Management Of Networked IoT Wearables – Very Large Scale Demonstration of Cultural Societal Applications
(Grant Agreement No 732350)

D5.3 Modelling of Complex Dynamics and Information Retrieval for Post-Event Analysis 1

Date: 2018-02-28

Version 1.0

Published by the MONICA Consortium

Dissemination Level: Public

Co-funded by the European Union's Horizon 2020 Framework Programme for Research and Innovation under Grant Agreement No 732350

Document control page

Document file: D5.3-Modelling_of_Complex_Dynamics_and_Info_Retrieval_v_01
Document version: 1.0
Document owner: KU

Work package: WP5 – Security Closed Loop Systems
Task: T5.3-T5.4 – Scene and Dynamic Modelling, Information Retrieval
Deliverable type: [OTHER]

Document status: ☒ Approved by the document owner for internal review
☒ Approved for submission to the EC

Document history:

Version	Author(s)	Date	Summary of changes made
0.1	Rob Dupre (KU)	2017-12-21	Initial Draft and Plan TOC
0.2	Rob Dupre (KU)	2018-1-12	Updated Plan
0.3	RD (KU) BW (VCA) SZ (LBU)	2018-1-26	Initial Merge of partner sections.
0.4	RD BM HS (KU)	2018-2-02	Addition of KU algorithms and results
0.5	DK (CERTH) AM (HWC) BW (VCA)	2018-2-08	Addition of CERTH and HWC Sections
0.6	RD BM HS (KU) AM (HWC)	2018-2-21	Changes based on review comments
0.7	RD (KU)	2018-2-27	Changes based on review comments FC (ISMB)
1.0		2018-2-28	Final version submitted to the European Commission

Internal review history:

Reviewed by	Date	Summary of comments
Sebastian Meiling (HAW)	2018-02-16	Overall, minor rewording. In Section 4.2.1 Storage requirements: clarify network setup and configuration (e.g. authorisation, redundancy, ...) in more detail.
Vincent Gissinger (ACOU)	2018-02-15	A few video concepts are still to be defined to clarify some points in this good document.

Legal Notice

The information in this document is subject to change without notice.

The Members of the MONICA Consortium make no warranty of any kind with regard to this document, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The Members of the MONICA Consortium shall not be held liable for errors contained herein or direct, indirect, special, incidental or consequential damages in connection with the furnishing, performance, or use of this material.

Possible inaccuracies of information are under the responsibility of the project. This report reflects solely the views of its authors. The European Commission is not liable for any use that may be made of the information contained therein.

Index:

1 Executive Summary

This deliverable documents the current progress relating to tasks T5.3 Scene and Dynamic Modelling and T5.4 Information retrieval over the last 12-1 months of the MONICA project. This is the first iteration of this deliverable with a second version due in month 30. Outlined below is an overview of the tasks T5.3 & T5.4 and the various applicable work undertaken thus far. This includes the application of existing techniques for the task goals as well as the development of new.

The aim of WP5 in the MONICA project is to use existing and newly-developed video-based sensors and algorithms to extract salient information from a scene (in this case a MONICA pilot site) to achieve the goals set out by the generated user requirements. T5.3 is focused on the creation of statistical models of scene dynamics. These models form the basis for detection algorithms, for example allowing the software to highlight the presence of a fight in a scene. T5.4 is specifically designed to develop post event analysis tools in an effort to reduce the time taken to extract footage or imagery of a particular incident or event.

The creation of statistical models is integral to the development of the algorithms utilised in the various other WP5 tasks (specifically T5.1) as such good progress has been made in this area already. As much of this is covered in the D5.1, rather than re-report this progress, the focus will be on the modelling processes themselves, the acquisition and use of data and some interim results on an example algorithm.

Progress on Information retrieval is also positive with some new research currently under review and some existing saliency techniques implemented. There is an outstanding issue relating to this task stemming from the lack of defined function requirements, this is due to be reviewed and on resolution will further drive progress in this task.

2 Introduction

2.1 Purpose, context and scope of this deliverable

The aim of WP5 in the MONICA project is to use existing and newly-developed video-based sensors and algorithms to extract salient information from a scene (in this case a MONICA pilot site) to achieve the goals set out by the user requirements established within WP2. Within this document clarity on what each task entails as well as how the tasks currently fit into the wider MONICA context is provided.

T5.3 Scene and Dynamic Modelling is focused on the creation of statistical models of scene dynamics, these models form the basis for detection algorithms, for example allowing the software to highlight the presence of a fight in a scene. T5.4 Information Retrieval is broken up into two distinct sections, firstly the process of Information retrieval, i.e. the searching and retrieving of previously recorded video/image data based on some search criteria. The second, Image/Video mining, is specifically designed to develop post event analysis tools in an effort to extract additional information from the saved image and video data.

2.2 Structure and content of this deliverable

This document is split into two distinct areas, relating to the two tasks covered within this deliverable.

- T5.3 Scene and Dynamic Modelling: the current techniques that require statistical modelling techniques, as well as their preliminary results. Additionally, an overview of the currently available MONICA data and publicly available datasets which could be used in the absence of relevant MONICA data.
- T5.4 Information retrieval: broken down into the tasks two component areas, Retrieval and Video Mining, with the current research being conducted into the area as well as overviews of existing applicable techniques.

For each of these sections, as well as an overview of the progress and current results applicable to that task, the wider MONICA context is provided. This elaborates on both the architecture and deployment impact as well as the relevant user requirements and solutions where these tasks are applicable.

3 Modelling of Complex Dynamics

3.1 Introduction

The modelling of complex dynamics provides the backbone from which recognition algorithms perform their tasks. For example, computer vision and machine learning algorithms which are designed to detect the presence of a human in an image should have some previous definition of what constitutes a person, this definition is statistically or mathematically described within the model. Such models are formal mathematical representations of the behaviour of real human beings or objects used to describe and make predictions about them. A given algorithm then utilises that model as a reference from which a decision about the presence of the modelled pattern in a scene can be made. Training data is used to learn such mathematical representations (models). The better the training data, the better the model representation and therefore the recognition performance.

