sequence. This data set is made up of 10 sequences with a top-view of the crowd, with length between 300 and 450 frames, rendered with a frame rate of 25 Hz and a 1820×720 resolution, showing 371 to 1,451 individuals.

**Table 3 – Main publicly available data sets related to crowd monitoring tasks.**

### 3 REQUIREMENTS

During the activities of work package WP2 a use case for the HCV tool had first been defined, starting from the goal and functionality proposed in the DoW. The use case was focused on the person re-identification and people search functionalities. It is reported in detail in deliverable D2.2, and is summarized in Sect. 3.1. User (LEA) and system requirements (see deliverable D2.1) have then been defined: they are reported in Sect. 3.2 and 3.3, respectively. Finally, Sect. **¡Error! No se encuentra el origen de la referencia.** addresses specific privacy issues related to the HCV tool.

#### 3.1 USE CASE

A use case was developed at the beginning of the project in work package WP2 to illustrate the person re-identification and people search functionalities. It is reported in deliverable D2.2 as UC-005. This use case

<sup>33</sup> <http://crcv.ucf.edu/data/tracking.php>

<sup>34</sup> <https://www.ee.cuhk.edu.hk/~xgwang/grandcentral.html>

<sup>35</sup> <http://www.cvg.reading.ac.uk/PETS2009/a.html>

<sup>36</sup> [http://www.ee.cuhk.edu.hk/~jshao/projects/CUHKcrowd\\_files/cuhk\\_crowd\\_dataset.htm](http://www.ee.cuhk.edu.hk/~jshao/projects/CUHKcrowd_files/cuhk_crowd_dataset.htm)

<sup>37</sup> <https://www.sites.univ-rennes2.fr/costel/corpetti/agoraset/Site/AGORASET.html>

<sup>38</sup> <https://github.com/tsenst/CrowdFlow>

focuses on two usage scenarios of video surveillance data: supporting LEAs reaction to some incident during event execution, and supporting forensic investigations in the post-event phase. Two main facts are pointed out by this use case: the aim of computer vision tools like the HCV tool is to support human operators, not to work autonomously in place of 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** were requested to implement the use case. The first is information from LEAs about the use of video surveillance systems during mass gathering events, which was collected through questionnaires (see Sect. 3.2), to guarantee the compliance with existing regulations on the use of video surveillance systems and the potential usefulness of the proposed tool for LEA operations. The second input is a set of videos acquired by video surveillance systems, satisfying system requirements (see Sect. 3.3); such videos are necessary both to develop the tools using data-driven computer vision techniques, and for their validation by LEAs, as well as for the practical demonstrations. Ideally, real videos recorded by LEAs' video surveillance systems during past relevant mass gathering events should be used. However, taking into account the difficulty of retrieving and using such data in the context of the LETSCROWD project (as pointed out by the LEA questionnaires), alternative solutions can be considered, like videos acquired by simulations of mass gathering events, and publicly available video data sets collected by the computer vision research community (see Sect. 2.4).

Three **pre-conditions** were planned: (i) The mass gathering location is monitored by a video surveillance system, including a system capable of recording and storing 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 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 events can occur:

- LEA operators 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 exhibiting a similar clothing appearance 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 than watching all the available 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 eyewitness describes 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 appearance is available. In both cases, LEA operators would like to immediately (i.e., 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 target appearance description using a predefined set of attributes (e.g., the specific colour and/or texture of the upper-body and lower-body garment, gender, the kind of sleeves – short or long), then the tool automatically searches for images of individuals exhibiting a similar 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 can allow LEA operators to find the images of the suspect individuals (if any) in a much shorter time than watching all the available 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 be considered:

- 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 use 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 an eyewitness): in this case LEA operators use 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 start 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 eyewitnesses. 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.

### 3.2 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 video surveillance systems by LEAs of different countries for monitoring mass gathering events, 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 the potential usefulness of the HCV tool to support LEAs during event execution and in post-event forensics investigations. The questions were subdivided according to the two considered event phases; they are reported below.



- Event execution phase: Do you use video surveillance systems to monitor crowded events? If so:
  - What are the laws (privacy, etc.) that rule the use of video surveillance systems by LEAs in your country (possibly depending on the kind of event, like a sport event, a demonstration, etc.)?
  - What kinds of video surveillance systems do you use/can be used in your country? (e.g., fixed cameras, cameras worn by officers, airborne cameras, etc.)
  - 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 operators 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 an eyewitness of some event of interest.)
- Post-event phase. If you use video surveillance systems to monitor crowded events:
  - Are recorded videos used (or can they be used) for forensic analysis in the post-event phase? (e.g., the kind of analysis involved in task T5.4: searching for a specific person observed in some video frames, or described by an eyewitness of 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 were obtained: Hochschule für den öffentlichen Dienst in Bayern – Fachbereich Polizei (**BayHfoeD**, 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:

- Video surveillance systems are used by all the six LEAs.
- Video surveillance systems with fixed cameras are used by all LEAs. Depending on the country regulations (e.g., about the type of event and the crowd size), other kinds of video surveillance systems are used by some LEAs: temporary video surveillance systems (LPV); mobile cameras worn by LEA officers (BayHfoeD, MAI, ERT, PSP), by special video teams (LPV), or by forensic operators in the 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 videos of CCTV systems 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 set by country-specific regulations on the purpose and on the storing time. Storing videos can be allowed to collect evidence of incidents (LPV), or if there is a concrete threat for the public security or an indication for a criminal behaviour (BayHfoeD). 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 BayHfoeD is defining the requirements for improving the practical use of video surveillance 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 provide stored videos for the design and demonstration of the HCV tool, also taking into account the limited storage time.

From the questionnaire outcomes summarized above **two main conclusions** were drawn: (1) the HCV tool as envisaged in the LETSCROWD project proposal would be compliant with existing regulations on the use of

video surveillance systems by LEAs of the six involved countries; (2) the three functionalities of the HCV tool (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 manually by LEA operators and investigators.

On the other hand, the difficulty for LEAs to make videos of real mass gathering events available to LETSCROWD partners, due to strict regulations on this matter, was pointed out as a significant limitation for the development and validation of the HCV tool, and in particular for the crowd monitoring module. For this reason the development of the HCV tool during the first year of the project has relied on publicly available image and video data sets collected by the computer vision research community (see Sect. 2.4). During the second year of the project some videos have been collected during three practical demonstrations where the HCV tool was tested (see deliverable D6.3), including two real mass gathering events. These videos have been used to validate and further develop the tool. Currently, synthetic videos generated by the CMP tool of task T5.1 by the Crowd Dynamics (CROWD) LETSCROWD partner are being used for further validation and development (see Sect. 5.3.2).

### 3.3 SYSTEM REQUIREMENTS

Fourteen system requirements had been defined for the HCV tool at the beginning of the project. They are reported in deliverable D2.1, and are listed below, sorted by requirement type. Requirement IDs used in deliverable D2.1 are 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 video surveillance systems by the LEAs, including related LEAs internal procedures. The rationale is that the use of video surveillance data affects privacy, and related regulations vary from country to country; moreover, compliance with LEAs internal procedures on the use of video surveillance 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; in particular, 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 computer vision techniques to be used in this tool are data-driven, and thus data as much representative as possible of the 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 video surveillance 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 computer vision 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 computer vision 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 video surveillance 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.

**Privacy-related issues**, which are the focus of the legal requirement HCV\_002 are of particular concern for the HCV tool. Currently, to manage their video surveillance systems LEAs use commercial or ad hoc software suites that are installed on LEAs computer facilities<sup>39</sup> and allow them both to view the videos in real time and to store them for further analysis, under country-specific regulations such as the ones mentioned in Sect. 3.2; for instance, videos can be stored only for a limited time (usually one month) unless they are useful for investigations on a crime. Accordingly, in a real deployment scenario the functionalities of the HCV tool are likely to be provided as part of the functionalities of a more complex software suite owned and managed by a LEA.<sup>40</sup> In this context, the following considerations can be made about the functionalities of the HCV tool.

- The **crowd density estimation** functionality does not perform any high-level image analysis, and provides only an estimate of the number of people in each frame of a streaming video, together with alerts about collective (i.e., not related to specific individuals) anomalous events such as overcrowding; this information is intended to be shown to LEA operators who are watching in real time the *same* videos, typically on CCTV screens in a control room. No specific privacy-related issues are foreseen for this functionality, as confirmed by the LETSCROWD partner LEAs involved in the practical demonstrations where the HCV tool has been tested.

<sup>39</sup> This emerged from discussions, workshops and practical demonstrations with the LETSCROWD LEA partners, and in particular ERT, INTERNO and LPV.

<sup>40</sup> In fact, there is already evidence that companies offering video analytics solutions for video surveillance systems targeted also on LEAs are integrating in their products functionalities analogous to person re-identification and people search (see Sect. 2.2.5).



- The **person re-identification** functionality is capable to detect a given individual in one or more recorded videos. To this aim an image (query) of the individual of interest has to be provided by the user, typically extracted from one of the available videos, or possibly coming from a different source (e.g., a picture taken during event execution by a LEA officer in the field, with a mobile device); moreover, all the pedestrian images automatically detected by a software tool which is part of the person re-identification system or of the software suite which includes it have to be temporarily stored, to be matched with the query image and shown to the user, together with contextual information such as the detection timestamp. No other information – in particular, no information about the identity of the detected individuals – is collected and stored. Images of detected individuals can be stored in the same computer facility hosting the person re-identification software, owned by the LEA, and need to be kept only until the user's search task is concluded. If necessary for further investigations, after their removal the same images can be obtained again by running the person re-identification system with the same query image as input. The above usage scenario refers to post-event forensic investigations on stored videos, about which no specific privacy-related issues emerged during the project or are foreseen for this functionality. On the other hand, a real time usage scenario can also be envisaged, to detect and to signal to LEA operators the presence of an individual of interest in streaming videos during event execution. In this scenario only the images of automatically detected individuals who match the query image need to be temporarily stored to be shown to the operator, and the ones who are not deemed to be relevant can be immediately deleted. However, as emerged during the practical demonstrations of this project, regulations in some countries may restrict the use of the person re-identification functionality only to a forensic investigation scenario for post-event analyses related to crime prosecution or investigation.
- Similar considerations as for person re-identification apply to the **people search** functionality. The only differences are that:
  - no image of the individual of interest (query) is available, and only an attribute profile related to clothing appearance and other attributes (including gender) needs to be provided by the user;
  - an estimated attribute profile is automatically computed and stored for each automatically detected individual in the available videos, together with their image.

## 4 DESIGN

This section describes the architecture of the HCV tool, the links to the tools developed in other tasks and work packages, the computer vision algorithms used to implement it, its user interface, and the implementation details, including a summary of the results of validation activities carried out on publicly available data sets and during practical demonstrations of the LETSCROWD project.

The HCV tool has been conceived as a prototype aimed at demonstrating two kinds of computer vision (or video analytics) functionalities for supporting LEA operators and forensic investigators in the use of video surveillance systems for monitoring mass gathering events and for post-event investigations: (i) searching for individuals of interest; (ii) estimating the size of a crowd and detecting related anomalous behaviours. As pointed out in previous sections, in real deployment scenarios these functionalities can be part of software suites like the ones that are already used by LEAs to manage their video surveillance systems. In the context of the LETSCROWD project it was not possible to integrate the HCV tool in the software suites of the partner LEAs, both because this would have required to get access to their computer facilities and to sensitive data, and because different software suites are used by the different partner LEAs. For these reasons the HCV tool has been designed and implemented as a stand-alone software prototype with a common GUI which gives access to the three modules of person re-identification, people search and crowd density estimation.

## 4.1 ARCHITECTURE

As planned in the system requirements (see Sect. 3.3) the HCV tool has been designed using a **client-server architecture** and a **Web-based GUI**. The GUI operates at the client side through a Web browser, whereas all the other software modules operate at the server side. This architecture was chosen at the beginning of the project to allow the partner LEAs to use the HCV tool during project activities (including practical demonstrations) without installing any software on their own computer facilities.

From a logical viewpoint the HCV tool is made up of three modules that share a common GUI, and provide the person re-identification, people search and crowd density estimation functionalities. Since the two former modules provide a similar functionality (searching for individuals of interest) they share some back-end software components such as the pedestrian tracking tool and the data base which stores the template gallery. The architecture of each of the the three modules is described in the following subsections.

### 4.1.1 Person re-identification module

This module provides the person re-identification functionality summarised in Figure 1 in a **end-to-end system**, which corresponds to a forensic investigation scenario in the post-event phase. The user can load one or more stored videos. The GUI shows only the last loaded video. The video currently playing in the GUI can be stopped at any time to select through a point-and-click interface a rectangular region (bounding box) containing an individual of interest (query image), and to run the retrieval process. The system then retrieves all the pedestrian bounding boxes automatically extracted from all the uploaded videos, and shows the user their images sorted by decreasing similarity to the query, together with contextual information such as the timestamp and the name of the video from which each image was extracted (in a real application scenario the camera ID could be shown, instead). The user can access each retrieved image and can play the corresponding video, starting from the frame where that image was extracted: this allows, e.g., to analyse the behaviour of the individual shown in the retrieved image. If no images of the query individual are retrieved, or if the user would like to search for additional images, the retrieval process can be refined by providing a *feedback* on any subset of the retrieved images, indicating whether the individual shown in each of such images has a similar or dissimilar clothing appearance to the query individual: an updated list of retrieved images is then shown to the user. This step can be repeated several times.

The above functionality is provided by the following software modules, depicted in Figure 5.

- The **pedestrian tracking** module processes in background all the loaded videos (including the ones that are not shown in the GUI) and outputs a *track* for each detected identity, i.e., a sequence of bounding boxes (one for each frame) assumed to contain the same individual tracked over a sequence of frames, each labelled with a same, univocal ID. The larger bounding box of each track is then selected and stored in the template gallery, together with the corresponding descriptor (see below). The choice of a *single* bounding box per track allows to reduce the processing time for the subsequent steps of feature extraction and matching, and avoids multiple and almost identical images of a same individual to appear in the retrieval results; the choice of the *larger* bounding box is motivated by the fact that it should correspond to the time instant when the tracked individual is *closest* to the camera, and therefore its image is likely to contain the highest amount of details. This module replaced the pedestrian *detector* used in the first version of the HCV tool, which extracted bounding boxes from each video frame independently on the other frames.
- The **feature extractor** module takes as input all pedestrian images (bounding boxes) coming from the pedestrian tracking module, computes a *descriptor* of each of them and stores it into the template gallery.
- The **template gallery** is a database that stores all the images produced by the pedestrian tracking module together with their descriptors and contextual information, including the source video and the timestamp. The template gallery is incrementally updated as soon as new pedestrian images are produced by the pedestrian tracking module.

- The **matching** module takes as input the descriptor of a query image chosen by the user, matches it with the descriptors of all the template gallery images, and returns a ranked list of template images sorted by decreasing similarity to the query, which is computed according to a predefined similarity measure.
- The **HITL** module it receives feedback information from the user on one or more template images in the ranked list provided by the matching module; the feedback consists in indicating whether the individual shown in each of the chosen images has a similar or dissimilar clothing appearance to the query individual. All the template images and then re-ranked such as the ones closer in feature space to the images labelled as 'similar' are pushed toward the top of the list, whereas the ones labelled as 'dissimilar' are pushed toward the end of the list.
- The **GUI** allows the user to upload one or more videos, to play any one of them (one at a time), to stop the playing video, to select from a still frame a bounding box containing an individual of interest as the query image, to start the retrieval process, to scroll the ranked list of retrieved template images and related contextual information, to inspect any template image and to play the corresponding video, and to optionally provide a feedback about any template images to re-rank all of them.

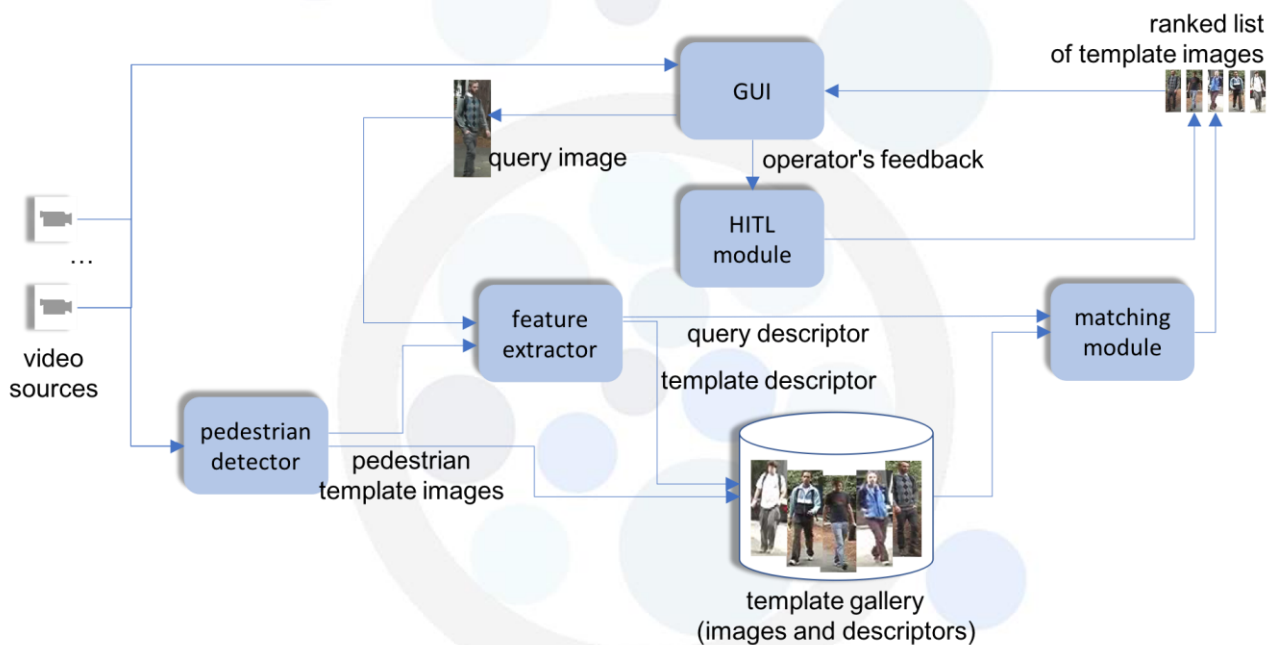


Figure 5 – Functional architecture of the person re-identification module.

