A framework for large-scale analysis of video “in the Wild” to assist digital forensic examination

Apostolos Axenopoulos, *Member, IEEE*, Volker Eiselein, Antonio Penta, Eugenia Koblents, Ernesto La Mattina and Petros Daras, *Senior Member, IEEE*

Abstract—Digital forensics departments usually have to analyse vast amounts of audio-visual content, such as videos collected from street CCTV footage, hard drives or online resources. The framework presented in this article, which has been developed in the context of the EU-funded project LASIE, aims to assist investigators in their everyday tasks, through the provision of innovative tools for image and video analysis, object detection and tracking and event detection. These tools exploit the latest advances in machine learning, including deep neural networks, to handle the challenges in processing content from real-world data sources. The framework is enhanced with advanced inference and recommendation capabilities, which filter-out inconsistencies and recommend additional evidence. An intuitive user interface allows exploiting the capabilities of the available tools in a user-friendly manner. The framework supports distributed processing, with easy deployment of the services in clusters of multiple workstations, making the proposed solution appropriate for big data analytics tasks.

Index Terms—video analytics; deep machine learning; digital forensics; big data.

1 Introduction

Law Enforcement Agencies (LEAs) need to identify, ingest, process, and analyse vast volumes of heterogeneous multimedia content in order to obtain valuable information and insightful knowledge that can allow them to rapidly react to a crime. However, the explosion of digital content coming from all kinds of portable digital devices and online activity cannot be covered by current methodologies, techniques and tools. As a result, when it comes to the investigation of a particular crime, the workload and the requirements of the law enforcement investigators and analysts are becoming unbearable. Analysts are often required to search in extremely large collections of forensic data, e.g. images and video, for evidence that could be accepted when presented in a court of law. This data analysis process typically takes a very long time, extremely high and costly human resources, if it is not automated or computer assisted. Thus, a challenge for LEAs is to keep their toolboxes up to date and be able to rapidly adapt and extend them, otherwise, the impact would be dramatic (e.g. important evidence missed, long investigation times).

To address the above problems and automate several parts of the investigation process, an extensive amount of research has been undertaken over the latest years. Specifically for image and video analysis, new technologies have been introduced exploiting the recent advances in computer vision and machine learning. These approaches achieve high levels of accuracy when tested in publicly available datasets that have been setup for research purposes. The datasets mainly consist of images and videos acquired in controlled environments (e.g. calibrated cameras, same illumination etc.). More comprehensive solutions incorporate several state-of-the-art computer vision algorithms for object detection and tracking to offer complete pipelines for large-scale processing, re-identification, search and retrieval [15]. However, concerning processing of video ”in the Wild”, such as footage from real CCTV street scenes, most of the methods fail. The need for video analytics tools that are robust in unconstrained survaillance enviroments is visible.

In this paper, a framework is proposed that offers a set of tools to manipulate, analyse and fuse heterogeneous forensic data acquired from different sources and in multiple formats. Then, through an iterative, interactive and highly efficient process the analyst will be able to browse and search the forensic data, using a user-friendly interface. The framework is equipped with appropriate knowledge structures that allow the system to provide recommendations to the analyst, guide the investigation process and perform inference based on evidence extracted from available data.

The framework has been implemented in the context of the EU-funded project LASIE[1]. The proposed technologies have been tested with promising results on a very challenging datastet consisting of real CCTV footage, which was made available by the London Metropolitan Police (MET). The majority of videos are of low resolution, low colour quality, from Pan-Tilt-Zoom (PTZ) cameras that move fast and blur the subjects, cameras that observe from far away

- *A. Axenopoulos and P. Daras are with the Information Technologies Institute, CERTH, 57001, Thessaloniki, Greece.*
E-mail: {axenop;daras}@iti.gr
- *V. Eiselein is with Technische Universität Berlin, Communication Systems Group, Einsteinufer 17, 10587 Berlin, Germany*
E-mail: eiselein@nue.tu-berlin.de
- *A. Penta is with the United Technologies Research Center, Cork, Ireland.*
E-mail: pentaa@utrc.utc.com
- *E. Koblents is with Universidad Politcnica de Madrid, Ramiro de Maeztu 7, Madrid-Spain.*
E-mail: ekl@gatv.ssr.upm.es
- *E. La Mattina is with Engineering Ingegneria Informatica spa, Palermo, Italy.*
E-mail: ernesto.lamattina@eng.it

[1] http://www.lasie-project.eu/

as well as cameras the are highly affected by light variations from day to night. The accuracy in detection of objects and events of interest, in search and retrieval of relevant items, as well as the efficiency in processing large volumes of data demonstrate that our solution can become a valuable tool for Law Enfocement Agencies that would speed-up their everyday work.

The paper is organised as follows: Section 2 provides an overview of the system and highlights the main contributions. Section 3 describes the core video analytics modules, while the proposed Evidence Search Engine is analysed in Section 4. The User Interface is presented in Section 6 and the methodology for large-scale analysis in Section 5. Finally, conclusions are drawn in Section 8.

2 Overview and Main Contributions

The workflow of the proposed framework is depicted in Figure 1. The system operates in two distinct phases, namely the offline and the online phase.