Within WP5, a number of techniques utilise created models of specific scene dynamics within the algorithms developed, many of these are covered within the Deliverable 5.1 and as such only cursory mentions will be given to those where relevant. Instead, within this section a detailed example is given of one such algorithm that utilises the modelling of complex dynamics. Within complex dynamics modelling, we are trying to model low level features such as detection of humans, objects, density and creating mid–level knowledge representations or sense out of them. For example, a large number of people in a small area (camera view) would be termed as densely crowded area. Furthermore, we are modelling complex interactions between objects (such as humans) as fight, protest, mob, etc. This would help us to make higher level understanding of the crowd behaviour.

To contextualise the use role of task 5.3 within the MONICA project, Table 1 outlines the use of dynamic scene modelling within the current MONICA solutions, the linked MONICA use cases, along with a brief explanation of how trained models will be utilised.

Table 1: Task 5.3 Scene and Dynamic Modelling within MONICA

MONICA Solution	MONICA Use Case	Use of Task in context
Locate Person/Object • Approximate - using 868MHz wristbands. Accuracy about 15 meters. The system allows for tracking 1000 wristbands each second. Hence, for 10.000 wristbands there will be a new location available every 10s. • High-precision - UWB wristbands. Accuracy about 50 cm. Maximum 2000 location updates per second.	• UC 4.2 Locate lost person • UC 4.4 Locate parent • UC 5.1 Locate staff member	The location model can track the approximate/precision position of a wristband dynamically, by using distributed anchors network. Therefore, if the required locating person who wears a wristband, it is easy to find using the location model.
Counting people Counting people at entrances and exits using digital tickets or cameras.	• UC 3.1 Detect high risk queues • UC 3.3 Monitor crowd based on capacity	The counting people model can be used to monitor the crowd density in various areas of an event. Based on these counts and other crowd properties (such as flow), predictions can be made on potentially high-risk queues or developing unsafe crowd situations.
Crowd behaviour analysis Crowd Abnormality Detection (vehicle entering the crowd, person going against the flow, etc.), and Fighting Detection using wristband or cameras.	• UC 7.1 Detecting an incident	The crowd behaviour analysis model can detect abnormal crowd behaviour e.g. fighting, falling, using human activity analysis algorithms.

3.2 Architecture

Within this section an architecture overview is given of the components relating to T5.3 Scene and Dynamic Modelling.

Figure 1 outlines the Deployment architecture as specified in the first iteration of the MONICA architecture (D2.2). Highlighted are the relevant components applicable to T5.3.