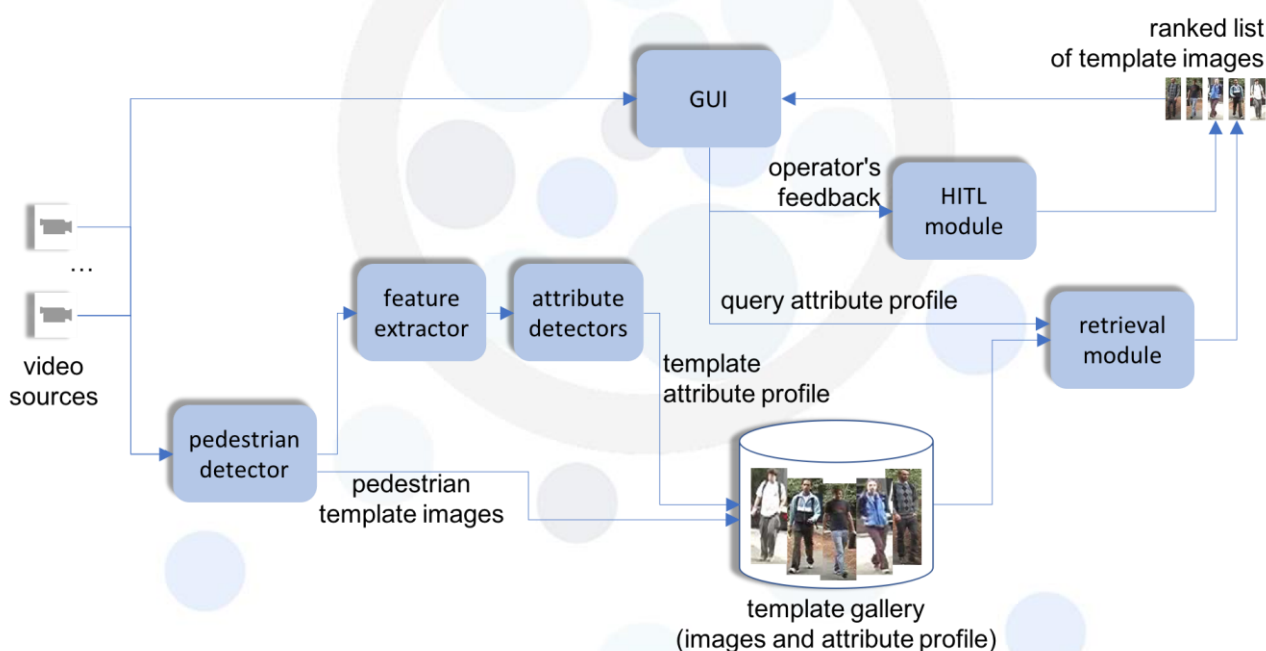
#### 4.1.2 People search module

This module provides the people search functionality summarised in Figure 2 in a **end-to-end system** that, similarly to the person re-identification functionality described above, corresponds to a forensic investigation scenario in the post-event phase. The user uploads one or more stored videos, inputs a description (*attribute profile*) of the target individual in terms of one or more predefined attributes, and starts the retrieval process. The system then retrieves all the pedestrian bounding boxes automatically extracted from all the uploaded videos, and shows the user their images, together with contextual information, sorted by decreasing similarity to the attribute profile. Using the same interface as the person re-identification module, the user can inspect each retrieved image, play the corresponding video, and can optionally refine the ranked list of retrieved images by providing the same feedback as in the person re-identification module, for one or more iterations.

The above functionality is provided by the following software modules, depicted in Figure 6.

- The **pedestrian tracking** module is the same used for person re-identification (see Sect. 4.1.1).

- The **feature extraction** module 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). Note that a different descriptor is used than the one for person re-identification, as described in Sect. 4.3.2.
- One **attribute detector** for each of the predefined attributes is used; each detector takes as input the descriptor of a pedestrian image belonging to the template gallery, produced by the feature extractor, and outputs a continuous score in a predefined range representing the likelihood that the corresponding attribute is present.
- The **template gallery** database is shared with the person re-identification module: in addition to the information described in Sect. 4.1.1, it also stores the descriptor of each template image computed by the above feature extraction module, and the corresponding attribute profile, i.e., the scores produced by the attribute detectors.
- The **retrieval** module takes as input a query consisting of the user-defined attribute profile, 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's attribute profile.
- The **HITL** module is shared with the person re-identification module: it receives from the user the same kind of feedback information described in Sect. 4.1.1, and re-ranks the template images accordingly.
- The **GUI** allows the user to upload one or more videos, to input a query attribute profile, to start the retrieval process, to scroll the ranked list of retrieved template images and related contextual information, to inspect any template image, to play the corresponding video, and to optionally provide a feedback about any template images to re-rank them.



**Figure 6 – Functional architecture of the people search module.**

#### 4.1.3 Crowd monitoring module

This module was initially envisaged to provide three real-time crowd monitoring functionalities at a **macroscopic** crowd level:

- **Crowd density estimation:** estimating the number of people, either in the whole camera view or in a region ("region of interest", ROI) defined by the user, focusing on crowded scenes with severe overlapping and occlusions, where no exact count is possible.



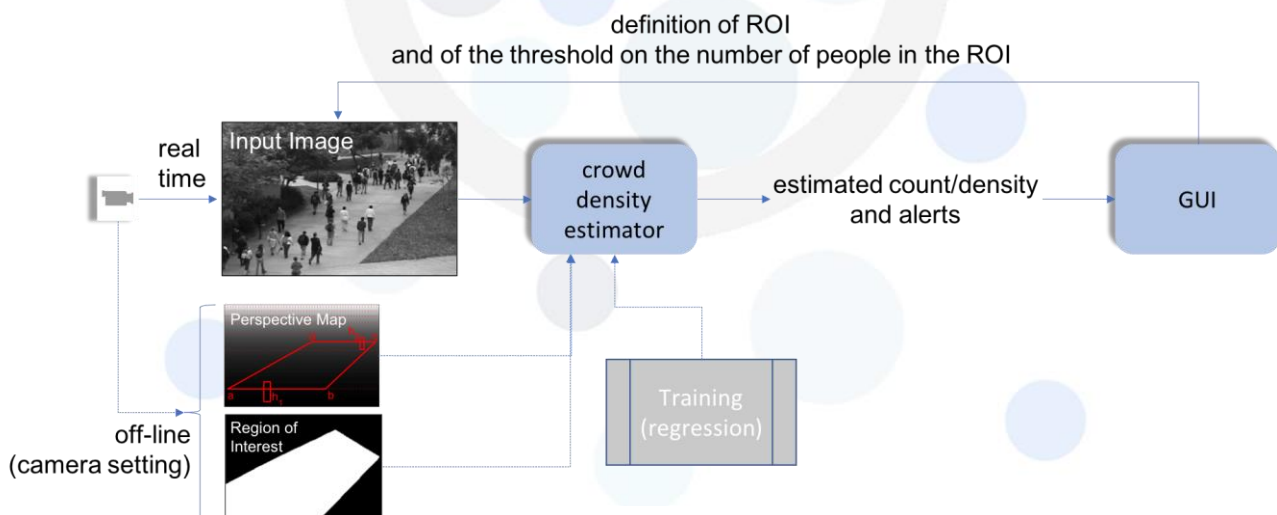
- **Detection of patterns of crowd movement:** detecting the main directions of crowd flow and their velocities.
- **Anomaly detection in crowd behaviour,** in terms of estimated crowd density and patterns of movement, such as: overcrowding with respect to a user-defined threshold; sudden increase or decrease in the estimated crowd density; anomalous 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, possibly using synthetic videos of normal (and possibly anomalous) crowd flows obtained from the CMP tool.

However, during the development of this module it was found that the accuracy of state-of-the-art methods and algorithms for detecting patterns of crowd movement are not yet satisfactory for the purpose of the HCV tool. It was therefore decided to implement only the crowd density estimation and the related anomaly detection functionalities. For the same reason, additional monitoring functionalities related to the analysis of a crowd at the **microscopic** level, indicated in the first version of this deliverable as potential functionalities of interest to the HCV tool, were not further considered: pedestrian tracking in crowded scenes, detection of groups of individuals, detection of specific events and of suspicious behaviours (e.g., people running in a slowly moving crowd, sudden group formation or breaking).

The module that has been developed for crowd density estimation and anomaly detection allows the user to load a stored video, to optionally define a ROI through the mouse (the default ROI being the whole camera view) and a threshold on the number of people inside the ROI, and to play the video. The estimated number of people in the ROI is shown in real time next to the video. The estimated crowd density is also represented as a heat map next to the video. An alert is shown to the user whenever the estimated number of people exceeds the user-defined threshold (if any), and when a sudden increase or decrease in crowd density is detected.

The estimated crowd density and the detected anomalous events can be sent in real time to the LETSCROWD server, and can then be used by the DRA and CMP tools, as further discussed in Sect. 4.2.

The architecture of the crowd density estimation module is shown in Figure 7.



**Figure 7 – Functional architecture of the crowd density estimation module.**

## 4.2 LINKS/INTERFACES BETWEEN TASKS/WORK PACKAGES

The following links between the HCV tool and other LETSCROWD tools have been defined.

- **LETSCROWD server:** the following information generated by the HCV tool is sent to the LETSCROWD server, from which it can be retrieved by other tools:

- the number of people (crowd density) estimated by the crowd monitoring module in the ROI of a given video;
- in a usage scenario of the person re-identification and people search modules during event execution, information about the detection of suspect individuals in different points of the event venue, generated by a LEA operator, can also be sent to the LETSCROWD server.
- **CMP tool** (task T5.1):
  - the estimated crowd density can be used, e.g., to calibrate and validate the simulation set up;
  - synthetic videos showing a simulated crowd, generated by the CMP tool, can be used off-line to train the machine learning-based algorithms used by the crowd density estimation module of the HCV tool.
- **DRA tool** (work package WP3):
  - the estimated crowd density can be used as one of the inputs to update in real time the risk level, according to the approach described in deliverables D3.6 and D3.8;
  - alerts generated by the crowd monitoring module and confirmed by human operators can be used by the DRA tool as **weak signals**,<sup>41</sup>
  - in a usage scenario of the person re-identification and people search modules during event execution, alerts about the presence of suspect individuals in different points of the event venue generated by LEA operators using the person re-identification and people search modules can be used by the DRA tool as weak signals.

More specifically, according to the DRA methodology described in deliverable D3.8, the following information can be sent to the DRA tool through the LETSCROWD application:

- 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 link to the video source of the detection.

### 4.3 ALGORITHM DESCRIPTION

In this section the algorithms used to implement the three modules of the HCV tool are described.

#### 4.3.1 Person re-identification module

This module has been developed using one of the main approaches proposed in the research literature for person re-identification systems, i.e., defining an ad hoc descriptor of pedestrian clothing appearance together with a suitable similarity measure, with no machine learning techniques involved. The HITL mechanism has been implemented using a content-based image retrieval approach with relevance feedback which was investigated in (11).

---

<sup>41</sup> A *weak signal* is defined in deliverable D3.8 as "A seemingly random or disconnected piece of information that at first appears to be background noise but can be recognized as part of a significant pattern by viewing it through a different frame or connecting it with other pieces of information", and is considered "the minimum quantum of information managed by the DRA."

- The pedestrian image **descriptor** is the one proposed in (79). The pedestrian image is first resized to 128×64 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 are used) and Lab (CIELAB) spaces, where the weights are defined by a Gaussian kernel centred at the image centre to reduce the contribution of pixels that are more likely to belong to the background. Subsequently, texture and edge features are extracted. To this aim the top and bottom image strips are not considered, since they are likely to contain the head and the feet as well as a relatively large background region. The three remaining upper strips are merged into a single image region, and the same is done for the three remaining lower strips. From both such regions the LBP texture descriptor and the HOG edge descriptor are then extracted; both of term are defined as histograms. The obtained colour, texture and edge histograms are then concatenated to form the final descriptor.
- The **similarity measure** between two descriptors (represented as column vectors)  $\mathbf{x}_1$  and  $\mathbf{x}_2$  is computed as the negative Euclidean distance  $-(\mathbf{x}_1 - \mathbf{x}_2)^t(\mathbf{x}_1 - \mathbf{x}_2)$ . In the first version of this deliverable (D5.4) the use of the Mahalanobis distance had been considered, due to the planned HITL algorithm, which has been later replaced by another one (see below).
- The **HITL** mechanism is implemented using a content-based image retrieval approach with relevance feedback described in (11). It consists in re-ranking the template images returned by the matching module in response to a given query image, on the basis of the feedback received by the user on one or more template images. For each template image the feedback can take two values: 'similar', if the clothing appearance is considered to be similar to the one of the query image regardless of the identity of the two corresponding pedestrians, or 'dissimilar', if clothing appearance is considered to be diverse from the one of the query image. The relevance feedback is implemented using the probabilistic approach of (80), which estimates the posterior probability that an image is relevant (in this context, 'similar') to the user's query using the nearest neighbour technique in feature space. To this aim the *Relevance Score* algorithm is used, which computes a score for each image  $I$  using its distances to its nearest relevant and non-relevant neighbours, defined as follows:

$$s(I) = \frac{\|I - NN_r(I)\|}{\|I - NN_r(I)\| + \|I - NN_{nr}(I)\|},$$

where  $NN_r(I)$  and  $NN_{nr}(I)$  denote the nearest relevant and non-relevant image to  $I$ , respectively, and  $\|\cdot\|$  is the metric used in feature space – in this case, the Euclidean distance. It is easy to see that the value of the relevance score  $s(I)$  equals 1 for images annotated as relevant ('similar') by the user, and equals 0 for images annotated as non-relevant ('dissimilar'). The rationale of the above relevance score technique is therefore that the closer an image to the nearest relevant image in feature space, the higher its relevance, and the closer an image to the nearest non-relevant image, the lower its relevance. After the user has provided a feedback on one or more template images, the relevance score is computed for all the other ones (the score for the selected images is set to 1 or to 0 as explained above), and the template images are re-ranked for decreasing values of the score. This step can be repeated for any number of times, until the user provides a feedback on the retrieved template images.

In the first version of this deliverable (D5.4) the use of the HITL algorithm of (14) was planned; however the implementation of that algorithm turned out to be rather complex, and not enough information was provided in (14) to this aim.

#### 4.3.2 People search module

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 CNN which takes as input a pedestrian raw image, and acts both as a feature extractor and as a classifier/detector. In the HCV tool the former, simplest approach has been used.

- The chosen **feature** descriptor is the Ensemble of Localised Features (ELF) used in (22). 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), which is made up of a collection of several benchmark person re-identification data sets, and is currently the largest available data set of pedestrian images taken both in outdoor and in indoor video surveillance settings. Each image is manually annotated according to 65 different attributes.
- For each attribute, a specific **detector** is implemented as a binary SVM classifier with the histogram intersection kernel. Note that all the attributes of the PETA data set are binary except for four 11-valued attributes, which have been converted into 11 binary attributes. The attribute detectors have been pre-trained on the whole PETA data set. Since for several attributes very few images are present in this data set, the corresponding attribute detectors exhibited a very low accuracy evaluated in terms of the Pr and Re measures. For instance, the 'accessory Headphone' attribute is present only in 31 out of 19,000 images; the corresponding *classification* accuracy was 0.9887, but Pr and Re were equal to 0.0175 and 0.0323, respectively; in other words, the label assigned by the 'accessory Headphone' detector (either present or not) was correct on more than 98% of the testing images, but among images labelled as exhibiting such an attribute only 1.75% actually exhibited it, and among the ones actually exhibiting it only 3.23% were correctly labelled as such. Accordingly, only the following subset of the attributes available in the PETA data set has been used in the people search module, selected among the ones exhibiting the highest Pr and Re values, corresponding to 37 binary attributes:
  - **upper body clothing colour**: black, blue, brown, green, grey, orange, pink, purple, red, white, yellow (11 binary attributes, one for each colour);
  - **upper body clothing style**: casual, formal (2 binary attributes);
  - **type of sleeve**: long, short (2 binary attributes);
  - **lower body clothing colour**: black, blue, brown, green, grey, orange, pink, purple, red, white, yellow (11 binary attributes, one for each colour);
  - **lower body clothing style**: casual, formal (2 binary attributes);
  - **pants type**: jeans, hot pants, trousers (3 binary attributes);
  - **type of shoes**: sneakers, leather (2 binary attributes);
  - **gender**: female, male (2 binary attributes);
  - **accessory – hat**: present, absent (1 binary attribute);
  - **accessory – backpack**: present, absent (1 binary attribute).
- The **HITL** approach has been implemented using the same method used for the person re-identification module, described in Sect. 4.3.1. As an additional, complementary method, the one envisaged in the first version of this deliverable (D5.4) can also be used. It consists of allowing the user to provide a feedback on the actual presence or absence of any subset of attributes for one or more of the retrieved template images, for a given query attribute profile. 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 can 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 as positive examples. To reduce the processing cost of the re-training phase, two possible solutions can 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. In any case the



classifiers can be re-trained off-line, and each detector can be updated only when the re-training phase is complete.

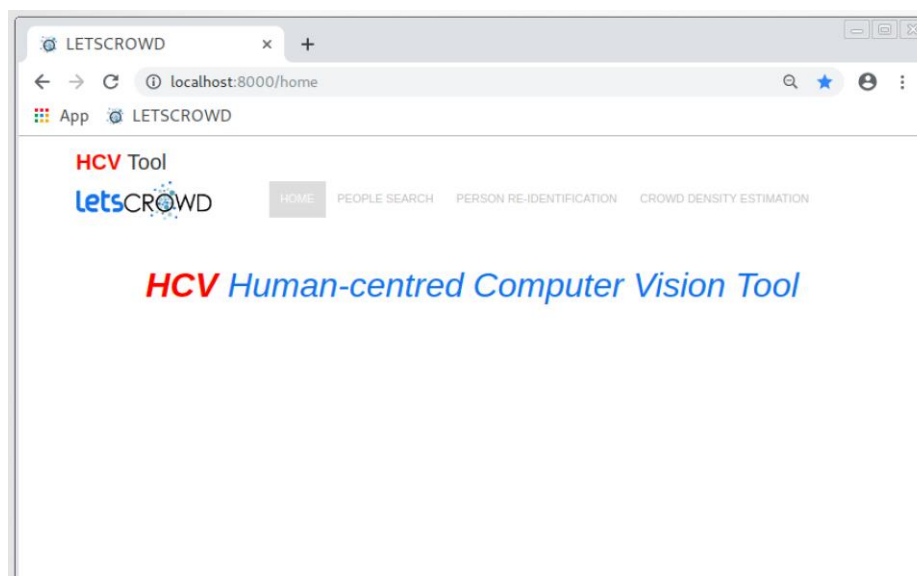
#### 4.3.3 Crowd monitoring module

Based on the guidelines provided in the survey of (33) for crowd density estimation in crowded scenes (see Sect. 2.1.3), an algorithm based on the counting-by-regression approach has been used to implement the crowd density estimation module. This algorithm provides an estimate of the number of people in a single video frame (scene), or in a user-defined ROI inside a frame, using static features. Its main processing steps are the following.