During the offline phase, the raw video data gathered from multiple sources (e.g. CCTV footage) are inserted to the system (ingestion service), segmented into clips of predefined length and given as input to the processing modules. The latter are running in parallel to fulfill different video analytics tasks: detection of faces, people, logos and vehicles, tracking of people and vehicles, extraction of soft biometric features, video summarisation and event detection. Distributed processing is followed by the creation of two basic structures that will facilitate search and retrieval of evidence at the online phase: the Visual Index and the Evidence Graph. The former takes as input the low-level features that are extracted automatically for each of the detected objects (humans, vehicles, logos) and generates an indexing scheme to enable rapid comparison between query and stored items. The Evidence Graph (EG) identifies high-level relationships among detected items, such as items that coexist in the same frame/scene or items with visual similarity and maintains a structure with all possible connections.

During the online phase (search), the Complex Query Formulator offers multiple search capabilities, such as query-by-example image/video, text-based video retrieval, or even fetch items (humans/objects) that coexist with the query in the same scene (e.g. using input from the EG). Additionally, the Inference and Recommendation modules exploit the knowledge accumulated in the EG to recommend further search directions, remove inconsistencies and produce more meaningful results.

The framework introduces the following innovative features:

High-accuracy object detection and tracking even in challenging environments (CCTV videos “in the wild”), exploiting the power of Deep Learning (Convolutional Neural Networks).

High-level detection of violent events, introducing novel video-level descriptors able to encode salient spatiotemporal motion patterns in video.

New capabilities for formulating complex queries, identifying new high-level relationships among detected items by traversing the EG.

Advanced inference and recommendation capabilities, which are able to filter-out inconsistencies and propose investigation directions through high-level reasoning on the EG.

A new approach for distributed processing of large volumes of videos, achieving optimal use of computational resources, thus, making the framework available to deal with big data analysis.

3 Forensic Data Processing

3.1 Object Detection

Given the vast amounts of footage available in forensic applications, object detection is of paramount importance in surveillance. Recent breakthrough in Deep Learning (DL) methodologies offers a promising new frontier for object detection. Convolutional Neural Networks (CNNs) achieve high quality object localization and classification in static images. Faster R-CNN [1] has been selected as a starting point for the object detection module due to its end-to-end trainable architecture and its performance both time and accuracy-wise. It is able to generate high quality region proposals and classify them detecting numerous objects according to previous training material. The Region Proposal Network (RPN) shares convolutional features and forms a unified network, exploiting the power of GPU computing. In our implementation [3], we utilise the latest state of art object detection model, the ResNet [2] with a 101-layer architecture pre-trained in the ImageNet dataset and fine-tuned using annotated examples from the LASIE dataset to better address the challenges of real world CCTV footage.

A second approach, much faster but less accurate, has been also integrated in the platform for very large scale analysis with low computational resources. This approach is appropriate for analyzing vast amounts of video data with a moderate hardware cost (i.e., low number of workstations without high-end GPUs), making use of the person detector algorithm proposed by Dollar *et al.* in [5]. This is based on a combination of an efficient extraction of highly discriminative features and a fast AdaBoost classifier which quickly examines only a few discriminative features in order to determine the presence of a person. With a single CPU core, this approach provides a throughput of 25 frames per second, more than two orders of magnitude higher than the one obtained with DL-based approaches such as the first one, if we use the same hardware (CPU only). This can be further sped up with GPU and CPU parallelization.

The aforementioned approaches have been employed for detecting humans, faces and vehicles. Another modality that could assist in identifying suspects within video footage is logo detection. Within the LASIE framework, it is used as enabling technology to identify the brand of a piece of clothing that an individual is wearing. In order to tackle the problems inherited from real life CCTV footage, LASIE deployed a CNN architecture for semantic pixel-wise classification, namely SegNet [14]. The proposed solution is a one step process: given an image as input, possible logo candidates are detected and classified accordingly. The end-user can filter the output by interactively setting a classification threshold. In order to obtain the best results on the challenging CCTV dataset the network was trained from