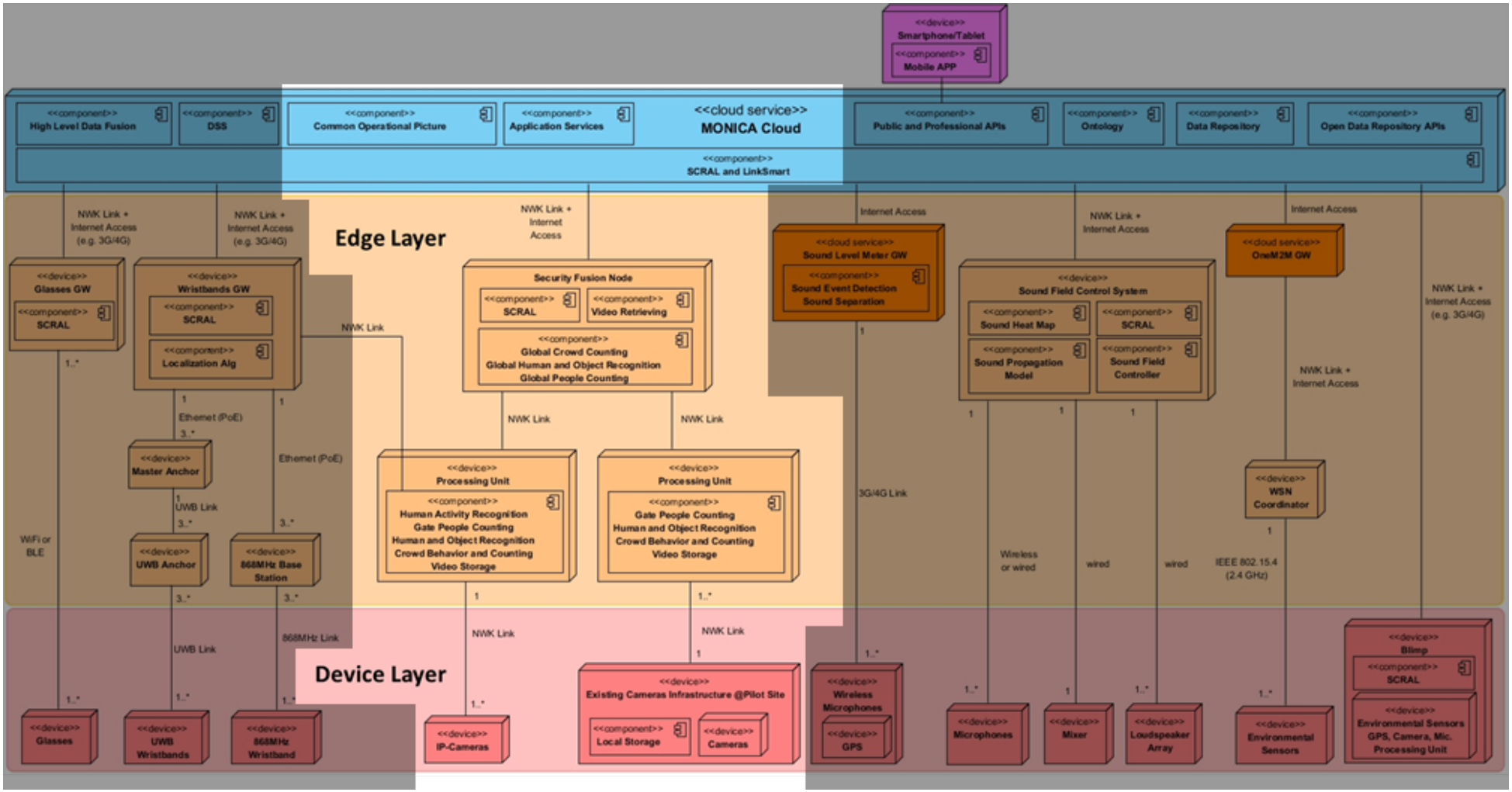

Figure 1: Detailed view of the MONICA Deployment Architecture (Figure 18 D2.2).

The created models themselves will sit on the components that run the algorithms, for example the people counting algorithm will make use of a number of crowd density and human detection models, as the processing of these algorithms is done on the processing nodes the copies of the relevant models will also reside within this component (Figure 2). The training and definition of these models is outlined in more detail in the following sections. The update process for these models is as-yet undefined and requires further discussion for the current MONICA platform as the concept of retraining the model after the completion of the project has not be raised in the user requirements. As such the models themselves are currently seen as static, with the option of manual updates based on the acquisition and ground-truthing of new data.

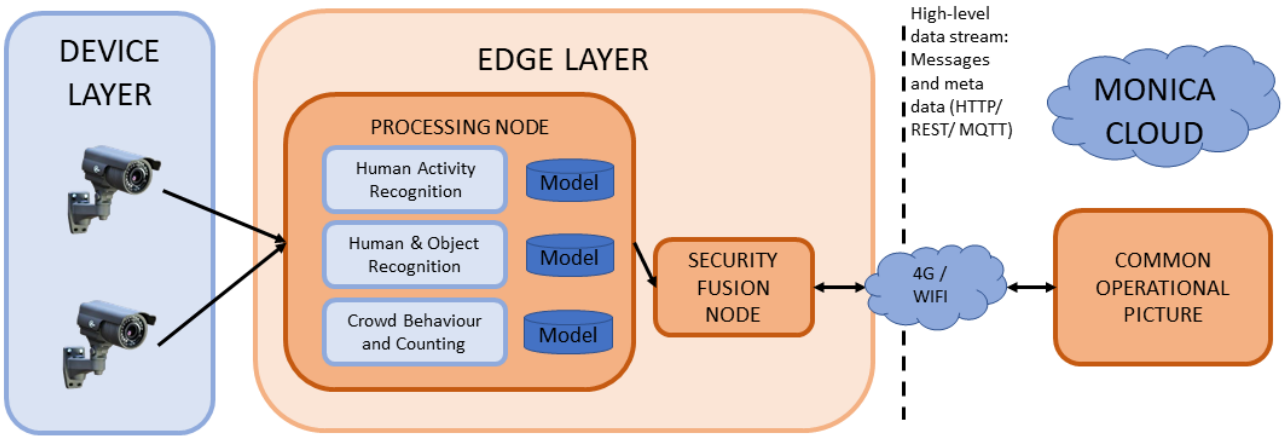

Figure 2: Extrapolated view of the WP5 components with respect to the storage and use of models with the various detection algorithms.

3.3 Crowd Behaviour Recognition

Within this section an example algorithm that utilises modelling techniques is given.

3.3.1 Crowd Behaviour Recognition Algorithm

Modelling complex dynamics in a minimally non-invasive way is a very challenging problem. Currently, deep learning-based models are being developed to automatically generate rich pattern representations of the crowd videos involving violent crowd behaviour. These patterns are then analysed using emerging machine learning algorithms to detect and recognize relevant crowd monitoring statistics. As illustrated in the architecture overview these statistics are then fed up to the MONICA cloud for further high-level processing from which security alerts can be generated.

Crowd Behaviour Recognition is a specific example of crowd analysis focused on the identification of certain crowd behaviours from a video sequence. Crowd analysis using visual data is becoming very common at various kinds of public events such as concerts, sport matches in stadiums, celebrations, protests, public gatherings at train stations or bus stops, to name a few. According to the popular reviews [Li2015, Julio2010], a large amount of work has been done to track, recognize and understand various behaviours in videos. These works have mainly focused on common scenes with low density of population, however, relatively little effort has been devoted to reliable classification and understanding of human activities in real-world crowded scenes. In heavily crowded scenes, often the detected objects (including humans) are very small, making the recognizing, analysing and characterizing of their interactions (such as behaviour) very challenging. In general, research has followed two ways of analysing them. Firstly, holistically, where individual components such as objects, places, scenes, their actions or interactions are not identified or classified individually, rather they are processed based on their whole appearance [Solmaz2012]. It is often advantageous to understand the crowd behaviour without knowing the actions of the individuals. Secondly, object-based approach, where individuals (human and objects) are detected and segmented to perform behaviour analysis [Zhou2012]. This kind of complex segmenting and tracking of individuals in crowded videos is a very challenging task. In our work we use the former, where individuals are not segmented or tracked, but the group of people are perceived holistically so as to recognize their behaviour.

In the context of MONICA, we seek to identify four violent crowd behaviours: *fight*, *mob*, *protest* and *protester* from *normal videos*. *Protest*, *fight* and *mob* are taken as events, whereas *protesters* are groups of people going to protest or in progress. All these are annotated in the publicly available database [Shao2015].