- **Perspective normalization** has been implemented as described in (33), 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** are extracted to represent a whole scene and no detection of single pedestrians is attempted, according to the counting-by-regression approach. Different kinds of features have been considered among the ones described in (33), where it was shown the effectiveness a given set of features depend on the kind of scene:
  - **foreground segment** features obtained through background subtraction, including the number of pixels in the segment and in its perimeter, the perimeter-area ratio, the perimeter edge orientation histogram, and the 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, including the number of edge pixels, the histogram of edge orientations, and the Minkowski fractal dimension of the edges; edges are extracted using the well-known Canny detector;
  - **texture** features, using the LBP descriptor;
  - and **gradient** features, using the grey-level co-occurrence matrix (GLCM).
- Linear and non-linear **regression models** have been used to map the (perspective normalised) feature vector extracted from a scene to an estimate of the number of people in the same scene. According to (33), more complex non-linear regression models provided often better results on benchmark data sets, when the testing images exhibited similar characteristics as the training images; however the simplest linear models turned out to be more robust to the mismatch between training and testing images which is likely to occur in real-world applications. In the HCV tool two versions of the linear regression model have been tested: 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. Two different non-linear models have also been tested: Random Forests and SVM with radial-basis function kernel.
- As a **training set** for the regression algorithm the following publicly available data sets have been used (see Sect. 2.4): QMUL Mall, UCSD Pedestrian Traffic and PETS 2009. Additionally, synthetic videos obtained from the CMP tool have been used, reproducing the PETS 2009 scenes and the scenes acquired during the practical demonstrations of the LETSCROWD project.

#### 4.4 USER INTERFACE

The HCV tool has been implemented as a client-server application with a Web-based interface. The starting window (Web page) is shown in Figure 8. The links at the top of the window give access to the three functionalities of person re-identification, people search and crowd density estimation.



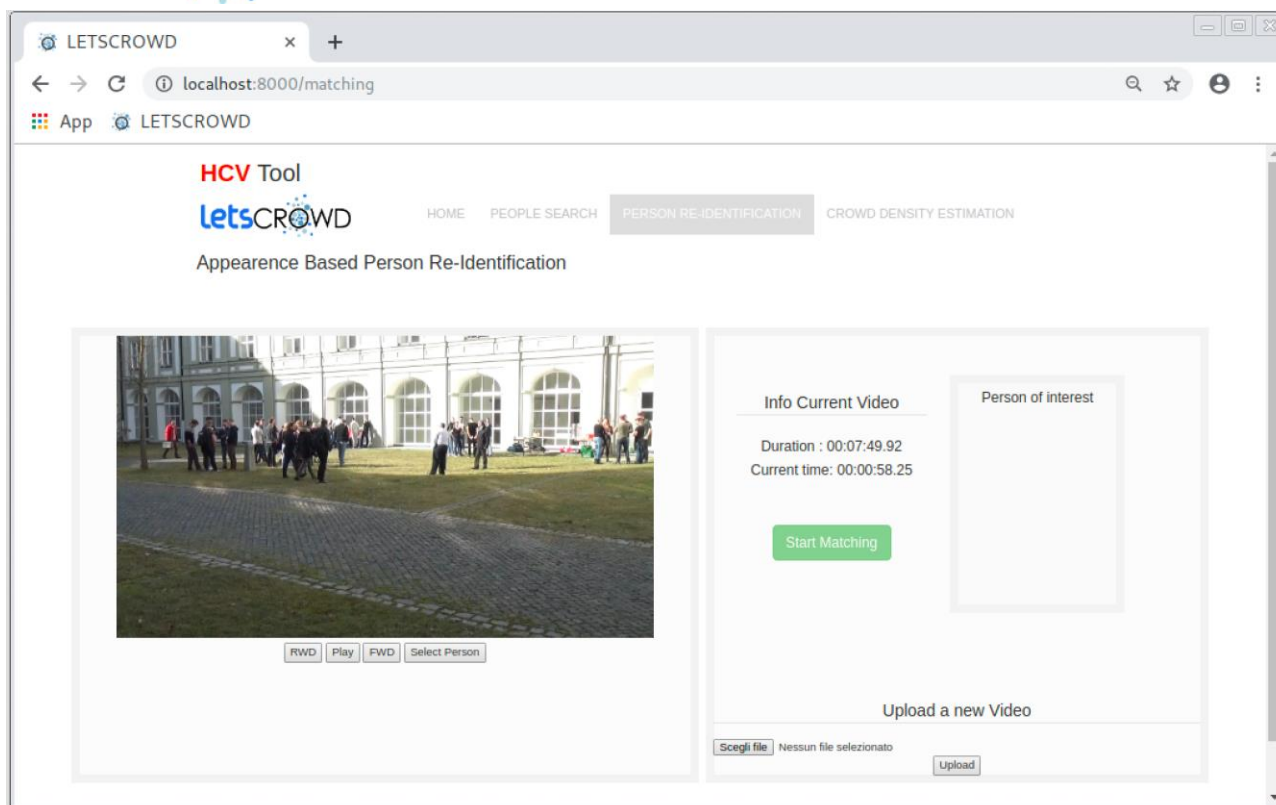
**Figure 8 – Initial window (Web page) of the HCV tool GUI. The three functionalities of person re-identification, people search and crowd density estimation can be accessed through the links at the top of the window.**

The GUI of the three modules are described in the following subsections.

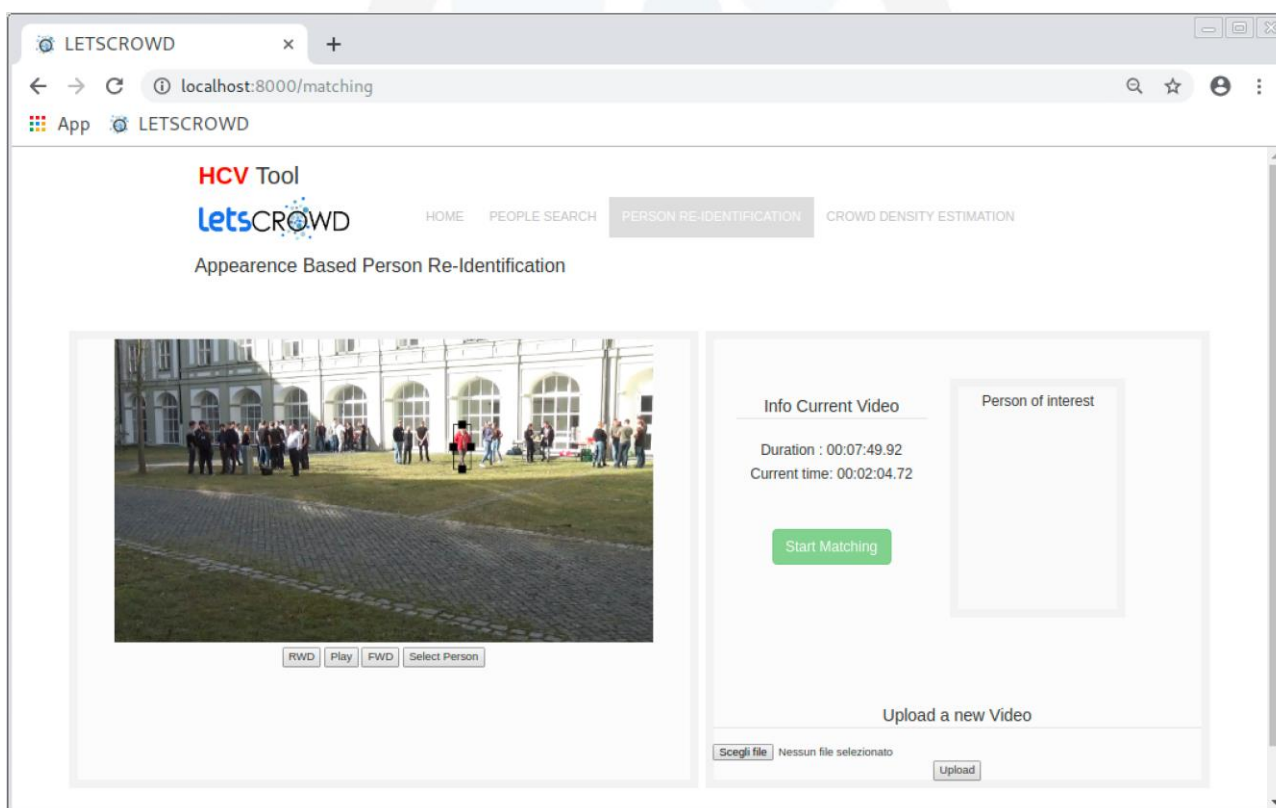
##### 4.4.1 Person re-identification module

The GUI is made up of three different views: the **main** window, the **results** window and the **template detail** window.

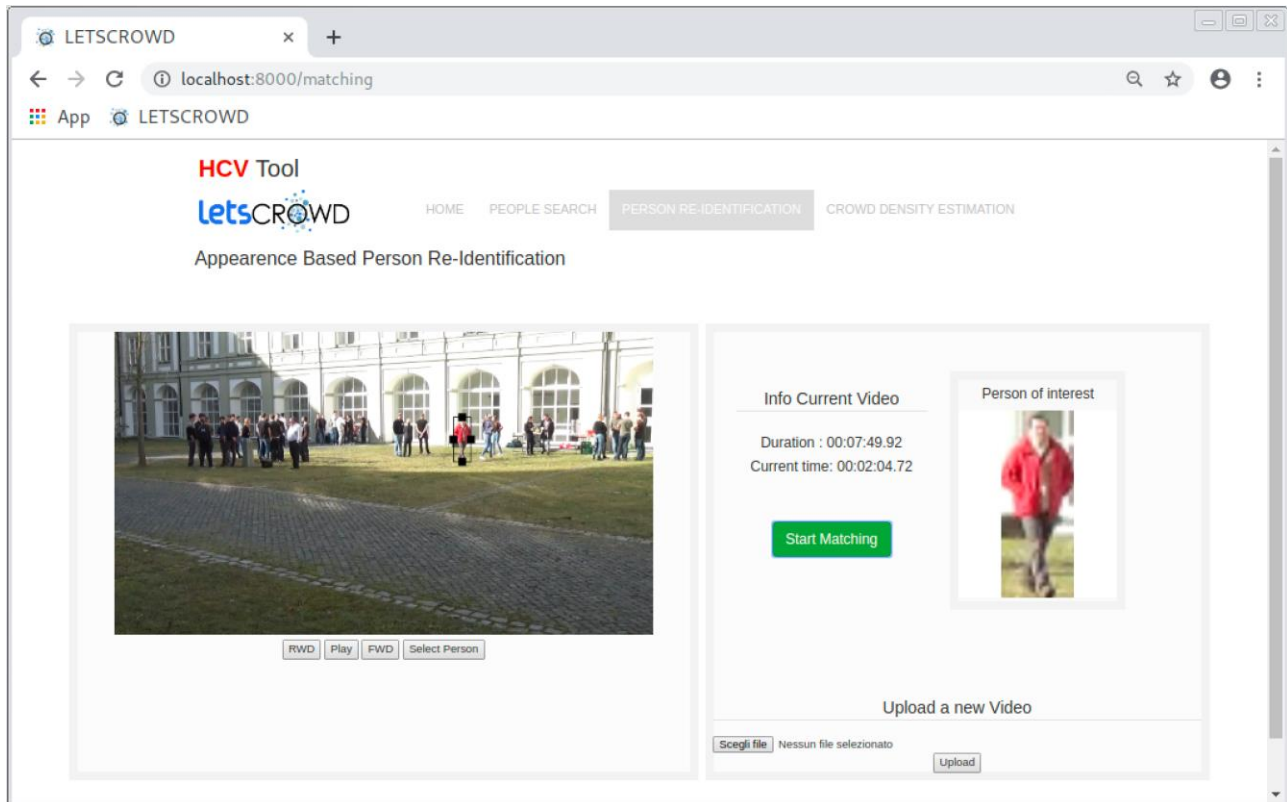
- The **main window** is shown in Figure 9. It allows the user to load one or more videos, and shows on the left the last loaded video. This video can be played and stopped using the controls shown below it. When a video is stopped the user can select on the frame shown in the GUI a rectangular bounding box containing an individual of interest, using the mouse; the bounding box can be resized by pointing, clicking and dragging the mouse on one of its corners (see Figure 10). Through the button 'Select person' next to the video controls the user can make the current bounding box the query image, which is shown on the right, and can start the retrieval process using the button 'Start matching' shown next to the query image (see Figure 11).



**Figure 9 – GUI of the person re-identification module: Main window.**



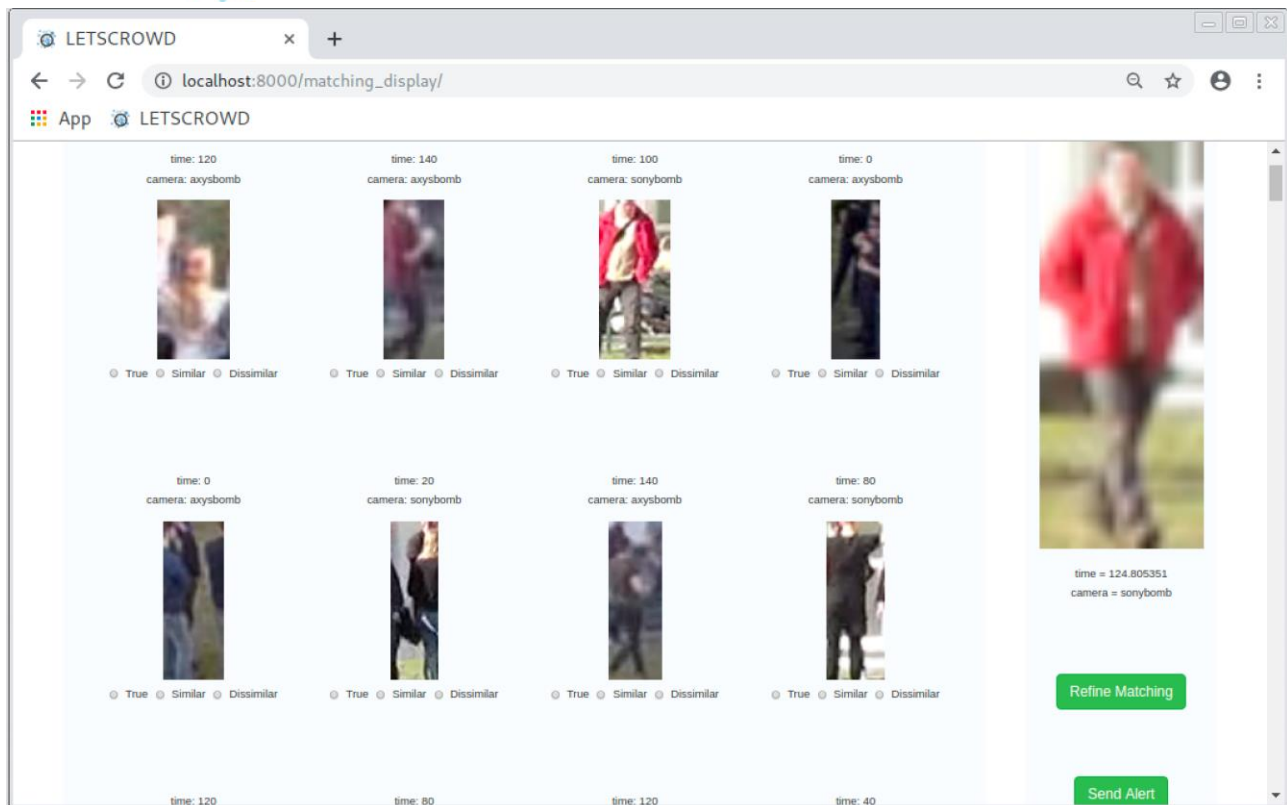
**Figure 10 – Selection of a bounding box containing the image of an individual in the Main window of the person re-identification GUI.**



**Figure 11 – Selection of the query image and start of the retrieval process from the Main window of the person re-identification GUI.**

- The outcome of the retrieval process is shown in a **results window** which replaces the main window, and is shown in Figure 12. The query image is shown on the right. The template images are shown in the main area of the window, ranked for decreasing similarity to the query, from top to bottom and from left to right. Context information is shown next to each template image (timestamp and name of the video file – in a real system this can be the ID of the camera from which the image was obtained), as well as two radio buttons which allow the user to provide an optional feedback on that particular image (either 'similar' or 'dissimilar' to the query image). The results window can be scrolled by the user to see all the retrieved template images. The operator can also further analyse any template image of interest by clicking on it, which opens the template detail window (see below). Next to the query image on the top of the results window the button 'Refine matching' allows the user to re-rank the retrieved template images based on the feedback provided by the user itself on one or more such images; if no feedback was provided this button is inactive, and an alert message is shown asking the user to provide a feedback on at least one template image. The re-ranked list of template images is then shown in the same results window. If the user finds template images of the same identity as the query, he/she can store information about those images (as well as the analogous information about the query image) by selecting them through the radio button 'True' (standing for 'True match', i.e., the same identity as the query image) next to the feedback radio buttons, possibly after inspecting the corresponding videos through the Template detail window (see below). In an application scenario during event execution, all the stored information can then be sent to the LETSCROWD server by clicking on the button 'Send alert' which is shown below the query image; this button opens a text field where the user can input an optional text to describe the alert generated. In a post-event, forensic investigation scenario, the same information can instead be stored for later use by the investigators. To return to the main window the user has to click on the link at the top of the page.





**Figure 12 – Results window of the person re-identification GUI: query image (right), and ranked list of the retrieved template images together with context information and user feedback controls.**

- When the user would like to further analyse a retrieved template image shown in the Results window, he/she has to click on that image, which opens the **Template Detail window** shown in Figure 13. This window shows the query image on the right, the selected template image on the top left together with the same contextual information shown in the results window (see above), and the whole video frame from which that image was extracted. This video can be played back by using the controls shown below it. This allows the user to analyse the individual in the template image (e.g., to see if the identity is the same as the one in the query image), as well as the behaviour of that individual (e.g., to get additional insights about whether the exhibited behaviour can be considered as suspect). By closing the template details window the results window is shown again.