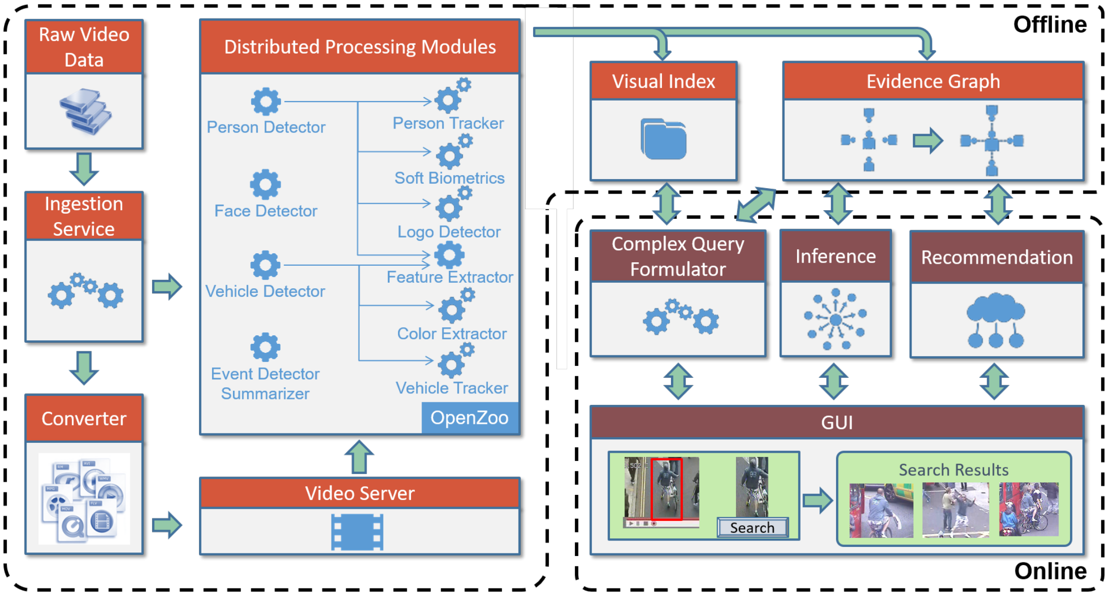


Fig. 1. The main workflow of the proposed framework.

scratch with ad-hoc and relevant data. This required the generation of a training dataset with thousands of images with metadata (i.e., information about which and where a logo is present in the image). Four different logos were trained: Adidas, Nike, the North Face, and Superdry. On average, 16,000 different images were used in the training phase of each brand.

3.2 Object Tracking

Object tracking is another cornerstone in surveillance applications offering advantages in multiple fronts. It constitutes an essential tool for investigators as it allows them to effortlessly track objects (including people and vehicles) in the examined scene, in an attempt to examine their unique identity or possible contribution to suspicious events. Moreover, object tracking contributes to better managing the extracted analytics for the examined footage. In large scale investigations the volume of the detected objects is posing important challenges to the indexing and retrieval functionality of the framework. Using object trajectories to prune the extracted analytics is critical for the smooth operation of the system.

However, object tracking is a challenging task, especially in dense multi-object environments. Occlusion, objects entering/exiting the scene and detection sparsity are the main challenges faced by tracking methodologies. In LASIE, the object tracking module is following the tracking-by-detection approach. Apart from the different object detectors, two methods have been explored for the tracking. In the first one, the tracker is modelled as data association problem and it solved by the well-known Hungarian algorithm approach. The algorithm receives as input a set of distances established between every pair of detections based on their spatial position and visual similarity. Once the matching cost is computed, the set of correspondences is found with minimum global cost. On the other hand, the second method uses only the spatial information. In this case in order to have a reliable tracking, a motion model is used to propagate the targets identity into the next frames. The velocity components of the target state are solved in an optimal way with the Kalman method. Using this filter is more important in this approach due to the fact that no visual appearance is used and thus the noise in the spatial position has more impact on the quality of the correspondences.

LASIE contribution enhances tracking focusing on the post-processing of the identified tracks. Post-processing of the tracks is perfomed for every distinct track ID in each single sequence. The first and the last detected box with the same track ID form a subsequence and in case of missing frame detections in this timeline, padding process fills those frames by creating adjacent detections with equal distances. Overall tracking is improved and experimental results show the efficiency of the contribution. The tracker runs almost at real time. For example, for a video where 5-10 people appear, the tracker processes at 5-6 fps (0.18s).

3.3 Soft biometrics

Soft biometrics are biometric traits that do not offer exact human identification, however, they can provide adequate information to narrow-down the search space and give valuable insights for the subject in question. In LASIE we followed the approach of [13] based on the concept of Exemplars, that is, to find matches of the examined subject over a labeled dataset, which is able to encode the quality, colour and light variations of the surveillance images. The core idea of the proposed approach is to have a fast and robust segmentation of each detection that will improve the sampling and the feature extraction process, thus, enhance the accuracy of the identified soft biometrics We used this for labelling of the subjects clothes colors, however, we have tested the approach to improve other soft biometric traits in similar conditions.

3.4 Event Detection and Video Summarization

Human video perception is limited when exploring huge amounts of data. The video summarization module in LASIE aims at reducing the amount of data for the user and provides an overview of the contents which highlights potential events of interest for further investigation. For this task, after the data upload to the LASIE system, videos are processed by automatic video summarization techniques. This step consists of a multi-cue video summarization module which incorporates higher-level information in an event-based approach following the methodology from [9]. The output of this module is a list of coherent split sequences with related key frames and assigned contextual events for each of them. For this purpose, the LASIE video summarization module is based on the following contextual events: a) camera motion, b) violent scenes and c) abandoned objects. The events identified in these steps determine the video splitting process and the generation of key frames for the split videos.

Camera motion events are detected in order to extract coherent scenes with static fields-of-view by applying sparse optical flow-based global motion estimation. Zoom-ins are also included since they indicate specific interest by the camera operator. Next, frame similarity is computed for the split scenes in order to extract the most significant key frames for each sequence. Color descriptors and point features can be used in this step due to their robustness and low computational load. Key frame candidates are assessed using a gradient-based crispness measure in order to avoid the extraction of key frames suffering from motion blur or frames with bad focus (e.g. after a zooming operation).