Normal videos constitute all other video where these 4 crowd behaviours are not present but would include other activities performed by a crowd, such as standing, walking, clapping, waving, etc. This would help us to generate alerts or summarize the crowd behaviour in good time (e.g. online processing involving a few seconds of analysis). Some of the images from the database [Shao2015] are shown below in Figure 3:

Figure 3: (Top) A fight, (Bottom) A protest extracted from the WWW Crowd Dataset

Our proposed framework is shown in Figure 4 below. As shown in the figure, we extract rich representations from crowd behaviour videos using fine-tuned deep convolutional neural residual network [Zhou2012] and then create subclasses within feature maps from each class. Features from these subclasses are then regularized using an eigen modelling scheme. This has helped to model the variances appearing from the intra-subclass variance information. Final low dimensional discriminative features are extracted after using between-class information. Dynamic time wrapping (DTW) is applied on the cosine distance measure to find the similarity measure between the two videos. 1-NN (nearest neighbour) classifier is used to find the respective classes from the normal videos. Experimental results from large crowd behaviour video database shows the superior performance as compared to the baseline and state-of-the-art methodologies.

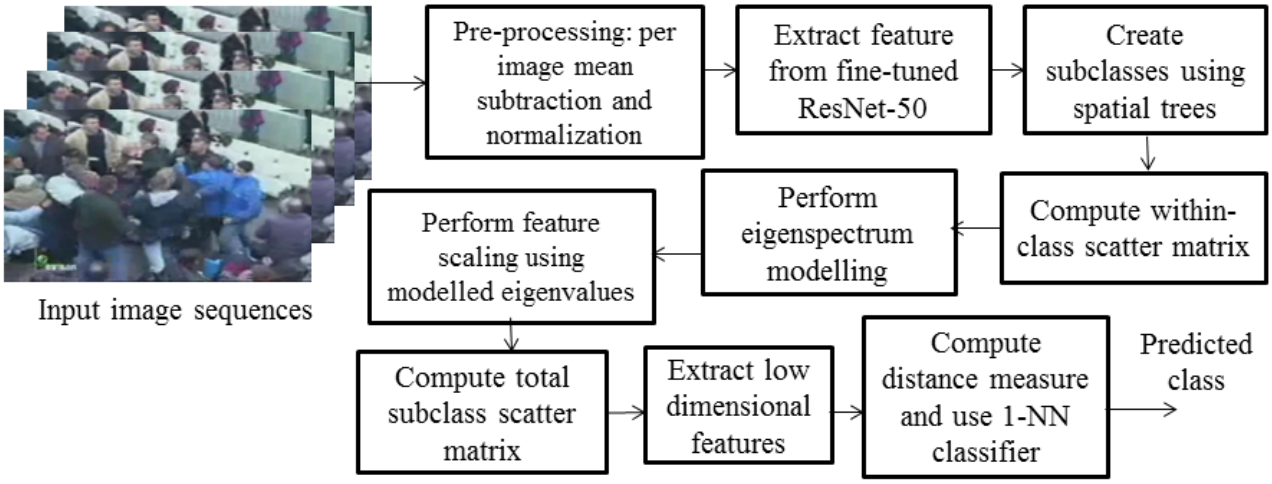

Figure 4: Block diagram of the proposed video analytic framework

Since we are using pre-trained (such as deep residual network with 50 layers (ResNet-50)) model, while training the network, each image is re-sized into a fixed size of 224 X 224 pixels. The images are then normalized by subtracting channel-wise mean intensity values from corresponding individual image channels in the colour image. Each channel of the colour image is also made to unit standard deviation. The ResNet-50 pre-trained network is fine-tuned using our dataset by retraining the whole network. Training only the higher-level layers helps in extraction of features specific to the present dataset. It also avoids overfitting, since there is a fewer number of images compared to ImageNet (http://www.image-net.org/) which contains millions of images. Additionally, it avoids overfitting and increases the robustness. Data augmentation is used in our framework to increase the performance of the network via a set of random transformation to the images. The augmentation operators include rotation (90, 180 and 270 degrees), translation, horizontal and vertical flipping of the original images. This is equivalent to adding new data in our training phase, so that trained models learn better representations and subsequently perform better recognition.

In order to model the large variances that appear in the intra-class (for example, within fight sequences of different scenarios) we used spatial partitioning trees (such as principal component analysis, k-dimensional (KD), and random portioning) are used to form subclasses. This helps us to approximate the underlying distribution with mixture of Gaussians and perform whole space subclass discriminant analysis among these subclasses. Our work uses a regularization methodology that enables discriminant analysis in the whole eigenspace of the within-subclass scatter matrix. Eigen feature regularization is performed using the method described in [Mandal2015]. Low dimensional face discriminative features are extracted after performing discriminant evaluation in the entire eigenspace of within-subclass scatter matrix.

Crowd behaviour analysis has inherent varying spatial-temporal structure. Different crowd groups would show the same behaviour (e.g. a protest) differently and even the same group is not ever able to produce the same behaviour exactly. For comparison between two behaviour events of different lengths we use dynamic time warping, which performs a time alignment and normalization by computing a temporal transformation allowing two behaviours to be matched. In the experiments of this work, Cosine distance measure and the first nearest neighbourhood classifier (1-NNK) are applied along with DTW to test the proposed approach for crowd behaviour recognition.

3.3.2 Experimental Results

We tested our approach on the world's largest crowd behaviour recognition database [Shao2015], comprising of 10,000 videos from 8,257 scenes. This database is constructed to answer “where", “who" and “why" (WWW Crowd database) questions and there are 94 crowd-related annotated attributes, such as stadium, concert, stage, fight, mob, parade, etc, to describe each video in the database. We selected a few normal crowd videos and 4 violent behaviours videos, such as fight, protest, mob and protester from this large database. As per the existing protocol [Shao2015], WWW database is randomly partitioned into training, validation and test sets in the ration 7:1:2, videos are converted to image sequences at 25 frames per second. This gave us a total of 219,094 number of images, the distribution for each of the selected attributes are shown in Table 1 below.