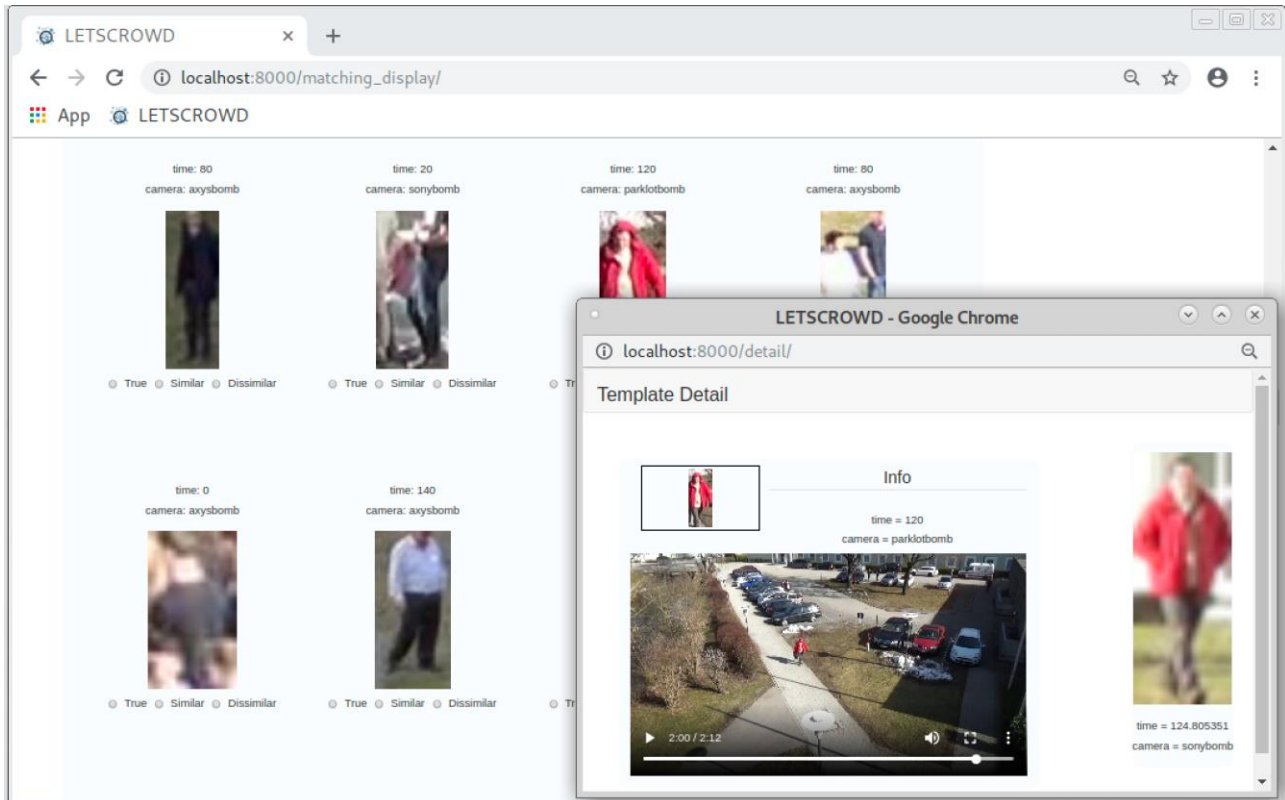
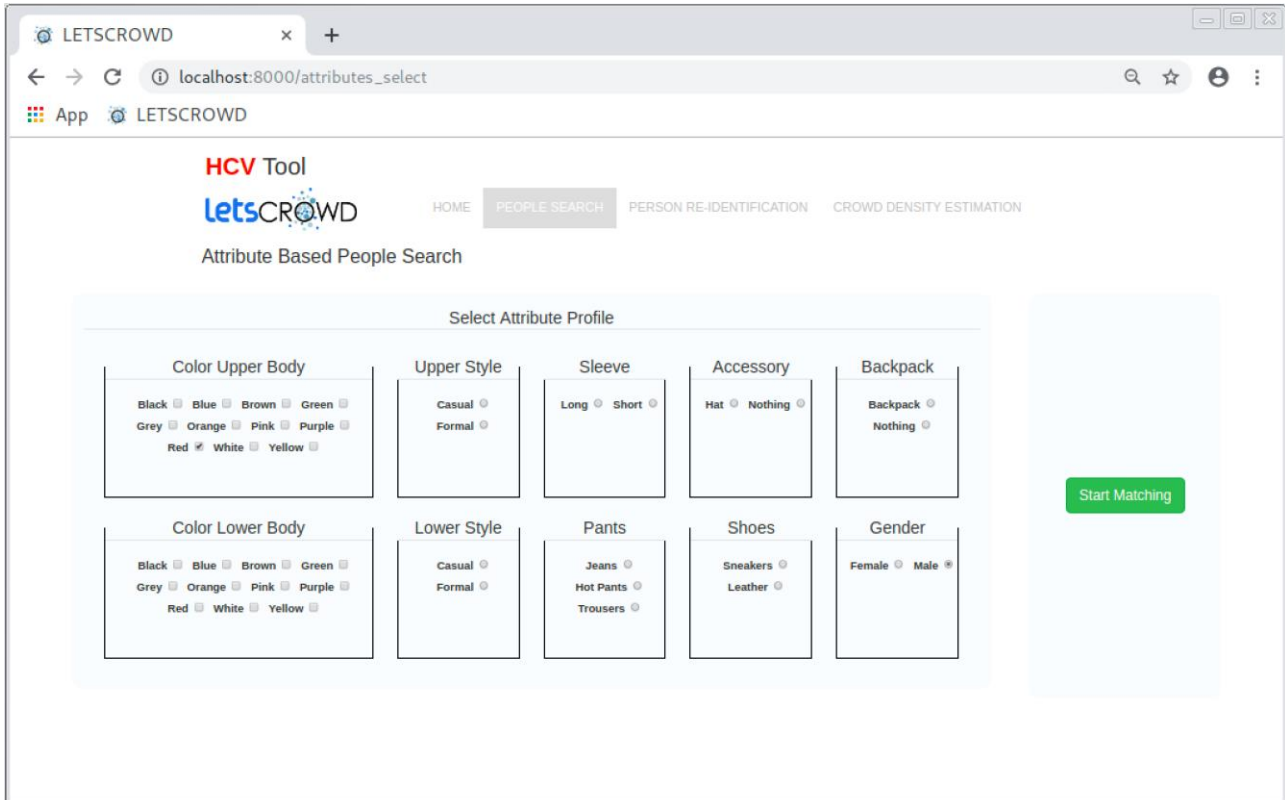


Figure 13 – Template Detail window of the person re-identification GUI.

#### 4.4.2 People search module

The GUI is made up of three different views: the **main** window, the **results** window and the **template detail** window.

- The **Main window**, shown in Figure 14, allows the operator to input the **attribute profile** which describes the clothing appearance and other attributes of the target individual. A predefined set of attributes is shown in the main window, grouped into the following categories: upper body clothing colour, upper body clothing style, type of sleeve, lower body clothing colour, lower body clothing style, pants type, type of shoes, gender, accessories (hat and backpack). Inside each category the available attributes are shown either as checkboxes or as radio buttons; checkboxes are used for non-mutually exclusive attributes like clothing colours (e.g., upper body clothing can exhibit more than one colour), whereas radio buttons are used for mutually exclusive attributes such as gender. By default no checkbox and no radio button is selected. After the target attribute profile is input by selecting at least one checkbox or radio button, the retrieval process can be started by clicking the 'Start matching' button. The retrieved template images are shown in the Results window (see below).



LETSCROWD

localhost:8000/attributes\_select

App LETSCROWD

HCV Tool

letsCROWD

HOME PEOPLE SEARCH PERSON RE-IDENTIFICATION CROWD DENSITY ESTIMATION

Attribute Based People Search

Select Attribute Profile

Color Upper Body

Black Blue Brown Green Grey Orange Pink Purple Red White Yellow

Upper Style

Casual Formal

Sleeve

Long Short

Accessory

Hat Nothing

Backpack

Backpack Nothing

Color Lower Body

Black Blue Brown Green Grey Orange Pink Purple Red White Yellow

Lower Style

Casual Formal

Pants

Jeans Hot Pants Trousers

Shoes

Sneakers Leather

Gender

Female Male

Start Matching

**Figure 14 – Main window of the people search GUI. The selected attribute profile corresponds to a male wearing some red item of clothing in the upper body.**

- The outcome of the retrieval process is shown in the **Results window**, which replaces the main window and is shown in Figure 15. This window is identical to the homonymous window of the person re-identification GUI, except for the following details (see Sect. 4.4.1 for the common details):
  - the chosen attribute profile is shown on the right, instead of a query image;
  - the retrieved template images are ranked for decreasing similarity to the attribute profile instead of to a query image;
  - if the user has flagged one or more template images as 'True' (i.e., the attribute profile of the corresponding individual matches the target one), the corresponding information can be either sent to the LETSCROWD server or stored for later use together with the target attribute profile, instead of the analogous information about a query image.

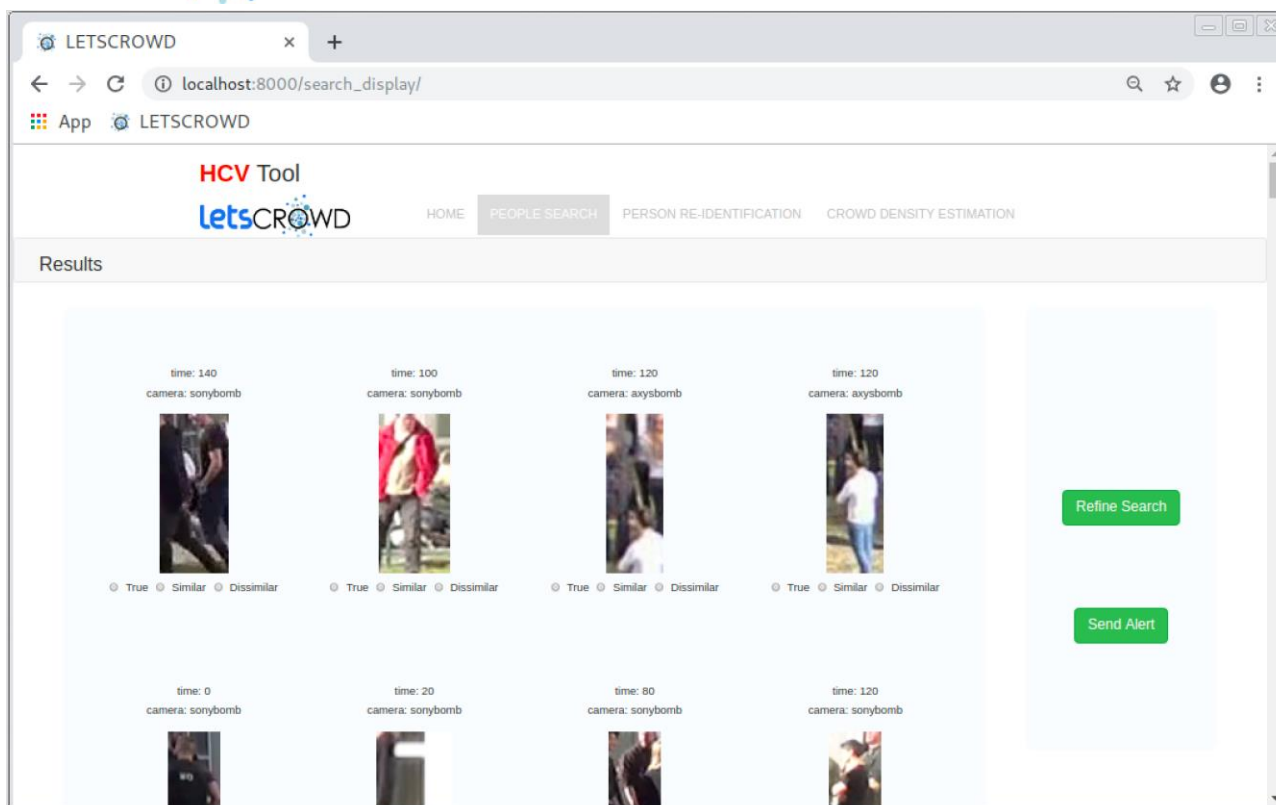


Figure 15 – Results window of the people search GUI, corresponding to the attribute profile shown in Figure 14.

- When a template image is selected by the user in the results window for further analysis, a **Template Detail window** opens, which is shown in Figure 16. This window is identical to the homonymous window of the person re-identification GUI.



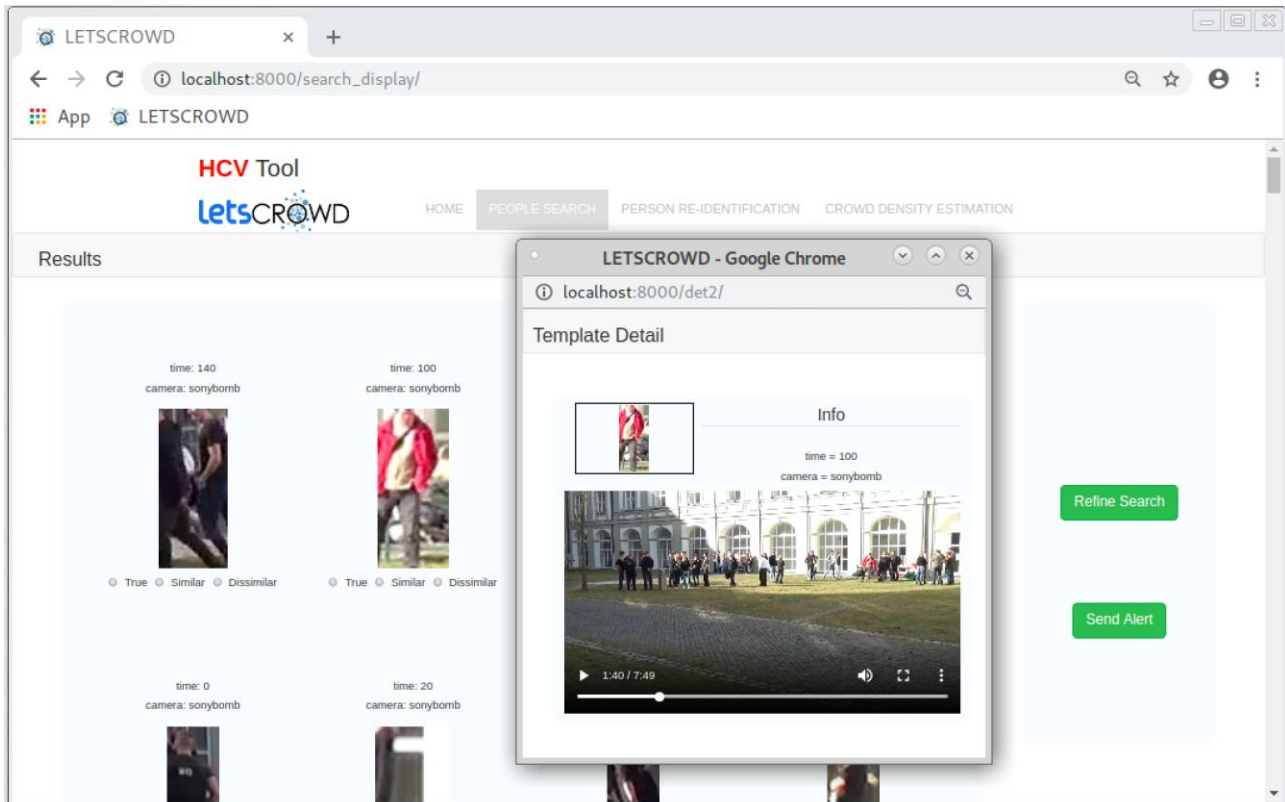


Figure 16 – Template Detail window of the people search GUI.

#### 4.4.3 Crowd monitoring module

The GUI of the crowd monitoring module is made up of the single window shown in Figure 17, which provides the crowd density estimation and the related anomaly detection functionalities. This window allows the user to load a video, which is shown on the left. The video can be played using the controls below it. The following information is shown below the video: total duration, elapsed time since the beginning and estimated number of people in the current ROI. The default ROI corresponds to the whole scene. The user can select a different, rectangular ROI by selecting the checkbox 'Select region of interest' below the video, then clicking with the mouse on any point in the video, which is considered as one the vertices of the ROI, and finally dragging the mouse and releasing it on the opposite vertex (see Figure 18). From that moment onward the estimated crowd density refers to the new ROI. The ROI can be reset to the whole scene by deselecting the above checkbox.

On the right of the window the same video is shown with a semi-transparent heat map superimposed, which represents the estimated density in predefined rectangular sub-regions of identical size.

Below the heat map the user can input a threshold on the number of people in the ROI, such that an alert is automatically generated and shown on the screen if the estimated crowd density exceeds that threshold (see below). The threshold can be removed by selecting a value of 0.

Alerts are automatically generated when the estimated number of people in the ROI exceeds the user-defined threshold (if any) and when a sudden increase or decrease of the estimated number of people in the ROI is detected. Such alerts are shown to the user as pop-up windows.