Violent scenes event detection automatically classifies the appearance of violent actions in the video footage. This approach is based on the Lagrangian video-level descriptor LaSIFT introduced in [6], which encodes salient spatio-temporal motion patterns in video and thus indicates violent behavior. The motion patterns are previously trained based on a labeled dataset and extracted in a temporal windowing approach for sets of 50 consecutive frames. The final classification is based on a support vector machine with non-linear kernel and provides a continuous violence score over the video. In order to maintain a better level of temporal consistency, the violence score can be refined using a median filter.

Abandoned objects can be important cues in a video in order to identify potential threats or to derive suspicious behavior. Related events are detected based on the approach from [8] which uses a Gaussian mixture dual foreground model in order to classify static objects.

4 The Evidence Search Engine

4.1 Image and Video Indexing

For each of the detected people and vehicles, visual features that encode color and texture are computed: i) an HSV color histogram in 5 horizontal stripes of the bounding box (colour) and ii) high dimensional vectors from the activations of the deep learning model (colour and texture). These features are used to search within the databases using the “query-by-example” approach. However, due to the extremely large scale of detections, an indexing structure that will reduce the search space and increase the retrieval speed-up is required. A typical approach is that of binary hashing, where a data-independent or data-dependent hash function is learnt and applied in the initial high-dimensional vectors to get binary vectors of much smaller dimensionality and storage size. Most importantly, these binary vectors can be easily compared with a query vector using the low-complexity hamming distance. In the case of crime investigations, it is typical that the queries of the investigator refer to a specific location or time period and this is obviously reducing even more the search space under consideration.

To support these requirements, in LASIE we developed a binary hashing index structure that is based on the spherical hashing algorithm [12] extended with location and time filters in the same structure. Each detection of the person or vehicle is a record in the index that holds the descriptor vector along with the location information of the camera from which the video came, as well as the time of capturing. The structure holds a selection bitset for locations and a temporal bitset that encodes time information to improve complex queries. In order to support both linear and circular time modes (i.e. “after 2 Feb. 2017” or “every Monday”) the temporal bitset for each record holds 74 bits. The 74-bit vector is structured as week days, month number, day of month, hours since midnight, etc., thus enabling fast bitwise operations to extract a small part of the dataset to work with.

4.2 Evidence Graph

The LASIE Evidence Graph (EG), which is basically a heterogeneous information network (HIN), is built from data produced by the processing modules. The modules detect and annotate a number of different entities (nodes), such as persons, vehicles, faces, etc. The goal of the EG is to connect these objects with various relationships (edges), in order to provide usable information to the analyst, as well as to the Inference and Recommendation modules. Each node may have attributes and each edge may have weights attached to them, carrying semantic information.

The EG consists of two main parts: the database and the knowledge extractor. The database stores the objects produced by the processing modules as graph nodes (e.g. Vehicle), along with their annotations as attributes (e.g. SubType Car, Color Red). Some trivial relations are stored as graph edges (e.g. ”Person X is part of Track Y”). The knowledge extractor identifies additional relations between the objects and creates the appropriate edges in the graph. Such relations include the connection between two tracks with at least one common frame, the connection of visually similar people, etc. The graph can be further enriched by a posteriori updating the knowledge extractor.

The popular graph database OrientDB[2] has been chosen for hosting the EG, since it achieves a good performance in graph-related functionalities, such as shortest path, traverse and aggregate operations.

4.3 Complex Query Formulation

The Complex Query Formulator (CQF) is the core component of LASIE for answering complex queries from different

[2] http://orientdb.com

modalities and combined in various ways. The term complex query refers to the fusion of multiple modalities of queries to produce a unified list of results from the evidence data.

To support search in large video collections, involving various scene types, LASIE is providing a set of tools for processing and analysing the videos, so that many different types of data are generated from each video, e.g. faces, people, tracks, vehicles, logos, scene classification, or video frame regions, etc. These are considered as the various modalities of content that were extracted from the videos. Having analysed the videos in the offline process, the end user (e.g. a Police officer) can search using content-based queries i.e., by selecting an area within a video frame and search for visually similar items. This type of query is considered as a uni-modal query since the content type of the query and of the database we are searching in are the same. However, there are many cases where a user may need to search for a certain modality by using a query of another modality, or by using a combination of modalities. These are considered as multimodal queries that permit combinations of modalities in the search process.

All modalities are inserted in the EG and create different edge types to link together graph nodes that are related in some manner. The EG-based multimodal search is considered the ideal configuration for combining different modalities together and enabling cross-modal and multi-modal queries, even with new modality types. The CQF is the module that builds the queries around the properties of the EG and formulates the different modalities as graph traversal queries. Having all the extracted data within a large graph, the possibilities of building different queries are endless. It is obvious that to exploit all these possibilities a query language is necessary. However, one of the goals of LASIE is to provide an easy to learn and easy to use user interface for the investigators. Thus, our final CQF provides a list of pre-selected queries in the front-end user interface, while it permits full configuration to support any other complex query the users need, with minimal configuration.