Table 2: Selected attributes and their images from the WWW crowd database.

Attributes	Normal	Fight	Mob	Protest	Protester
# Images	15,631	14,059	14,609	87,241	87,554

Figure 5 (left) shows the recognition rate (%) versus the number of features used for classification of various crowd behaviours. We have also implemented the baseline approach, which uses the features from the ResNet-50 model fine-tuned using the images from the WWW crowd database. It is evident that our proposed learning framework outperforms the baseline method for all the violent activities. For protest and protester, the recognition rate improvement is small, but for fight and mob our approach outperforms the baseline method significantly.

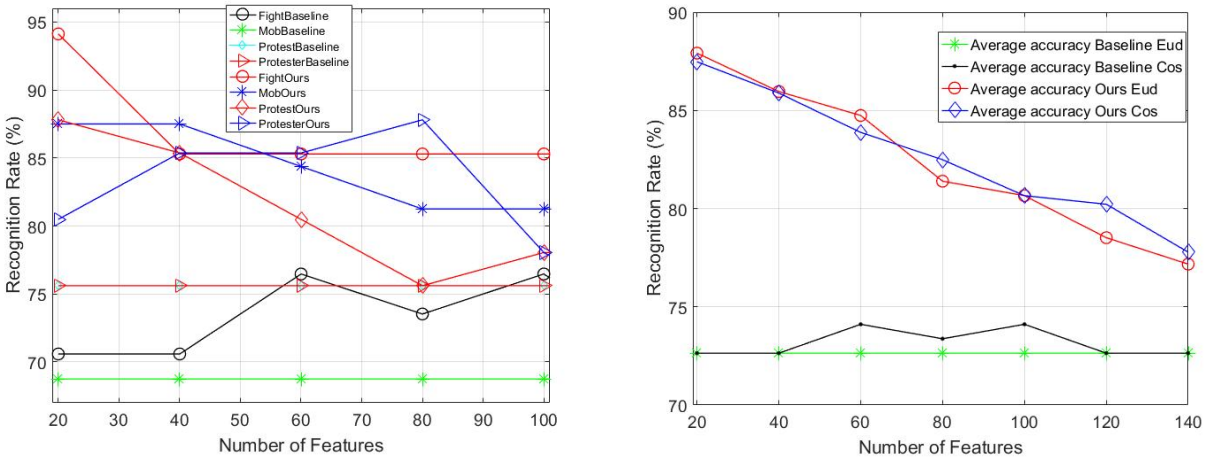

Figure 5: (Left) Recognition rate (%) of the selected attributes on WWWcrowd database. (Right) Average recognition rates (%) of the selected attributes on WWWcrowd database.

The average recognition rates (%) of all the violent behaviours are shown in Figure 5 (Right). It can be clearly seen that our proposed method outperforms the baseline method for both the Cosine (Cos) and Euclidean (Eud) distance measures. The performance gain is higher for small number of features.

3.3.3 Summary

Our work proposed a fine-tuned deep convolutional neural residual network framework that creates subclasses within feature maps of each class using spatial partitioning trees. Eigen feature regularization is used to weight the features of the whole within-class eigenspace of the crowd behaviour videos. This has helped to model the variances appearing from the intra-subclass or within-class variance information. Low dimensional discriminative features are extracted from total class scatter matrix to represent the various crowd behaviours videos. Dynamic time wrapping is applied on the cosine distance measure to find the similarity measure between the two videos. Experimental results from a large crowd behaviour video database shows that with proper training our algorithm, on an average, could correctly recognize the four violent behaviours (fight, protest, mob, protester) from the normal videos with a success rate of >87% using only 20 features as shown in Figure 5 right.

3.4 Fight Detection

Another example where modelling of complex dynamics is utilised in MONICA is in the Fight Detection module. In order to detect real-time events in video streams, the fighting module was developed. The proposed system follows a modular architecture with different components being responsible for operations such as video capturing and event detection. It should be noted that computer vision algorithms constitute the core of the system and video analytics components - procedure for extracting regions with motion, the trajectories` computation mechanism and feature extraction process - are the major characteristics of the module. More details on this has be given in deliverable “D5.1 - Sensor Analytics and Information Fusion 1 - Section 3.1.3.2”.

3.5 Data

Within this section an overview will be provided of the current state of available data for use in modelling complex dynamics within MONICA. It must be stressed that when the concept of *data* is discussed, it is often references two specific concepts, firstly the raw video or image data produced from the various camera setups in the pilot sites, the second and more troublesome is the annotation or ground truth associated with those videos and images. These annotations can take many forms, in the context of modelling complex dynamics this will be descriptions detailing what is in the raw footage in line with what an algorithms desired output will be. For example, Figure 6 illustrates an example frame that would be evaluated by the algorithms (and related models) in the Processing Node. To effectively train those models the annotation data is needed to teach a model what it is looking for. Annotations can include bounding boxes highlighting specific areas of a scene, in this case location of people or child. Additionally, annotations can also include data that summarises a scene, such as the global count.

It is worth noting here that the outputs of the MONICA system, namely the messages sent from the various components could also be utilised as annotations for data. For example, the output of the crowd counting algorithms could be used to annotate scenes that were not used when training the model utilised by that algorithm. This would allow for the creation of more data which could be used in the future training of models, however it is important to make the distinction between ground truth data and annotations created from outputs of algorithms. Put simply there is no guarantee of accuracy with the latter and as such should be used carefully so as not to reduce the effectiveness of an algorithm because the model it uses is trained on inaccurate data.