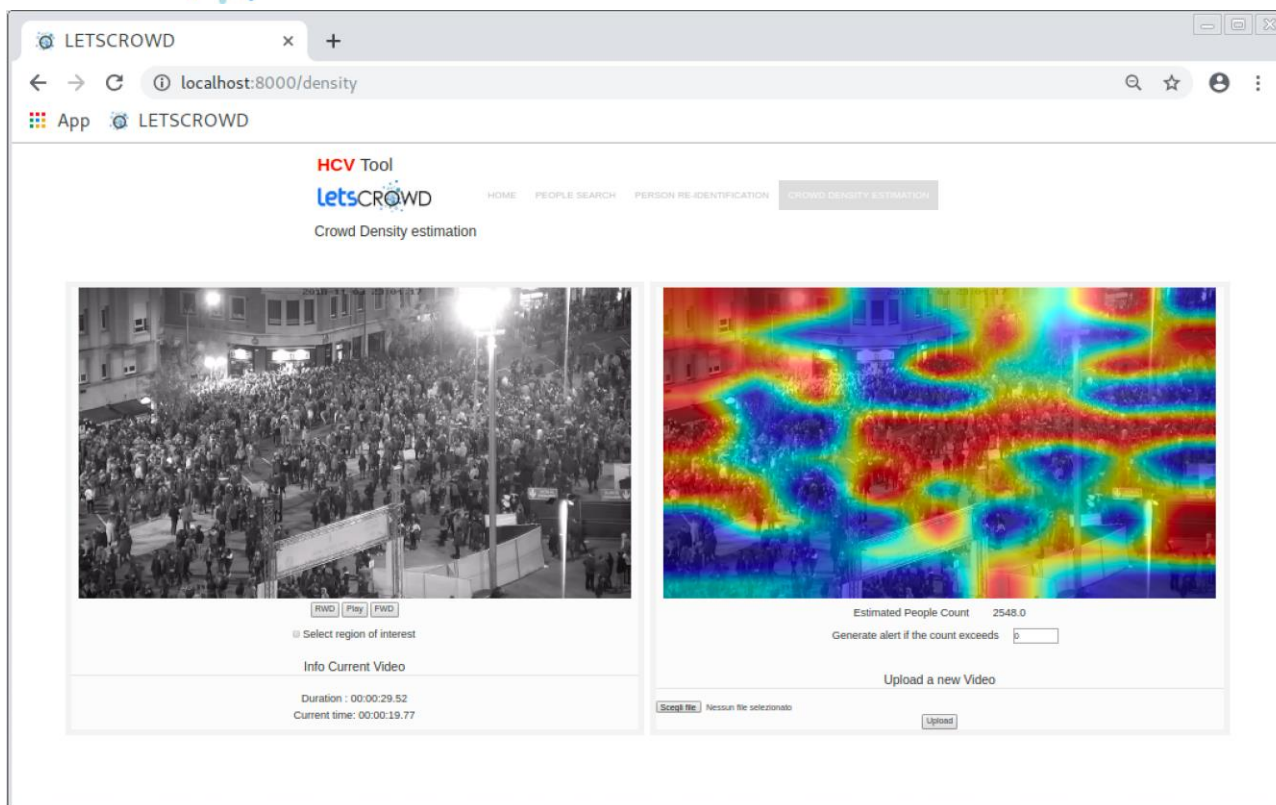


Figure 17 – GUI of the crowd density estimation module.

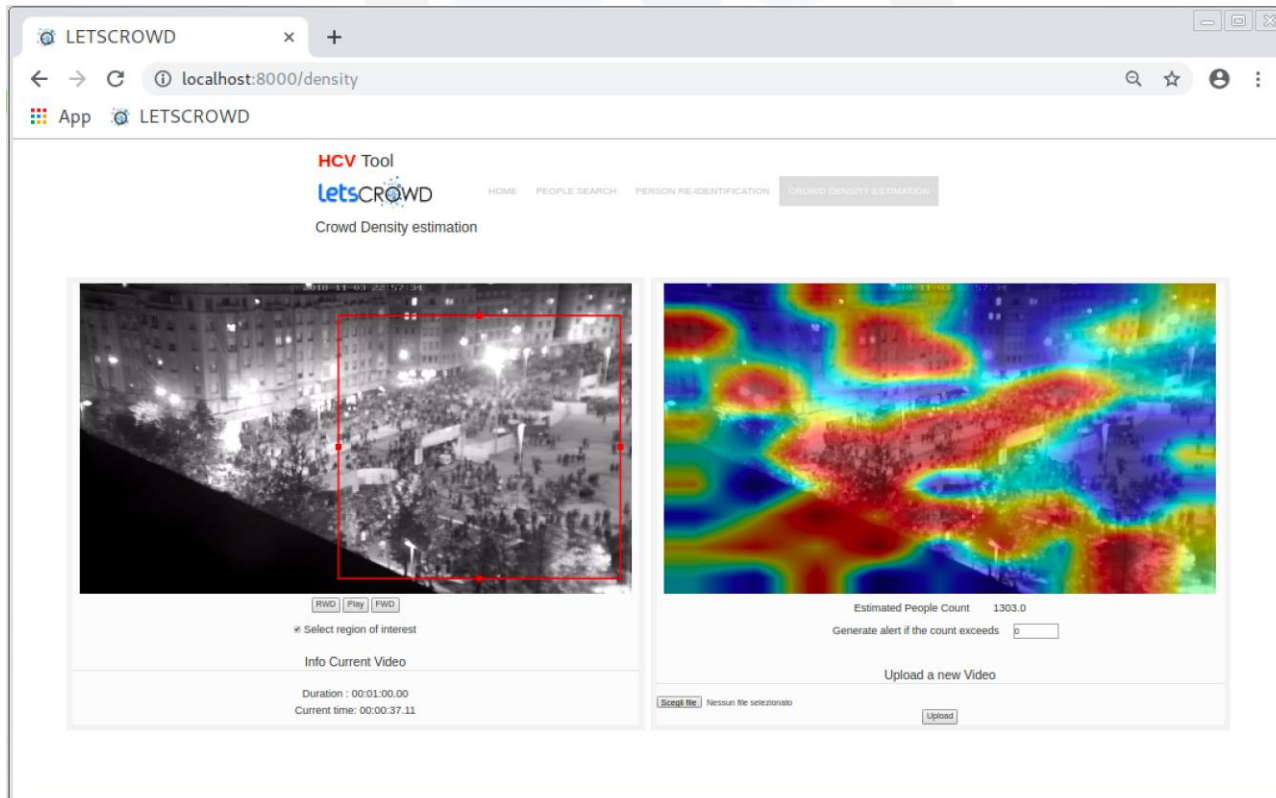


Figure 18 – GUI of the crowd density estimation module: selection of a region of interest (red rectangle on the left).

## 5 IMPLEMENTATION

This section reports details about the implementation of the three modules of the HCV tool, including the GUI.

### 5.1 PERSON RE-IDENTIFICATION MODULE

#### 5.1.1 Implementation details

The person re-identification module has been implemented using the following programming languages, libraries and frameworks:

- The Python<sup>42</sup> (v. 2.7) open-source language, developed under an OSI-approved open source license which makes it freely usable and distributable.
- The following open-source Python libraries for plotting functionalities and for computer vision and machine learning algorithms:
  - matplotlib<sup>43</sup> (v. 2.1.2): a 2D plotting library;
  - numpy<sup>44</sup> (v. 1.15.1): a package for scientific computing;
  - skimage<sup>45</sup> (v. 0.13.1): an image processing toolbox.
- The OpenCV<sup>46</sup> (Open Source Computer Vision) library (v. 3.3), released under a BSD license, freely available 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 open-source pedestrian tracking software available on <https://github.com/ambakick/Person-Detection-and-Tracking>.
- The template gallery database has been built using MongoDB<sup>47</sup> (Community Edition), a widespread non-relational ("NoSQL") document-oriented database management system (DBMS). The PyMongo<sup>48</sup> driver and PyMODM<sup>49</sup> framework have been used to access MongoDB from Python code.
- The server side of the tool has been implemented using the high-level Python Web framework Django.<sup>50</sup> The communication with the client side has been implemented through JavaScript, using the Ajax pattern,<sup>51</sup> the JQuery library,<sup>52</sup> and the Json format to encode the server-to-client communication.
- The client side (GUI) has been implemented in HTML.

With reference to the functional architecture shown in Figure 5, the only component already fully available as open source software was the pedestrian tracking tool. All the other components have been specifically developed for the LETSCROWD project. This includes the feature extraction and HITL algorithms taken from

---

<sup>42</sup> <https://www.python.org/>

<sup>43</sup> <https://matplotlib.org/2.1.2/>

<sup>44</sup> <http://www.numpy.org/>

<sup>45</sup> <http://scikit-image.org/>

<sup>46</sup> <https://opencv.org/>

<sup>47</sup> <https://www.mongodb.com/>

<sup>48</sup> <https://docs.mongodb.com/ecosystem/drivers/python/#python-driver>

<sup>49</sup> <https://api.mongodb.com/python/current/tools.html>

<sup>50</sup> <https://www.djangoproject.com/>

<sup>51</sup> [https://www.w3schools.com/xml/ajax\\_intro.asp](https://www.w3schools.com/xml/ajax_intro.asp)

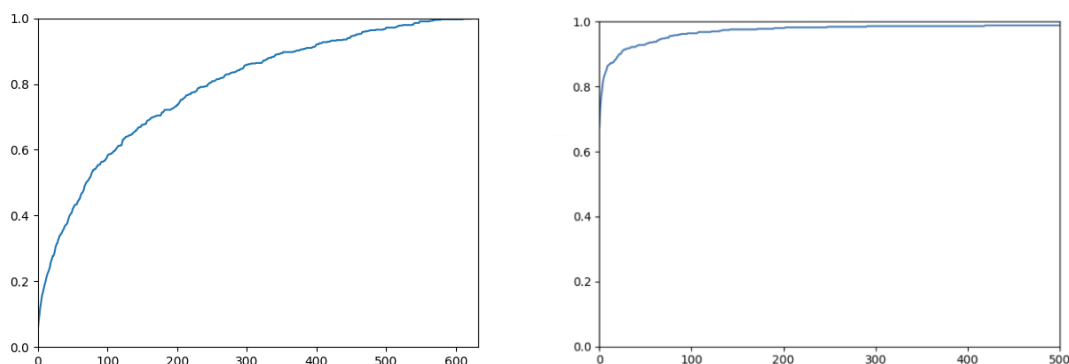
<sup>52</sup> <https://jquery.com/>

(79) and (11), respectively, for which no software implementation was provided by the authors of the respective papers.

### 5.1.2 Validation

During the **design phase** the validation of the person re-identification algorithm described in Sect. 4.3.1 has been carried out on two publicly available and widely used benchmark data sets mentioned in Sect. 2.4: VIPeR and Market-1501. VIPeR consists of manually extracted bounding boxes of 632 pedestrians, with a size of 128×48 pixels; 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. 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. Market-1501 consists of 32,668 automatically extracted and manually verified bounding boxes of 1,501 pedestrians, and of a number of bounding boxes corresponding to false detections, named "distractors", which have been included to mimic real application scenarios. All bounding boxes have been resized to 128×64 pixels.

The standard experimental set-up has been used for VIPeR: one randomly chosen image of each identity has been used as a query, and the other one has been used as a template. For Market-1501 the set-up suggested by the authors of this data set has been considered: 750 identities have been used as query individuals, and one image of each of them was randomly chosen from each camera, resulting in 3,368 query images (note that not all individuals appear in images from all six cameras); the template gallery was made up of 19,732 images. In both data sets for each query identity there is at least one template image in the template gallery (for VIPeR, exactly one image). 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 curves obtained using the chosen descriptor and similarity measure are shown in Figure 19. Note that the CMC curve for Market-1501 is limited to the first 500 ranks, since the top ranks are the most relevant ones, and the recognition rate after the 500th rank is very close to 1. As an example, the CMC curve on VIPeR 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 of templates; for about 40% of the query images it was among the top 50 positions, and so on. The recognition rate for Market-1501 is about 0.7 at the first tank, and exceeds 0.9 at rank 50. It is worth noting that the recognition rate on Market-1501 is higher than in VIPeR, despite the template gallery of Market-1501 is larger. The performance for both data sets is similar to the one reported in the literature for the same kind of descriptors and similarity measures (79), (14).



**Figure 19 – CMC curves obtained on the VIPeR (left) and Market-1501 (right) data sets, using the image descriptor and similarity measure described in Sect. 4.3.1. Note that the CMC curve for Market-1501 shows only the first 500 ranks out of the 19,732 template images.**



The person re-identification module has been then tested "in the field" in two of the **practical demonstrations** of the LETSCROWD project by officers of the organizer LEAs, that took place before April 2019:

- A cyclo-cross race – a real mass gathering event – in OostMalle (Belgium, Feb. 23-24, 2019), where the practical demonstration was organized by the Lokale Politie Voorkempen (LPV) with the support of the Belgian Federal Police for the installation and management of an ad hoc video surveillance system (five PTZ cameras and a command post);
- Simulations with volunteers at the main campus of the Hochschule für den öffentlichen Dienst in Bayern – Fachbereich Polizei (BayHfoeD) in Fürstenfeldbruck (Germany, Feb. 25-27, 2019), where several videos were acquired through video cameras provided by the LETSCROWD technology providers involved (including UNICA) and by BayHfoeD.

In both practical demonstrations (including the one in OostMalle, carried out in the context of real mass gathering event) volunteers were recruited to play the role of 'suspect individuals', to guarantee that they appeared in the videos of all the available cameras.

The details about these practical demonstrations, including the outcomes of the validation process involving several aspects beside the recognition accuracy of the tested tools, are reported in deliverable D6.3.

As regards **recognition accuracy**, in these tests the individuals in the query images appeared at ranks between 10% to 20% the size of the template gallery, which was approximately 200 and 500 for the practical demonstrations in Fürstenfeldbruck and in OostMalle, respectively. On the one hand, these results can be considered somewhat analogous to the ones observed in laboratory experiments on benchmark data sets (see above). On the other hand, the LEA officers who tested the person re-identification module felt that the position of the query identity in the ranked list of template images was often too high, i.e., too many false positives were present. In particular, in some cases several template images apparently very different from the query image were ranked higher than the template of the same identity as the query, which was found counterintuitive. These results were partly due to the following reasons, that contributed to the number of false positives:

- The template gallery was populated in a fully automatic way using all the bounding boxes provided by the pedestrian *detection* module of the HCV tool (see Sect. 4.3.1). A non-negligible number of such bounding boxes turned out to be false detections, or to contain only a part of an individual, or more than one individual; moreover, several bounding boxes contained also a significant portion of the background.
- In the version of the HCV tool that was used in these practical demonstrations a pedestrian *detector* was used instead of a pedestrian *tracker*. This choice was made to ease the implementation, since it did not require to select one or more bounding boxes from each detected track. On the other hand, since bounding boxes were selected from each frame independently on the other frames, several images of a same individual detected in subsequent frames, very similar to each other, were stored in the template gallery. Moreover, some false detections may be filtered out by a *tracker*.
- The HITL module, which exploits user's feedback to improve the ranking of template images of the query identity (when they are present in the template gallery), was not available yet in the HCV tool.

To sum up, the general feeling from the practical demonstrations is that a higher recognition accuracy would be required by LEAs to a person re-identification tool in real mass gathering events. Beside the possible directions for improvement discussed above, another potential direction is the use of more discriminant appearance descriptors (features), for instance by exploiting deep learning techniques. On the other hand, more complex descriptors usually require a higher processing time, which can be an issue in real-world applications where the size of the template gallery is huge due to the large number of cameras and to the length of the recorded videos.

Finally, a related practical issue is that of **camera setting** (position, resolution, etc.). As emerged from user requirements (Sect. 3.1) LEAs may use also cameras belonging to external CCTV systems; moreover, some

camera settings are regulated by country-specific rules. Accordingly, system requirements (Sect. 3.3) included only general indications: tilt angle with horizontal plane less than 45 degrees, height of about 3 m or less. With regard to resolution, images of most benchmark data sets have a 128×64 pixel size or similar, which can be considered to be approximately the minimum required image size for the matching algorithm of person re-identification systems to be effective. This can be translated into a constraint on the maximum distance of pedestrians to the camera which is useful for person re-identification. The above image size can also be used as a reference to filter the images (bounding boxes) coming from the pedestrian detection or tracking tool to be stored in the template gallery: images of smaller size should be discarded as they are likely to be too small to allow an accurate matching with a query image; images with a significantly different aspect ratio should be discarded as well, as they are likely to be false detections, or to contain more than one individual or a large background region, which can in turn degrade matching accuracy. In videos acquired by mobile cameras the presence of blur reduces image quality and is likely to severely degrade matching accuracy, although this issue has not been addressed in the literature, and no mobile cameras were used in the practical demonstrations involving the HCV tool.

## 5.2 PEOPLE SEARCH MODULE

### 5.2.1 Implementation details

The same programming languages, libraries and frameworks as in the person re-identification module have been used (see Sect. 5.1.1). Additionally, the following Python libraries were used to implement the attribute detectors:

- sklearn<sup>53</sup> (0.19.2): a machine learning toolbox;
- tensorflow<sup>54</sup> (1.8.0): a machine learning toolbox for deep neural networks.

With reference to the functional architecture shown in Figure 6, the only components already fully available as open source software was the pedestrian tracking tool (as for the person re-identification module) and the machine learning software used to train the attribute detectors. All the other components have been specifically developed for the LETSCROWD project.

### 5.2.2 Validation

The attribute detectors for each the 37 binary attributes listed in Sect. 4.3.2 have been trained on the PETA data set described in Sect. 2.4. Their detection accuracy was evaluated at **design phase** on the same PETA data set, and was similar to the one reported in the literature (59). For instance, for the upper body colours the values Pr and Re ranged from 0.18 and 0.41 for 'Yellow' to 0.71 and 0.72 for 'Black'; for the gender Pr and Re equalled both 0.74 for 'Male', and 0.69 and 0.71 for 'Female'; for the backpack accessory Pr and Re were equal to 0.48 and 0.44.

The people search module was then tested in the same **practical demonstrations** as the person-re-identification module mentioned in Sect. 5.1.2, using the same 'suspect individuals' (volunteers). The results in terms of **retrieval accuracy**, i.e., the first rank at which template images of individuals exhibiting the query attribute profile appeared, were altogether similar to the ones of the person-re-identification module. Also in this case the LEA officers involved in the test felt that the number of false detections was too high. The general feeling is that a higher retrieval accuracy would be required by LEAs to a people search tool in real mass gathering events, in terms of a better attribute detection accuracy. To this aim the main direction for improvement is analogous to one of the directions discussed in Sect. 5.1.2 for the person re-identification functionality, i.e., finding more discriminant features for pedestrian attribute detection; deep learning techniques, which are potentially capable to act as automatic feature extractors, appear currently as a promising solution. Moreover, the availability of a large and representative training set for each attribute of

<sup>53</sup> <https://scikit-learn.org/>

<sup>54</sup> <https://www.tensorflow.org/>

interest is necessary: existing data sets collected by the computer vision research community (see Sect. 2.4) are still too limited for real-world applications.

As regards the **camera setting**, the same considerations made in Sect. 5.1.2 for the person re-identification module apply also to the people search module.

## 5.3 CROWD MONITORING MODULE

### 5.3.1 Implementation details

The same programming languages, libraries and frameworks as the people search module have been used (see Sect. 5.2.1). With reference to the functional architecture shown in Figure 7Figure 6, the only component already fully available as open source software was the machine learning software used to train the regression models. All the other components have been specifically developed for the LETSCROWD project.