As an example, a complex query that the investigator would like to execute might look like the following one: “Fetch me all people that were **next to** someone like the one in the visual query and any other person that might look like them in any video and have green upper body clothes”. This query can be built by the user as follows: [Visual Query] + [Co-exists] + [Visually Similar] + [upper body: green], where the [Visual Query] is drawn directly from the input video, while [Co-exists], [Visually Similar], [upper body: green] can be selected at the appropriate fields of the LASIE User Interface (Figure 4). The steps of this query in the CQF are depicted in Figure 2. First, we get a list of tracks using the visual query. Then, for each one of the selected tracks, we traverse the EG with **co-exists** edges to fetch people next to the one in the visual query. Next, we also traverse the EG with edges **similar-to** to fetch any other track that is similar to these ones and are in any video in the collection. Finally, we keep only those nodes that match the attribute **upper-body=green**.

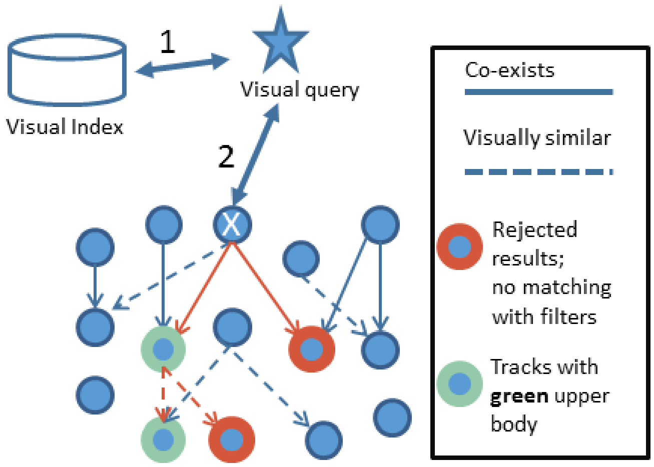

1
Visual Index
Visual query
2
X
Co-exists
Visually similar
Rejected results; no matching with filters
Tracks with **green** upper body

Fig. 2. An example Complex Query Formulator flow.

4.4 Inference

The role of the inference engine in the LASIE platform is to help the police officer in validating investigative hypothesis about crime events by selecting the appropriate evidences. An investigator is formulating a hypothesis by asking *who-where-when*-based questions involving entities such as persons, objects, scenes, places. These entities are retrieved using their visual appearance, which can be recognised by the LASIE computer vision modules, and/or by exploiting the knowledge stored in the EG and in the inference module. For example, the inference module will process semantic concepts (e.g. X supporter of team Z) by associating their meaning to their visual representations in the physical world (e.g. since black is the official colour of team Z, if someone wears black shirt it is likely to be supporter of team Z). Let us consider a murder case where an investigative hypothesis could involve an entity X, which is a person dressed in a red jacket and an entity Y, which is a supporter of the team Z. The instances associated to X can be retrieved by the visual attributes (i.e. red jacket), while the ones related to Y can be found inferring from the EG and the inference module that the supporters of the team Z are dressed in a black shirt. In this context, the hypothesis is formulated by expressing a set of queries involving the previous entities such as: (i) *Who* was close to X during the violent action? (ii) *Where* X and Y were detected close to pubs around the city? (iii) *When* X and Y were entering into the stadium? The inference will help the officer by selecting the most relevant evidences that can support his/her hypothesis and let him/her to classify it as consistent (i.e. there is clear interepretation of the evidences), relevant (i.e. it clarifies an important aspect of crime), persuasive (i.e. it can be used to handle objections effectively).

The inference engine is based on two main algorithms. The first one has the goal of filtering the evidences by selecting the ones that are more relevant to the entities and at same time closer to the spatial areas involved in the hypothesis definition. The algorithm is based on an Integer Liner Model, where the objective function is defined in order to maximize the coverage of relevant evidences (e.g., suspects) around points of interest (e.g pub, tube station, stadium). The second one has the goal to associate the most probable instances to the entities linked to the hypothesis and it is based on probabilistic rules in Markov Logic Network [4]. In particular, it is implemented using the services provided by the Tuffy engine[3].

4.5 Recommendation

The LASIE recommender system performs reasoning on top of the EG, which is modelled as a large-scale Heterogeneous Information Network (HIN). Its main goal is to identify in the vast collection of multimedia content contained in the EG, nodes and regions potentially relevant for the investigation. To this aim, this module performs online local analysis of the EG starting from a query node that is, explicitly or implicitly, selected by the user. The query node corresponds to some piece of evidence (a person, a vehicle, etc) that has previously received the users' attention. The user can provide relevance feedback to the system by saving for further analysis recommendations that are indeed relevant.

The LASIE recommendation functionalities for the EG include similarity search, relevance recommendation, local clustering and ranking [10]. The implemented techniques rely on the meta-path concept, which represents a sequence of nodes and links carrying a relevant semantic meaning.

Similarity search allows to identify semantically related objects in HINs and is the basis of many data mining tasks, such as recommendation, clustering and ranking. The LASIE recommendation module supports the computation of PathSim and HeteSim similarity measures. PathSim evaluates the similarity of same-typed objects based on symmetric paths, while HeteSim computes the relevance of same or different-typed objects under arbitrary meta-paths. The LASIE system also implements the HeteRecom [11] recommendation algorithm for HINs, which aims at identifying multi-typed items related to a query object in terms of meaningful semantic relationships.