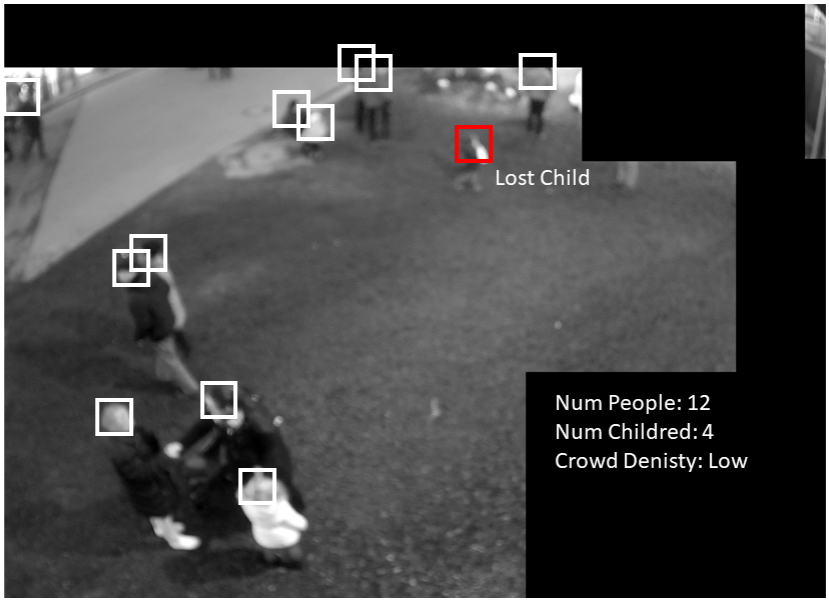

Figure 6: Example image from Winterdom Hamburg with some possible annotations.

Without a large amount of well annotated specific MONICA data, algorithm performance could be affected. As is often the case in computer vision and wider modelling tasks, results are directly linked to the quality and similarity of data to the final application. Below is some quantitative data detailing what WP5 and the wider consortium has access to.

Table 3: Description of currently available MONICA specific data.

Provider	Description	Format		Annotations
YCCC/LR	Pan/tilt/zoom CCTV footage of entrance and exit locations during a Leeds Rugby match	Currently in a proprietary format and not suitable for processing. However, review of the video data is possible with specific application		N/A
City of Bonn (BONN)	Footage taken during the Pützchens Markt from 10 cameras around the site at various times of the day. Both Analogue and digital CCTV cameras.	Analogue 704x576 25fps H.264 2523kbps	Digital 640x360 25fps H.264 1739kbps	N/A
Free and Hanseatic City of Hamburg (FHH-SC)	Footage from 2 cameras taken during the Winterdom event.	2048x1536 25fps H.265 80000kbps		N/A
Comune Di Torino (TO)	Footage captured from 6 cameras during the Kappa Future Festival 2017	1920x1024 30fps ITU H.264 39718kbps		Some preliminary manual annotations preformed

Although the data that has been made available is a great start, the principle concern here is the lack of annotations. This is a laborious and sometimes challenging task, one which a single partner or WP cannot really be responsible for. As such the suggestion has been made to utilise so called 'crowdsourced' data annotation services such as Amazon Mechanical Turk or Crowdflower, to annotate the available MONICA data for use in training of the algorithm models. This is subject to ethical approval and will require extensive discussion as to the quantity of data required and the timeframe in which to have it completed by. However, this process is vital and will need to be done if MONICA data is to be used at all.

The use of these crowdsourcing tools is a logical answer to the annotation problem, a defined set of images and videos from across the available data can be sent off and the various annotations defined. Each image can be annotated several times to ensure the annotations returned are representative and accurate. The principle concern with this is the financial cost of these tools. It is the belief of WP5 that this cost should be distributed amongst the pilot sites. The result of this process will be a highly general and scientifically useful dataset covering the crowd behaviour and counting problems. This would lead to new and exciting research into the areas, providing an important resource.

In the absence of or (at this stage) limited ground truth (labels) annotated data in MONICA project, we plan to use publicly available databases. In our selection we have noticed that most of the existing public crowd datasets contain only one or two specific scenes and the second largest one [Shao2014] provides only 474 videos from 215 crowded scenes. We propose to use the largest and most diverse publicly available crowd behaviour recognition database [Shao2015], comprising of 10,000 videos from 8,257 scenes. While making this database, the researchers did not include any keywords referring to specific places (such as Time Square, Grand Central Station) but used functionalities of places or generic cues (such as landmark, street, pedestrian, walking, mob, swim, stadium, stage, outdoor, indoor, ceremony) instead. The key words were searched in many public video search engines like Getty Images, Pond5 and YouTube. Some examples representation violent crowd behaviours are shown in Figure 7.