### 5.3.2 Validation

Validation of this module during **design phase** has been carried out on three benchmark data sets for crowd density estimation among the ones listed in Sect. 2.4: UCSD Pedestrian Traffic, QMUL Mall and PETS 2009. The results in terms of the Mean Deviation Error (MDE, see Sect. 2.2.3), across the different linear and non-linear regression models considered and different combinations of the considered features (see Sect. 4.3.3) can be summarized as follows:

- on QMUL Mall the MDE ranged from 0.10 to 0.29;
- on UCSD Pedestrian Traffic the MDE ranged from 0.10 to 0.35;
- on PETS 2009 an excessive MDE was obtained, instead, ranging from 0.30 to more than 1.0.

It was also observed that the use of perspective correction did not provide sensible improvements; on the contrary, in a few cases it considerably worsened the accuracy.

Notably, on **cross-data set** experiments (i.e., training on one data set and testing on a different one) the results considerably worsened, as expected. As discussed in previous sections, this points out that this kind of computer vision task is still very challenging, and existing algorithms can be expected to perform reasonably well only if the training images are representative of the ones processed during operation. This confirms that using synthetic videos of crowds obtained from crowd modelling and computer graphics tools is an interesting solution to investigate to attempt reducing the gap between currently achievable performance and performance requirements of real application scenarios.

The crowd density estimation module was then tested in the same **practical demonstrations** mentioned in Sect. 5.1.2, as well as in another practical demonstration organized by the Ertzaintza (ERT) partner LEA during a real mass gathering event, the MTV European Music Awards 2018 in Bilbao (Spain, Nov. 3, 2018). In Fürstentfeldbruck several scenes with a relatively small number of people (about 70) were acquired using fixed cameras. In OostMalle two scenes with a few hundreds people were acquired by two PTZ cameras when the public left the event venue at the end of the cyclo-cross race. Similarly, in Bilbao two scenes with several hundred people were acquired in a large square outside the event venue (a football stadium), when the public was leaving the stadium at the end of the event: one from a PTZ camera and one from a RPAS. In these tests the exact number of people shown in the videos, which is needed to evaluate crowd density estimation accuracy, was easy to compute only for the simulations in Fürstentfeldbruck, whereas it could only be roughly estimated in the Bilbao and OostMalle real mass gathering events. Similarly to the validation experiments on benchmark data sets, these tests have been carried out using different regression models and different combinations of the considered features. As expected, with some exceptions the observed accuracy was generally rather low, given that the scenes in the videos acquired during the three practical demonstrations are very different from the ones in the benchmark data sets used to train the regression algorithms; indeed, in this respect these tests correspond to cross-data set experiments.

As for the person re-identification and people search tools, the general feeling of LEAs from the practical demonstrations was that the accuracy of the crowd density estimation module is still not sufficient for real-world mass gathering events. In this case the representativeness of the image data set used to train the underlying regression model is clearly the key issue to guarantee a good accuracy. In real-world application scenarios where it is not possible to collect beforehand *real* videos representative of the scenes that will be processed during operation, a very promising solution is the one envisaged in the LETSCROWD project and discussed in previous sections, i.e., using *synthetic* videos generated by computer graphics and crowd modelling tools. This solution is currently being investigated by ongoing experiments, using synthetic videos generated by the CMP tool of the Crowd Dynamics LETSCROWD partner; such videos are being generated by reproducing the same camera views, crowd size and crowd configurations as in the original videos acquired during the practical demonstrations. These experiments are being carried out at the time of submission of this deliverable, and their results will provide insights on the improvements achievable using this kind of solution.

With regard to **camera setting**, according to system requirements (Sect. 3.3) fixed and PTZ cameras have been considered. The suggested angle with the horizontal plane is of about 45 degrees or more, and the height should be about 5 m or more. For crowd analysis at the macroscopic level relatively far views are preferred. For reasons similar to the ones discussed in Sect. 5.1.2, more specific indications have not been included in system requirements. We point out that crowd density estimation algorithms may need a perspective map. For fixed cameras the perspective map can be defined beforehand, or off-line. For PTZ cameras real-time information of the current view is needed, instead, either 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 had also been considered in the system requirements as a potential additional video source, since its regulation at the EU level was still in progress, although RPASs are already allowed in some countries as indicated by LEAs in the user requirements (see Sect. 3.1). In fact, one video coming from a RPAS was provided by ERT during the Bilbao practical demonstration.

## 6 CONCLUSIONS

The human-centred computer vision tool is a software prototype which provides functionalities aimed at supporting LEA operators and forensic investigators in two kinds of tasks commonly carried out using video surveillance systems in mass gathering events: (i) estimating in real time the size of a crowd and detecting related anomalous events such as overcrowding, and (ii) searching for individuals of interest in recorded videos, during forensic investigations, starting either from an image or video of the individual of interest or from a description provided by an eyewitness.

The above tasks are still very challenging for state-of-the-art computer vision algorithms in real-world application scenarios such as the ones considered in LETSCROWD, related to mass gathering events. In fact, results observed in the practical demonstrations that took place before April 2019 showed that the accuracy achieved by the HCV tool would not be satisfactory yet for LEAs in real mass gathering events. As a distinctive feature with respect to the state of the art, the HCV tool includes two potential solutions that are currently being investigated to improve the accuracy of the underlying computer vision algorithms in this kind of application scenario: exploiting user's feedback in the person re-identification and people search modules to re-rank the retrieved images of pedestrians, such that images more similar to the query image or target attribute profile are "pushed" toward the top ranks; and exploiting synthetic videos of a crowd generated by the CMP tool to train the crowd density estimation algorithm using images similar to the ones that will be processed during operation.

For demonstration purposes the HCV tool has been implemented as a client-server application with a Web-based graphical user interface, using open-source programming languages and software libraries. In a real deployment scenario the functionalities of the HCV tool can be provided as additional functionalities of



commercial or ad hoc video analytics software suites, such as the ones that are already used by LEAs and other organizations to manage video surveillance systems.

## 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. *RGB-D Based Multi-attribute People Search in Intelligent Visual Surveillance*. **Liu, Wu, et al.** s.l. : Springer, 2012. MMM 2012. Vol. LNCS Vol. 7131, pp. 750-760.
18. *Attribute-based People Search: Lessons Learnt from a Practical Surveillance System*. **Feris, R., et al.** s.l. : ACM, 2014. ICMR '14.

19. *Searching for people using semantic soft biometric descriptions*. **Denman, S., et al.** 2015, Pattern Recognition Letters, Vol. 68, pp. 306-315.
20. *People search based on attributes description provided by an eyewitness for video surveillance applications*. **Frikha, M., Fendri, E. and Hammami, M.** 2019, Multimed Tools and Applications, Vol. 78, pp. 2045-2072.
21. *Describing objects by their attributes*. **Farhadi, A., et al.** : IEEE, 2009. Proc. Int. Conf. on Computer Vision and Pattern Recognition. pp. 1778-1785.
22. **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.
23. *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.
24. *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.
25. *Person Re-identification by Deep Learning Attribute-Complementary Information*. **Schumann, A. and Stiefelhausen, R.** : IEEE, 2017. Proc. Int. Conf. on Computer Vision and Pattern Recognition Workshops. pp. 1435-1443.
26. **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.
27. **Su, C., et al.** Multi-type attributes driven multi-camera person re-identification. *Pattern Recognition*. 2018, Vol. 75, pp. 77-89.
28. **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.
29. **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.
30. **Zhan, B., Monekosso, D.N. and Remagnino, P. et al.** Crowd analysis: a survey. *Machine Vision and Applications*. 2008, Vol. 19, pp. 345-357.
31. **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.
32. **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.
33. **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.
34. **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.
35. **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.
36. *Fully Convolutional Network for Crowd Size Estimation by Density Map and Counting Regression*. **Wu, B. and Lin, C. s.l.** : IEEE, 2018. IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 2170-2175.
37. *A survey of recent advances in CNN-based single image crowd counting and density estimation*. **Sindagi, V.A. and Patel, V.M.** 2018, Pattern Recognition Letters, Vol. 107, pp. 3-16.
38. *Granular-based dense crowd density estimation*. **Kok, V.J. and Chan, C.S.** 15, 2018, Multimedia Tools and Applications, Vol. 77, pp. 20227-20246.
39. *A texture based manifold approach for crowd density estimation using Gaussian Markov Random*

- Field*. **Lamba, S. and Nain, N.** 5, 2019, Multimedia Tools and Applications, Vol. 78, pp. 5645-5664.
40. **Isard, M. and Blake, A.** CONDENSATION conditional density propagation for visual tracking. *Int. Journal of Computer Vision*. 1998, Vol. 29, 1, pp. 5-28.
41. **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.
42. **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.
43. **Helbing, D. and Molnár, P.** Social force model for pedestrian dynamics. *Physics Review E*. 1995, Vol. 51, 5, pp. 4282-4286.
44. **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.
45. **Wu, S., et al.** Crowd Behavior Analysis via Curl and Divergence of Motion Trajectories. *Int. J. Computer Vision*. 2017, Vol. 123, pp. 499-519.
46. **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.
47. *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.
48. **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.
49. *Analysis-by-synthesis: Pedestrian tracking with crowd simulation models in a multi-camera video network*. **Jin, Z. and Bhanu, B.** 2015, Computer Vision and Image Understanding, Vol. 134, pp. 48-63.
50. *LCrowdV: Generating labeled videos for pedestrian detectors training and crowd behavior learning*. **Cheunga, E., et al.** 2019, Neurocomputing, Vol. 337, pp. 1-14.
51. *Person Re-identification in the Wild*. **Zheng, L., et al.** : IEEE, 2017. Proc. Int. Conf. on Computer Vision and Pattern Recognition. pp. 3346-3355.
52. **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.
53. **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.
54. **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.
55. **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.
56. **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.
57. *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.
58. **Hand, D.J.** Classifier Technology and the Illusion of Progress. *Statistical Science*. 2006, Vol. 21, 1, pp. 1-14.
59. **Deng, Y., et al.** Pedestrian Attribute Recognition At Far Distance. *ACM Multimedia*. 2014, pp.



789-792.

60. **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.
61. **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.
62. **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.
63. **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.
64. *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.
65. *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.
66. *Scalable person re-identification: A benchmark.* **Zheng, L., et al.** s.l. : IEEE, 2015. Proceedings of the IEEE International Conference on Computer Vision. pp. 1116-1124.
67. *Pedestrian Attribute Classification in Surveillance: Database and Evaluation.* **Zhu, J., et al.** : IEEE, 2013. Proc. Int. Conf. on Computer Vision Workshops. pp. 331-338.
68. **Li, D., et al.** A Richly Annotated Dataset for Pedestrian Attribute Recognition. *CoRR*. 2016, Vol. abs/1603.07054.
69. *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.
70. *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.
71. *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.
72. *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.
73. *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.
74. *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.
75. *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.
76. *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.
77. **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.
78. *Optical Flow Dataset and Benchmark for Visual Crowd Analysis.* **Schröder, G., et al.** s.l. : IEEE, 2018. Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1-6.
79. *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.
80. *Ten Years of Relevance Score for Content Based Image Retrieval.* **Putzu, L., Piras, L. and**



**Giacinto, G.** [ed.] P. Perner. s.l. : Springer, 2018. Machine Learning and Data Mining in Pattern Recognition. Vol. LNCS Vol. 10935, pp. 117-131.

81. **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.

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

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

84. *Searching for people through textual and visual attributes*. **Fabian, J., Pires, R. and Rocha, A.** 2012. 25th SIBGRAPI Conference on Graphics, Patterns and Images. pp. 276-282.

85. *Describable Visual Attributes for Face Verification and Image Search*. **Kumar, N., et al.** 10, s.l. : IEEE, 2011, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, pp. 1962-1977.

## 7.2 ACRONYMS

### Acronyms List

CCTV	Closed-circuit television
CMP	Crowd modelling and planning
CNN	Convolutional neural network
DBMS	Database management system
DoW	Description of work
DRA	Dynamic risk assessment
FP	False positive
fps	Frames per second
GLCM	Grey-level co-occurrence matrix
GUI	Graphical user interface
HCV	Human-centred computer vision
HITL	Human-in-the-loop
HOG	Histogram of oriented gradients
LBP	Local binary patterns
LEA	Law Enforcement Agency
Pr	Precision
Re	Recall
RPAS	Remotely piloted aircraft system
SVM	Support vector machine
TP	True positive

**TABLE 4 – Acronyms**

## 8 ANNEX A – HCV TOOL USER GUIDE

The HCV tool is accessed through a Web browser. The **starting window** (Web page) is shown in Figure 20. The links at the top of the window give access to the three functionalities of person re-identification, people search and crowd density estimation. The button 'Clean Database' allows the user to remove the videos previously uploaded in each of the three modules, together with the pedestrian images automatically extracted from them and the related information (see Sect. 4.1), stored in the *template gallery* database.

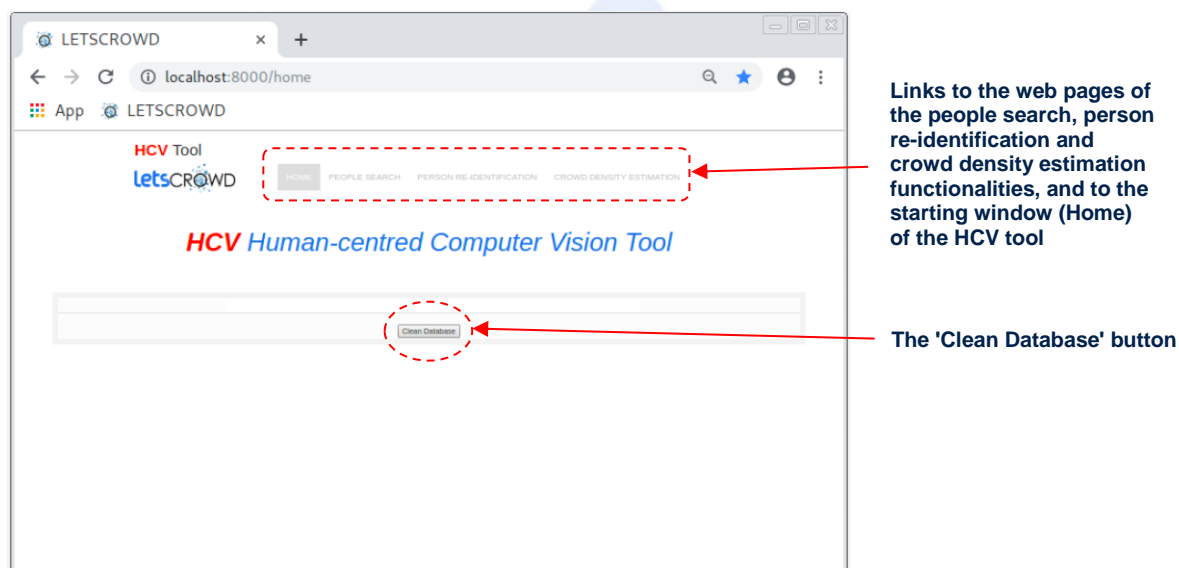
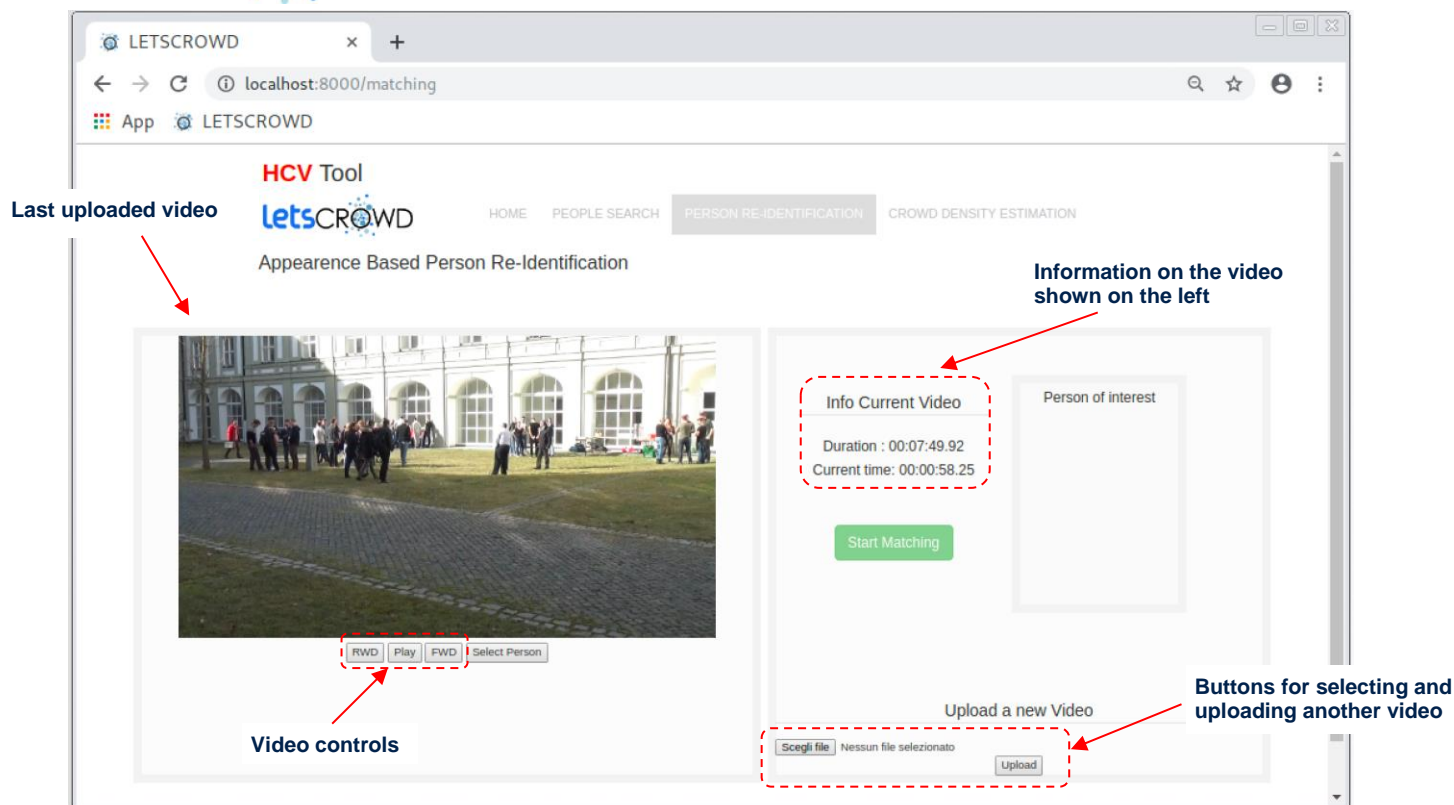


Figure 20 – Starting window (Web page) of the HCV tool.