On the other hand, the LASIE system addresses the problem of HINs clustering, which consists of inducing densely connected heterogeneous sub-networks. A representative example of clustering in HINs is the NetClus [10] global algorithm that discovers net-clusters and provides ranking distributions. However, in the LASIE framework, a local clustering approach is preferred, as it allows to extract the cluster or community of one or several query nodes in real time, without processing the complete EG. The proposed algorithm is an iterative procedure that infers the highly connected sub-network of interest, based on the similarity measures PathSim and HeteSim. Global HIN ranking [10] is then applied to the resulting sub-network in order to identify outstanding EG nodes which can be potentially relevant for the investigation.

5 Large-Scale Analysis

For mitigating the parallelization needs of LASIE, OpenZoo[4] has been utilized, a framework for distributed stream and batch processing with emphasis on multimedia content, partially developed within the LASIE project. The framework is designed to hide the deployment and communication complexities of large distributed and heterogeneous topologies, aiming to permit developers and researchers to focus on the intelligence of their applications. OpenZoo uses MongoDB[5] for persistence and RabbitMQ[6] as a communication medium and supports Java, Python and C++ for service implementation, allowing service deployment on Windows and Linux based systems. All distributed processing modules (Figure 1) have been implemented within OpenZoo, allowing multiple instances of each module to operate simultaneously.

The model followed includes a producer service, which locates the I-frames of a specific input video file and creates task messages, each of them containing a video segment, in form of a start I-frame position and an end I-frame position. Overlapping sections between the segments allow the produced tracks to be merged at the end of the process. In our case, the desired video segment length is set to 2 min and the overlapping to 10 secs (5 secs at the beginning and 5 secs at the end of each segment). The use of this processing model eliminates manual labour, connected to the balanced supply of all the processing module instances with video input.

Each task message is sent to all processing modules. Since more that one modules can be deployed on each server, a cache has been developed to minimize network traffic, by keeping a downloaded segment enough time to be processed by all modules, before it is deleted to save space. The tracks produced for each segment are sewed back together at the overlapping positions and a representative detection is produced per track, using the detection confidence, as well as the detection's position in the track and the size of the bounding box. Finally, the features extracted from the detections are sent to the Visual Index and all aggregated tracks and detections are sent to the EG.

Regarding scalability, OpenZoo's limits coincide with RabbitMQ's limits: a) the system can handle as many servers as consumers can be connected to a RabbitMQ queue and b) the system can handle as many subtasks, as messages can be posted to a RabbitMQ host. Apart from that, given enough disc space available for hosting the raw video, the database and the queues, there is no other central service or resource that can negatively affect scalability.

6 The LASIE User Interface

The LASIE User Interface provides an ergonomic virtual environment, designed in close collaboration with police officers and forensic experts to support investigation management, data ingestion, query formulation, forensic hypotheses management and exporting evidence. Card design pattern is adopted to provide a suitable end-user experience thanks to their flexibility, responsiveness and dynamic structure able to summarize heterogeneous contents within limited size.

LASIE provides an open source framework, OPENNESS.sec, to support forensic activities during all stages, from the creation of the investigation to the export of the evidence for the court. All resources collected and analysed for

[3] http://i.stanford.edu/hazy/tuffy/
[4] http://openzoo.org
[5] https://www.mongodb.com
[6] https://www.rabbitmq.com

investigations are organised according to specific Forensic Smart Folder and visualised on the top of the Investigation Wall (Figure 3).

The morphology of a card consists of the following parts: thumbnail, type, title, short description and actions bar. Specific infographics are used to indicate the type of resource, the level of dangerousness and the relevance for the investigation. Users have the opportunity to annotate any type of resource thanks to a dedicated annotation panel. Smart widgets have been designed to allow investigators to visualise and interact with the results of the analysis; others provide graph management capabilities to make available functionalities for adding and removing links between people and criminal events, including spatio-temporal and causal relationships.

During query formulation, users are assisted using an interactive interface (Figure 4). Queries can be composed by keywords and/or by selecting a bounding box related to the target object or person in the user interface. Investigators have the opportunity to filter selecting a geographical area, a time range and other distinctive characteristics as such as logos, colors of the clothes, etc.

7 Evaluation

Although the scope of the paper is to present a full framework for large-scale analysis of "video in the wild", it is worth mentioning that each separate low-level processing module, which has been integrated in the framework, has advanced the state-of-the-art in the respective area. These advances are summarised in this section.

Object detection: The faster method presented in Section 3.1 has been tested in MOT2D challenge[7], where it achieved a Multiple Object Tracking Accuracy (MOTA) of 24.61%, excelling among same-speed methods on comparable hardware.

Logo detection: the performance of the logo detection method presented in Section 3.1 was evaluated on footage from CCTV cameras. The recordings are 704x624 pixels in resolution; captured at 12fps, and encoded with MPEG4 at around 150 KBps. The performance was measured in terms of recall and accuracy. Additionally, the results of the People Tracking module were used in order to reduce the number of false positives obtained. In fact, candidate detections not overlapping with a person (e.g., logos on cars, or buildings) are not relevant for the LASIE scenario. The logo detection module achieves high accuracy and recall, in all four categories of logos (Adidas, Nike, the North Face, and Superdry). More specifically, the lowest accuracy is achieved in Adidas logos (82.54%) and the lowest recall in Nike logos (57.74%), while the highest accuracy in Nike logos (91.55%) and the highest recall in Adidas (98.48%).