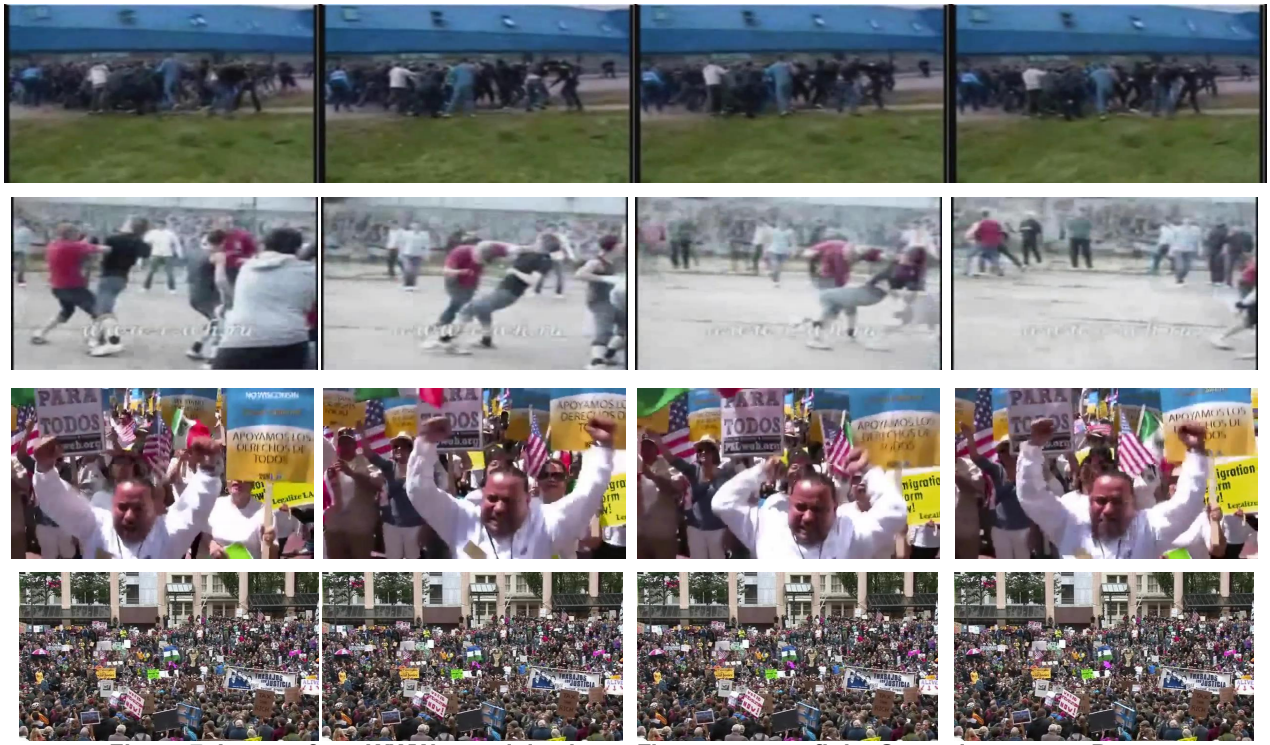

Figure 7: Images from WWW crowd database. First two rows: fight, Second two rows: Protester

This database is constructed to answer “where is the crowd", “who is in the crowd" and “why is crowd here" (WWW Crowd database) questions and there are 94 crowd-related annotated attributes, such as stadium, concert, stage, fight, mob, parade, etc, to describe each video in the database. 16 annotators were hired to label attributes in the WWW crowd database and another 3 annotators to refine labelling for all videos. The attributes used in this database are commonly seen and experienced in our daily lives, so these annotations do not require special background knowledge. In every round, a 10 seconds video clip was shown to each annotator and was asked to label at least one attribute from each attribute list without time constrained. In our opinion this database is very close to our requirement, although the video quality is not that high as compared to the current state-of-the-art MONICA cameras.

4 Information Retrieval

4.1 Introduction

Information retrieval is required for in-depth post-event analysis of incidents and can be a time consuming, manual task. During an event, captured video data is stored in 'video storage' via a Video Management System (VMS) in the edge layer (see Figure 1) to be reviewed or processed later. Additionally, meta-data generated during the event from the various WPs is also stored, for example, messages from WP5 regarding crowd behaviour or fight alerts from the fight detection algorithm. Information retrieval is a search through video storage utilising this meta-data to find specific video data relevant to the desired task. Tasks like, finding all time episodes in the stored video where the crowdedness is higher than a threshold, or detecting all instances of a specific event (like a fight).

In addition to the task of recalling existing video based on already created meta data is the concept of video mining, where by additional correlations and patterns, previously unknown, can be extracted in post processing techniques. In MONICA we are interested in analysing video clips to come to conclusions about some queried behaviour. A video mining module will be part of the edge layer and has immediate access to both the stored data and the 'meta-data storage'. The main goal of this module will be analysing these large amounts of data to find correlations, interesting patterns, and insights. It helps to explain and understand past behaviour by finding interesting, human-interpretable patterns that describe the data. In MONICA, the video mining module would be a post-process which can be called any time after the event. All the additional knowledge derived through this module can then be stored in the ‘meta-data storage’ for future Information retrieval requests or further processing.

To contextualise the role of task 5.4 within the MONICA project, Table 4 outlines the possible use of the information retrieval systems within the current MONICA solutions, the linked MONICA use cases, along with a brief explanation of how the process could be utilised. Currently, the concept of mining recorded video footage for further meta data post event is not reflected in the current solutions and use cases. As such, further discussion is needed to identify if this is required.

Table 4: Task 5.4 Information Retrieval within MONICA

MONICA Solution	MONICA Use Case	Use of Task in context
Person or object retrieval	UCG4 analysis for lost person	Person/object retrieval model will analyse the recorded and stored video based on a given time window and try to provide useful feedback for the lost procedure.
Post-event retrieval	UCG7 / UCG8 event or incident analysis	Event or incident retrieval model can analyse the stored videos based on the event time window and provide some details of the event.

4.2 Architecture

Within this section an architecture overview is given of the components relating to T5.4 Information Retrieval.

Figure 8 outlines the architecture components utilised as part of Task 5.4. As the MONICA system is running on the video feeds, their respective frame data will be recorded and stored locally on site. Additionally, the meta data and messages produced by the various MONICA components and sent up to the MONICA cloud will also be logged. The action of creating an information retrieval request is done so from the high-level MONICA architecture, whereby the Common Operational Picture (COP) provides the interface from which a search request can be generated by a user. As the COP has access to all the MONICA events and messages (meta data) created, and all these events have timestamps, this database can be utilised to create the search queries. This query is then sent though to a Video Management System (see Section 4.3.1) in the form of specific timestamps and camera feeds allowing the raw data to be retrieved and sent up to the MONICA cloud for review. Additional meta data can be generated as a result of post event mining processes, which could identify previously undetected events as a result of more in-depth comparison or analysis. Please note that at time of delivering this, the architecture for post processing and video mining is

unconfirmed due to the lack of requirements and wider discussion within the consortium. As the proposed architecture is designed to b