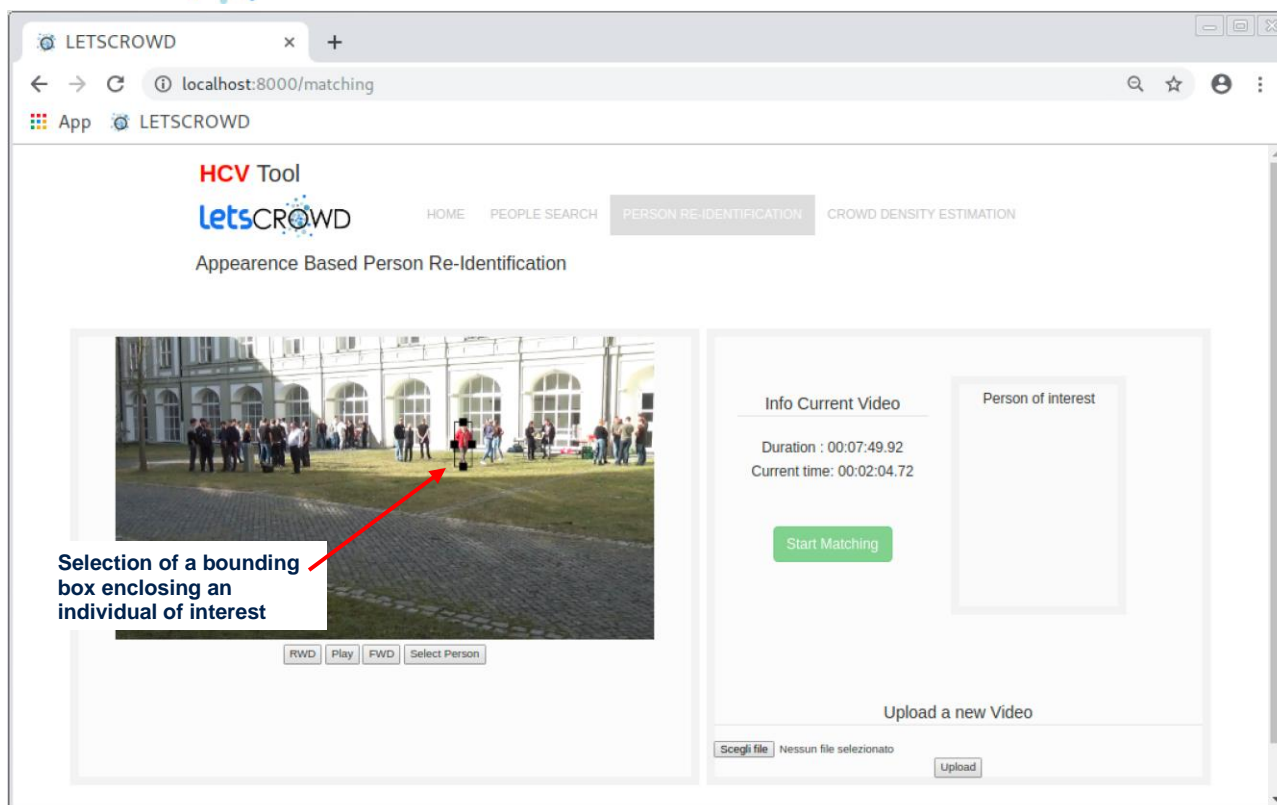
### 8.1 PERSON RE-IDENTIFICATION FUNCTIONALITY

The **Main window** of the person re-identification module is shown in Figure 21. It shows on the left the last uploaded video and below it the button to play and stop it, and to move frame-by-frame forward and backward. On the right the information about the same video is shown (duration and current time). The buttons on the bottom-right allow the user to select and upload another video.



**Figure 21 – Main window of the person re-identification functionality.**

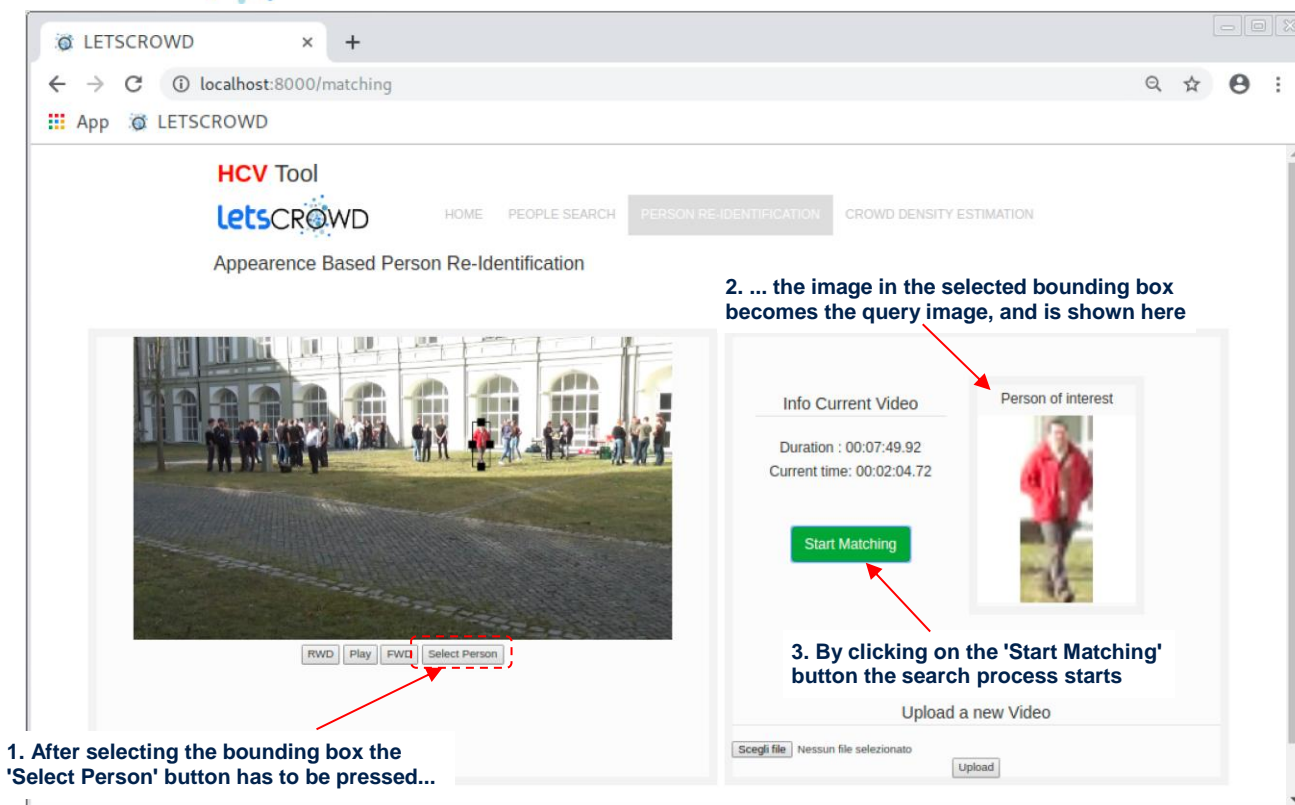
To select an image (bounding box) of an individual of interest who appears in the video, the user has first to stop the video and to draw, using the mouse, a rectangular bounding box containing that individual, as shown in Figure 22. If necessary, the bounding box can be resized by pointing, clicking and dragging the mouse on any of the black squares shown on each of its sides.



**Figure 22 – Selection of a bounding box containing the image of an individual of interest.**

Then, by clicking on the button 'Select person' below the video the current bounding box becomes the query image, which is shown on the right of the main window (see Figure 23). Finally the retrieval operation can be started by clicking on the button 'Start matching' shown on the left of the query image.





**Figure 23 – Choice of the *query* image of the individual of interest, and start of the retrieval operation.**

The query image is automatically matched to all the *template* images in the *template gallery* database, i.e., the pedestrian images automatically extracted by the person re-identification module from all the uploaded videos. These images are then shown to the user in a **Results window** which replaces the main window, and is shown in Figure 24. In the results window the query image is shown on the top-right. The retrieved template images are shown in the main area of the window, ranked for decreasing similarity to the query image, from top to bottom and from left to right. Context information is shown above each template image: timestamp and name of the video file from which it was extracted. The results window can be scrolled by the user to see all the retrieved template images.

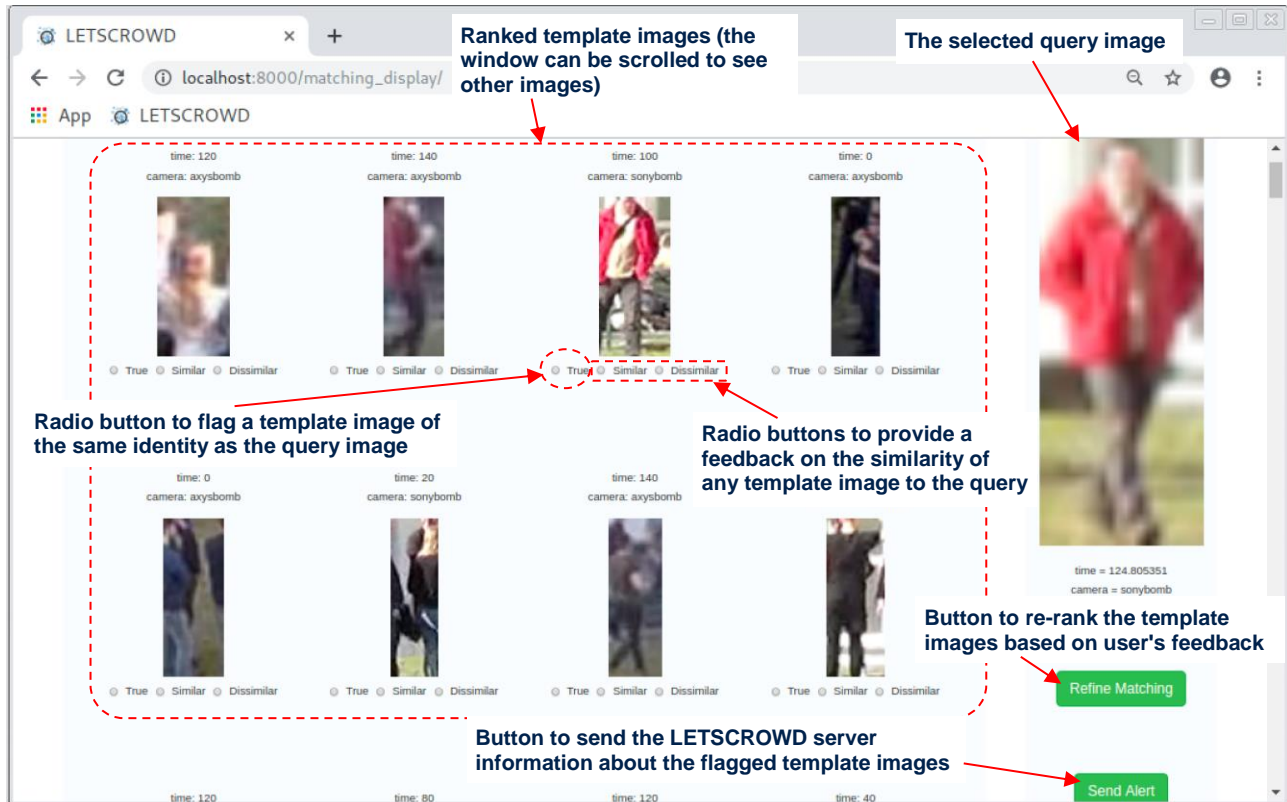


Figure 24 – Results window of the person re-identification module.

Below each retrieved template image three radio buttons are shown: 'True', 'Similar' and 'Dissimilar'.

- The 'True' button (where 'true' stands for 'true match') allows the user to flag a template image as showing the same identity of the query image. By selecting this radio button the information about the corresponding image is temporarily stored by the person re-identification module. In a later moment, stored information about all template images that have been flagged as 'true matches' can be sent to the LETSCROWD server, together with the analogous information about the query image, by clicking on the button 'Send alert' which is shown below the query image; this button opens a text field where the user can input an optional text to describe the alert generated.
- When the number of template images in the template gallery database is very large, the user may be willing to inspect only the top-ranked ones in the results list. If the user does not find any template image of the same identity as the query image among the inspected ones, or if he/she would like to search for further images of that identity, the search can be refined as follows. First the user has to provide a feedback on the similarity or dissimilarity of one or more template images to the query image, through the radio buttons 'Similar' and 'Dissimilar'; then he/she has to click on the button 'Refine matching'. The template images are automatically re-ranked based on the above feedback, and are shown in the same results window.

The user can also further analyse any template image of interest by clicking on it, which opens the **Template Detail** window shown in Figure 25. This window shows the query image on the right, the selected template image on the top-left, together with the same contextual information shown in the Results window (see above), and the whole video frame from which the template image was extracted. This video can be played back by using the controls shown below it. This allows the user to analyse the individual in the template image (e.g., to verify that the identity is the same as the one of the query image), and the behaviour of that individual (e.g., to get additional insights about whether the exhibited behaviour can be considered as suspect). By closing the Template Detail window the results window is shown again.

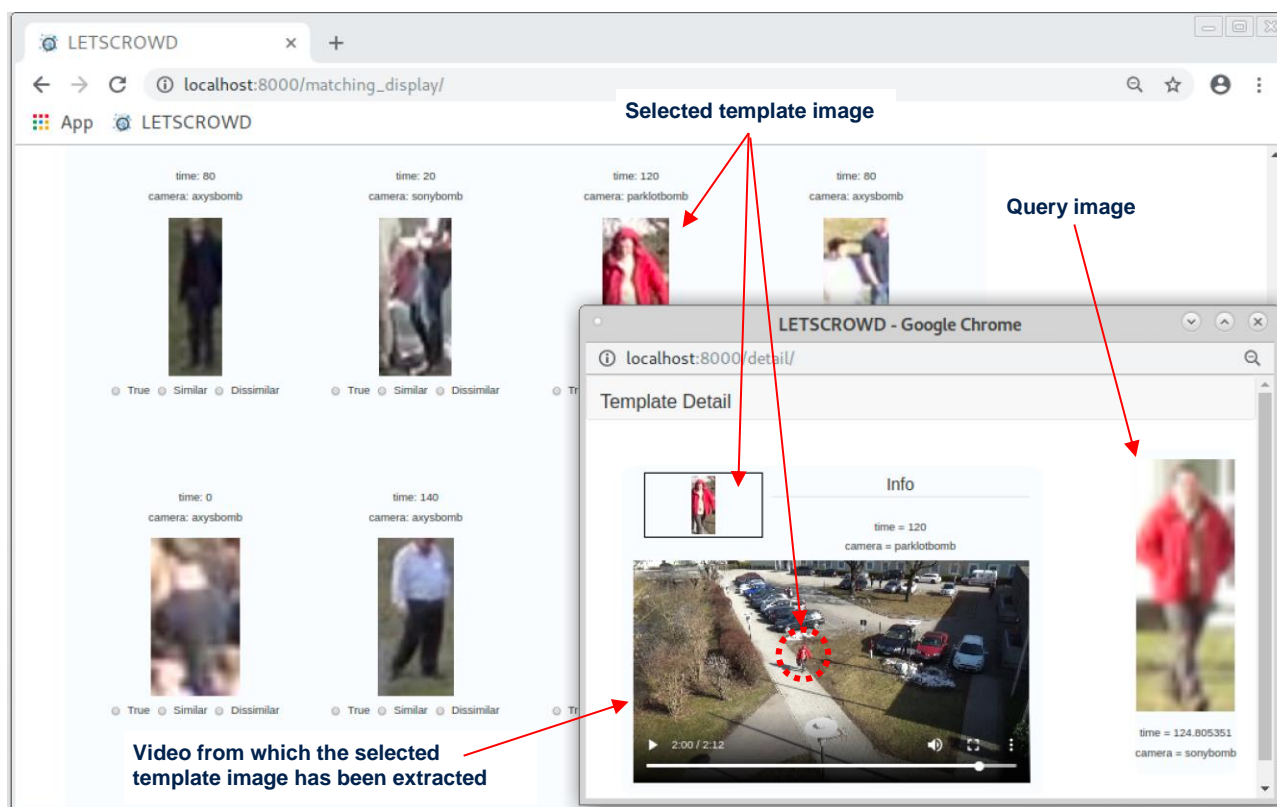


Figure 25 – Template Detail window of the person re-identification module.

To return to the Main window of the person re-identification module the user has to click on the corresponding link at the top of the Results window.

## 8.2 PEOPLE SEARCH MODULE

The **Main** window is shown in Figure 26. It shows a predefined set of attributes that the user can choose to input the *target attribute profile*, i.e., the description of the appearance of the individual of interest to be searched among the same template gallery images as the person re-identification module. The attributes are grouped into the following categories: upper body clothing colour, upper body clothing style, type of sleeve, lower body clothing colour, lower body clothing style, pants type, type of shoes, gender, accessories (hat and backpack). The available values for each attribute are shown either as checkboxes or as radio buttons; checkboxes are used for non-mutually exclusive attributes like clothing colours (e.g., upper body clothing can exhibit more than one colour), whereas radio buttons are used for mutually exclusive attributes such as gender. By default no checkbox and no radio button is selected. After the target attribute profile is input by selecting at least one checkbox or radio button, the retrieval process can be started by clicking the 'Start Matching' button.