Event detection: the algorithm based on the new LaSIFT descriptor that classifies automatically the violent actions has been tested on the MET dataset. Compared to the baseline method MoSIFT [7], the proposed LaSIFT achieves higher accuracy (84.00 ± 7.42 compared to 72.38 ± 11.86 of MoSIFT). Exemplary violent event detection results are presented in Figure 5.

Regarding the modules presented in Sections 4 and 5, quantitative evaluation, i.e. comparing each algorithm with a baseline on a specific dataset, is not available, since there are no forensic tools with similar functionalities, to the best of our knowledge. The novelty of these methods is that they offer totally new capabilities to the investigator for analysis, search and retrieval of relevant evidence, and that they exploit uniquely the capabilities of the forensic data processing modules (Section 3).

8 Conclusion

A new framework for large-scale exploitation of forensic data has been presented. Several tools for video analytics are proposed, appropriate for use in video "in the wild", i.e. taken from real CCTV street scenes. The tools perform automated object (human, face, vehicle, logo) detection and tracking, video event detection and summarisation and are robust in low-resolution, low colour quality, motion blur and lighting variations. The proposed Evidence Search Engine offer numerous capabilities for retrieval of relevant evidence, allowing for complex query formulation, presenting recommendations and filtering out inconsistencies to reduce the search space. An appropriately designed user interface facilitates the use of the analysis tools by the investigators. The framework supports distributed processing of large volumes of video, through an intuitive approach, exploiting optimally the available computational resources. The framework has been demonstrated in real content provided by the London Metropolitan Police, showing very promising results.

Acknowledgments

The work presented in this paper was supported by the European Commission under contract FP7-607480 LASIE. The authors would like to thank the London Metropolitan Police, UK, for providing the CCTV video footage and for giving permission to process the dataset for research purposes.

Additional authors: Theodoros Semertzidis, Michalis Lazaridis, Anastasios Dimou, Tobias Senst, Jaume Amores, Federico Alvarez, Lukasz Kondrad, Giuseppe Vella, Antonino Randazzo and Pietro Pomo.

References

[1] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

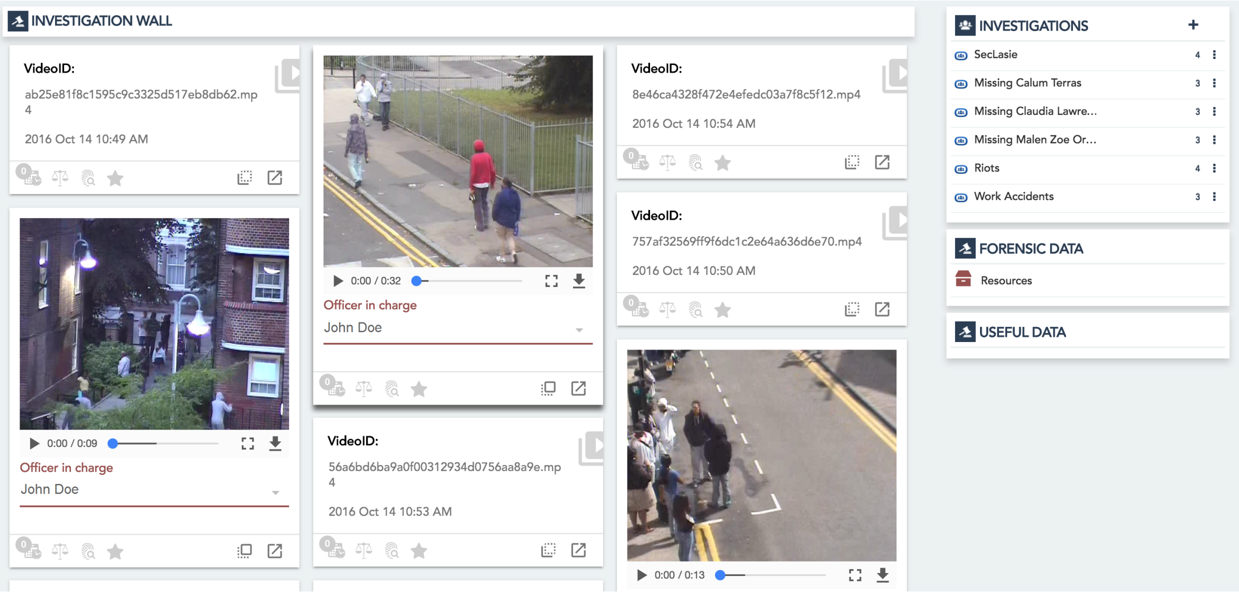
[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[3] Anastasios Dimou, Paschalina Medentzidou, Federico Alvarez and Petros Daras, "Multi-target Detection In CCTV Footage For Tracking Applications Using Deep Learning Techniques", *IEEE International Conference on Image Processing, ICIP 2016*, Sept 25-28, Phoenix, Arizona, USA.