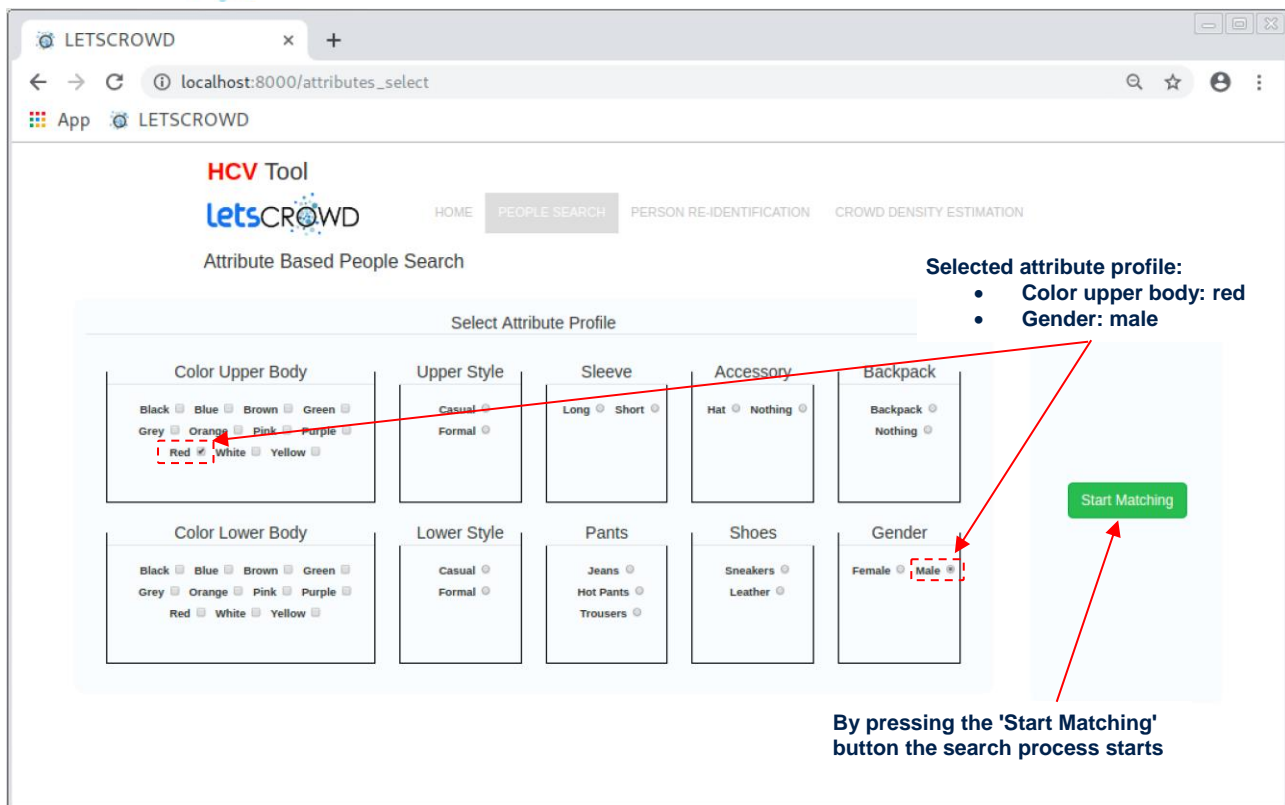


Figure 26 – Main window of the people search module.

The retrieved template images are shown in the **Results window**, which is similar to the homonymous window of the person re-identification module. The Results window is shown in Figure 27. In this case the template images are ranked for decreasing similarity to the attribute profile selected in the Main window. The same options as for the person re-identification module are available to the user:

- Providing a feedback ('similar'/'dissimilar') on any subset of template images through the corresponding radio buttons, and then re-ranking such images accordingly by pressing the 'Refine Matching' button on the top-right.
- Flagging any subset of template images as exhibiting the target attribute profile through the radio button 'True', and then sending to the LETSCROWD server the corresponding information, together with the target attribute profile, through the 'Send Alert' button shown on the top-right.
- Inspecting any template image by clicking on it, which opens the **Template Detail window** that is in turn similar to the homonymous window of the person re-identification module, and provides the same options to the user. The Template Detail window is shown in Figure 28.



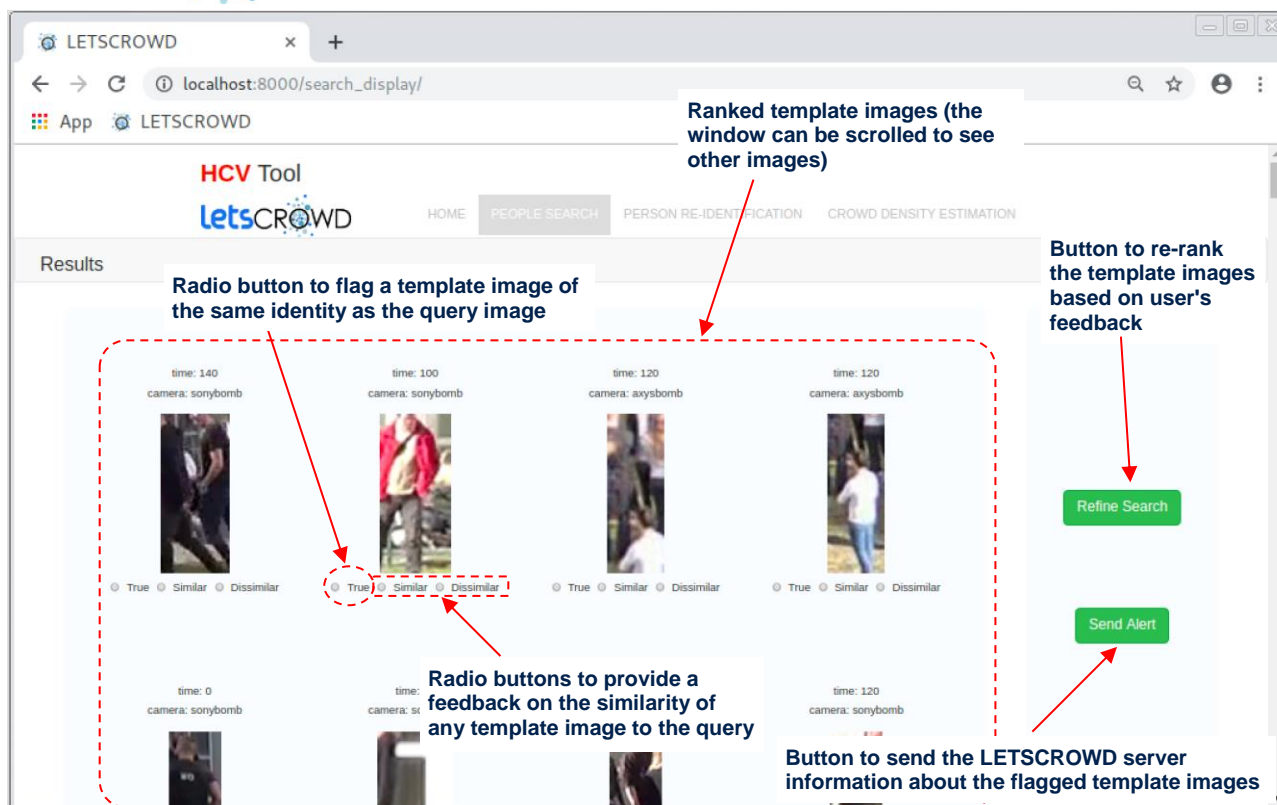


Figure 27 – Results window of the people search module.

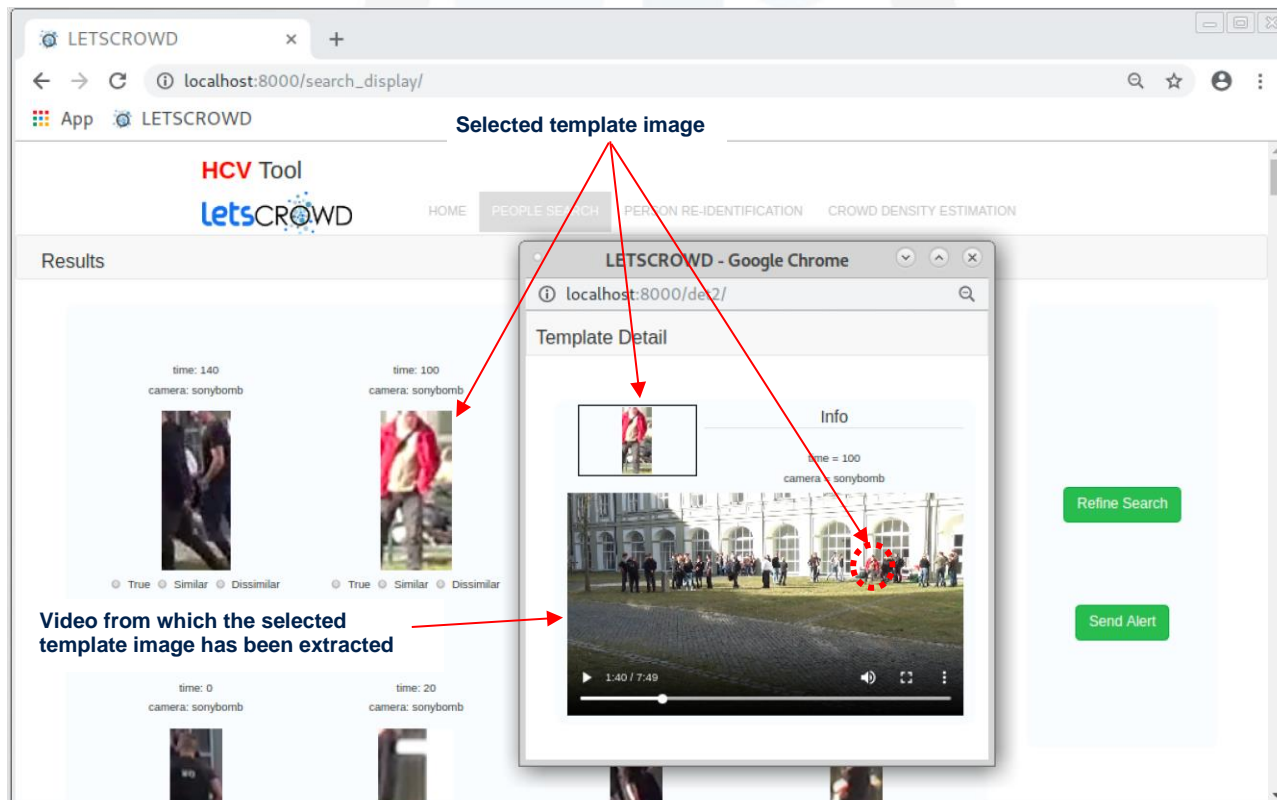


Figure 28 – Template Detail window of the people search module.

### 8.3 CROWD MONITORING MODULE

The crowd monitoring module is accessible through the **Main window** shown in Figure 29. The user can upload a video using the buttons in the bottom-left. The video is shown on the left part of the Main window. Below the video there are three buttons which allow the user to play and stop it, and to move frame-by-frame forward and backward. Below such buttons the duration of the video and the elapsed time since its beginning are shown. In the right part of the Main window the same video is shown, with a superimposed semi-transparent heat map which represents the estimated relative crowd density in different areas of the scene, using colours from blue (low density) to red (high density). The estimated number of people is shown below the heat map, and is updated in real time, frame by frame. The user can also set a threshold on the estimated number of people above which an alert will be automatically generated and shown on the screen: this can be done by entering the threshold value (any positive integer number) in the text field below the heat map. The default threshold value is 0, which means that alerts are disabled. After any threshold has been set, alerts can be disabled by the user by entering a value of 0 in the same text field. Alerts are also automatically generated when a sudden increase or decrease of the estimated number of people is detected.

Alerts are shown to the user in pop-up windows. The user can confirm an alert by clicking on the 'Send alert' button in the same window: in this case the alert is sent to the LETSCROWD server as a description of the anomalous event, and the following contextual information: the name of the video file (or camera ID), the timestamp, a snapshot of the corresponding video frame, the corresponding estimated number of people and, if the alert refers to exceeding the user-defined threshold, the value of such a threshold.

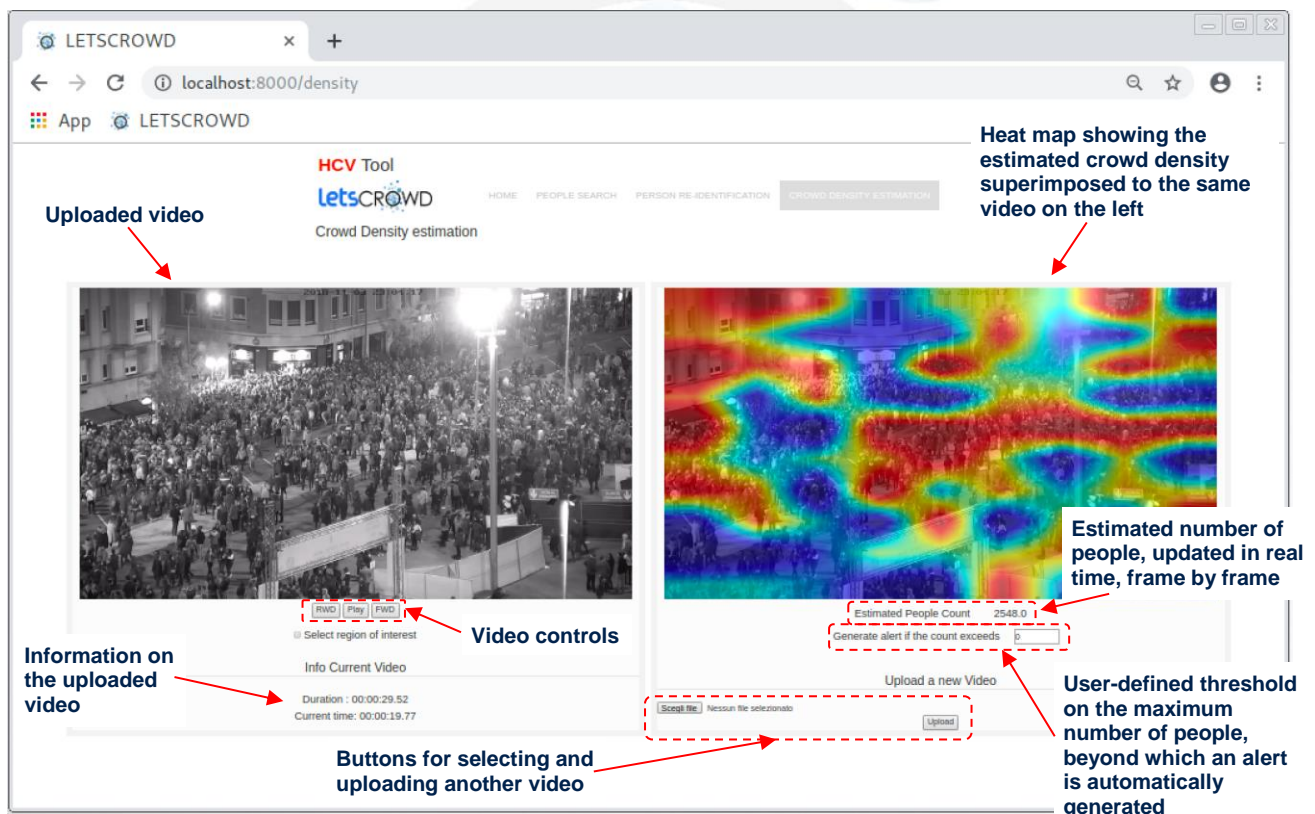


Figure 29 – Main window of the crowd density estimation module.

The user can also choose to monitor only a rectangular sub-region of the whole scene, named "region of interest" (ROI). To this aim the checkbox 'Select region of interest' below the video on the left has to be selected first; then the ROI can be selected by clicking with the mouse on any point in the video, which is considered as one the vertices of the ROI, and finally dragging the mouse and releasing it on the opposite vertex. The selected ROI is highlighted in red in the video on the left, as shown in Figure 30. From that

moment onward the estimated crowd density, as well as the user-defined threshold on the number of people (if any) and the alerts automatically generated, refer to the selected ROI. The default ROI corresponds to the whole scene. Any selected ROI can be reset to the whole scene by deselecting the above checkbox. A given ROI can also be changed by deselecting and re-selecting the checkbox 'Select region of interest', and then drawing the new ROI.

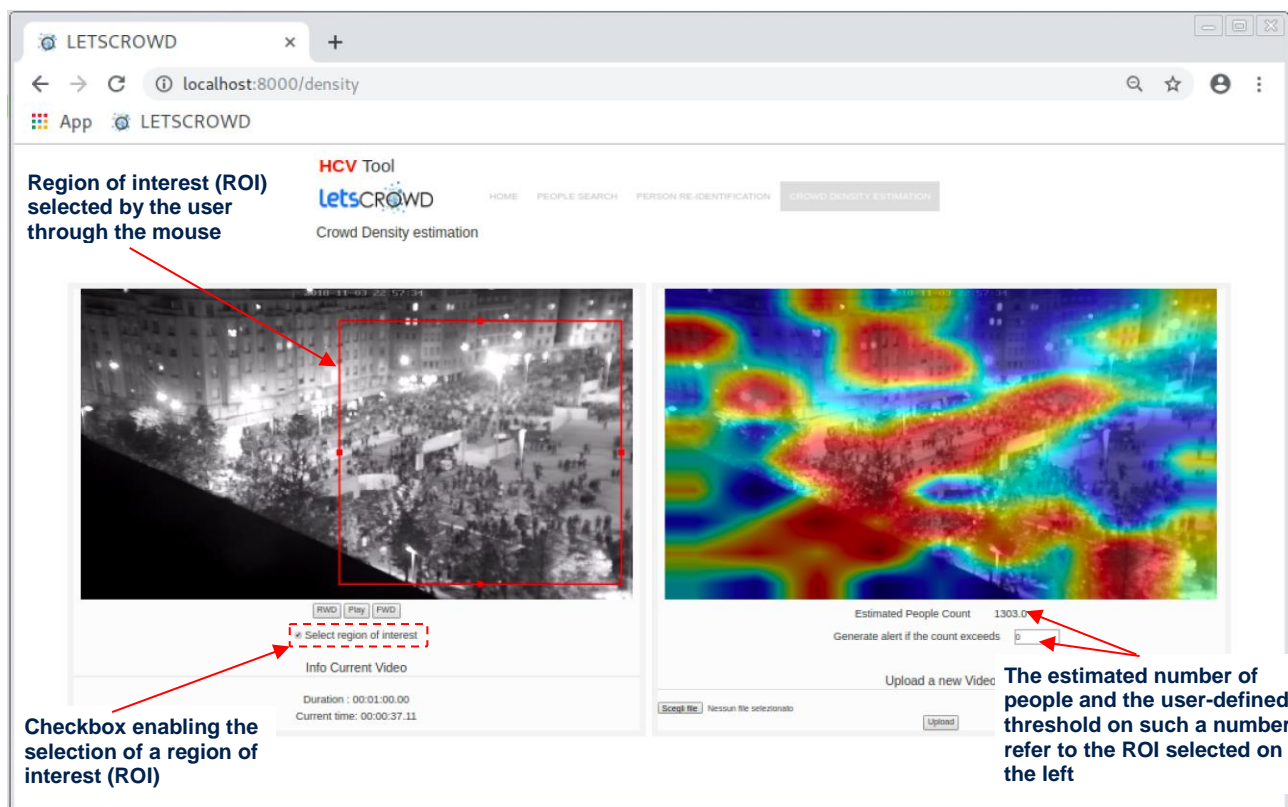


Figure 30 – Selection of a region of interest in the crowd density estimation module.