[4] Matthew Richardson and Pedro Domingos. "Markov logic networks". *Machine Learning Journal*, 62, 1-2, pp 107-136, 2006.

[5] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.

[7] https://motchallenge.net/data/2



INVESTIGATION WALL

VideoID:
ab25e81f8c1595c9c3325d517eb8db62.mp4
2016 Oct 14 10:49 AM

0:00 / 0:09
Officer in charge
John Doe

0:00 / 0:32
Officer in charge
John Doe

VideoID:
56a6bd6ba9a0f00312934d0756aa8a9e.mp4
2016 Oct 14 10:53 AM

VideoID:
8e46ca4328f472e4efedc03a7f8c5f12.mp4
2016 Oct 14 10:54 AM

VideoID:
757af32569ff9f6dc1c2e64a636d6e70.mp4
2016 Oct 14 10:50 AM

0:00 / 0:13

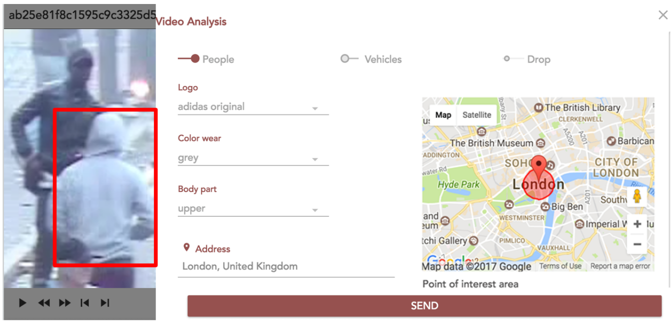
INVESTIGATIONS +

SecLasie	4
Missing Calum Terras	3
Missing Claudia Lawre...	3
Missing Malen Zoe Or...	3
Riots	4
Work Accidents	3

FORENSIC DATA
Resources

USEFUL DATA

Fig. 3. The Investigation Wall of the LASIE UI.



ab25e81f8c1595c9c3325d5
Video Analysis

People ○ Vehicles ○ Drop

Logo: adidas original
Color wear: grey
Body part: upper
Address: London, United Kingdom
Point of interest area

SEND

Fig. 4. The Query Formulation of the LASIE UI.

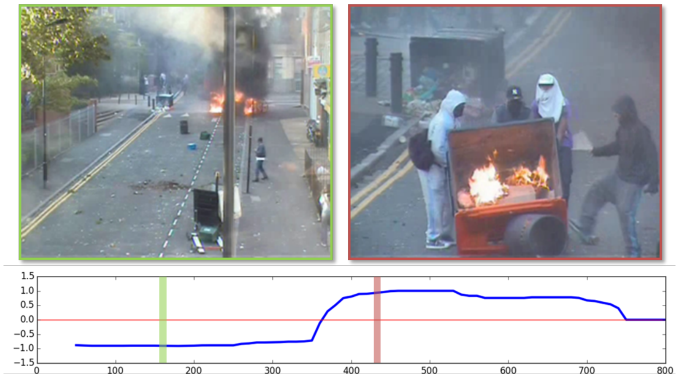


Fig. 5. Exemplary violent event detection results: frame before visible violence (left, green marker), aggressive behaviour recognized (right, red marker). Blue plot indicates frame-wise violence score.

[6] Tobias Senst, Volker Eiselein, and Thomas Sikora. A local feature based on lagrangian measures for violent video classification. In *IET International Conference on Imaging for Crime Detection and Prevention*, pages 1–6, 2015.

[7] Enrique Bermejo Nievas, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence detection in video using computer vision techniques. *Computer Analysis of Images and Patterns*, pages 332–339, 2011.

[8] Rubén Heras Evangelio and Thomas Sikora. Static object detection based on a dual background model and a finite-state machine. *EURASIP Journal on Image and Video Processing*, Vol. 2011:11, 2011.

[9] Rubén Heras Evangelio, Tobias Senst, Ivo Keller, and Thomas Sikora. Video indexing and summarization as a tool for privacy protection. In *IEEE International Conference on Digital Signal Processing (DSP 2013)*, 2013.

[10] C. Shi, Y. Li, J. Zhang, Y. Sun and P. S. Yu, "A survey of heterogeneous information network analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 17–37, 2017.

[11] C. Shi, C. Zhou, X. Kong, P. S. Yu, G. Liu and B. Wang, "Heterecom: a semantic-based recommendation system in heterogeneous networks," *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1552–1555, 2012.

[12] Heo, J. P., Lee, Y., He, J., Chang, S. F., and Yoon, S. E. (2012, June). Spherical hashing. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 2957-2964). IEEE.

[13] T. Semertzidis, A. Axenopoulos, P. Karadimos, P. Daras, "Soft Biometrics in Low Resolution and Low Quality CCTV Videos", 7th International Conference on Imaging for Crime Detection and Prevention (ICDP-16), Madrid, 23-25 November, 2016

[14] V. K. Badrinaravan, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation", CVPR, 2015.

[15] Y. Lu, , A. Chowdhery and S. Kandula, "Optasia: A Relational Platform for Efficient Large-Scale Video Analytics", Proceedings of the Seventh ACM Symposium on Cloud Computing. ACM, 2016.