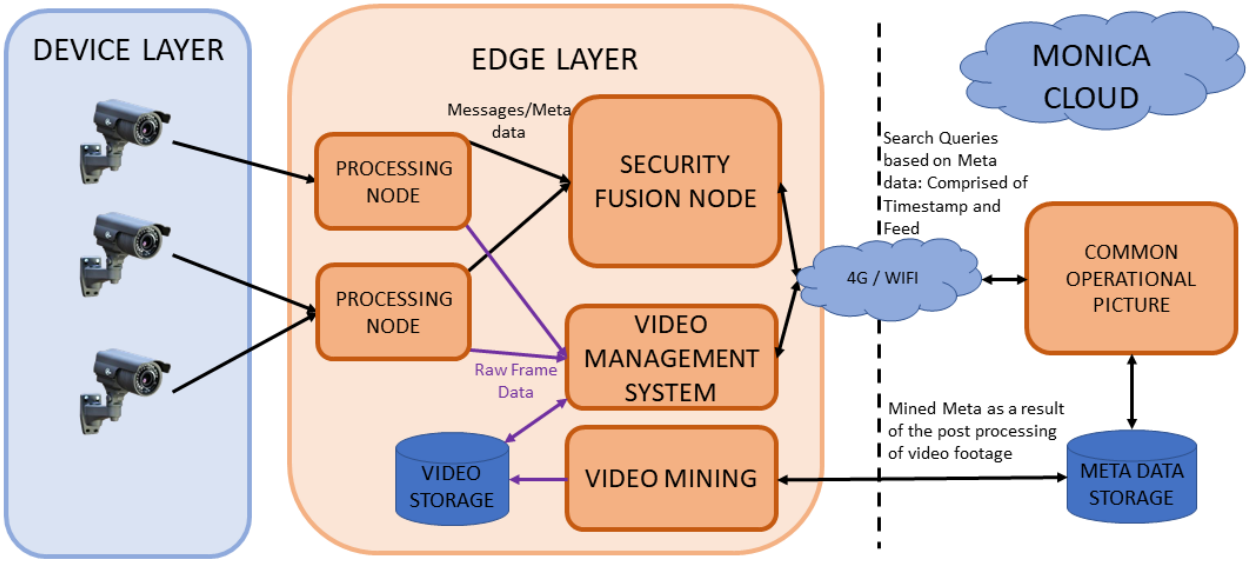

Figure 8: Extrapolated view of the WP5 components with respect to Information retrieval requests.

4.2.1 Storage requirements

Since there will be a potentially large number of cameras running on a site at one time it is critical to understand where the data is to be stored for the purposes of data retrieval or data analytics. Two main locations have been considered within MONICA:

- **Locally at the pilot site:** the main requirement that needs to be met first is where to store the data locally for easy retrieval and analytics. This location depends on the existing infrastructure at the site. For example, if the site is well structured to accommodate storage of data, then the challenge is whether MONICA has access to the local storage unit or not. Otherwise, Network Attached Storage (NAS) will need to be deployed. NAS is a file-level computer data storage server self-contained solution for storing or sharing data over a network and providing access to a heterogeneous group of clients.
- **The Cloud:** The cloud is expected to store the meta data produced by the various MONICA components, while the raw data (Images/Video) will be stored locally at the pilot site. This is for two reasons. The first, is for privacy and ethical considerations, the second reason is to save the network bandwidth which is needed to accommodate other communications needs. The data will be sent to the cloud through an internet connection that is provided through an internet gateway at the pilot site.

In general, the video storage unit must provide:

- **Rapid Deployment:** The fact that there are multiple sites forces the deployment to be flexible in term of installation. This requirement calls for ready-to-go storage unit that is plug-and-play and could be used in all pilot sites and all events.
- **Central Management:** To avoid unauthorized access and ensure simple management, the storage unit has to be located in the control room within the venue. The word “central” refer also to the star topology of the network to and from the storage unit.
- **Reliability and Availability:** Since the data will be stored and used later for analytics, it is critical to have a backup unit in case the data in the main storage unit failed to be loaded. A backup unit is not necessary, but important to be considered. The main storage unit should be connected to the backup unit and data should be transferred in real-time. This feature reflects the need for reliable data storage and retrieval.

The challenges though are the geographical spreading of the components and hence the deployment needs to be scalable to accommodate the high number of deployed components. Another consideration is the prevention of unauthorised access to the video data being stored. The video storage unit should be assigned a static IP address within the DHCP backbone network or be addressed by URL for remote access. This may

introduce a single point of failure; however, this should be more manageable and secure in terms of access. The IP address will need to be known for the different sensors gateway with in the pilot site.

On the other hand, retrieving data from the cloud or the storage unit is slightly different from sending the data to it. The retrieving of the data happens between the data analytics server and the storage units which should be in close proximity so as to facilitate fast data retrieval and processing.

4.3 Information Retrieval

As outlined in the architecture and introduction to this section the stand-alone task of retrieving data from video storage based on a search criterion is well defined. To provide further context of how information retrieval is currently utilised a brief example of the work flow currently implemented at the Leeds pilot site is outlined below along with an example use case.

The Leeds Stadium is monitored by many cameras; some record the matches, and some record the rest of the stadium such as seating, etc. The match event is recorded across all the cameras and monitored from a control room. The control room remotely direct and zoom in the closest camera to any reported incident. With the help of the cameras, the control room will then locate and direct the stewards to any reported incident. The incident details and timestamps are manually logged.

If there is a need to review footage post event, the current information retrieval process is performed manually, accessing the footage from their existing video management system using timestamps extracted from the incident log. Due to the use of Pan/Tilt/Zoom (PTZ) cameras, often the footage recorded is of a focused location or event, this is useful when responding to an event, however limits the information that can be extracted post event, as the constrained view may well miss important details in the surrounding area.

A Current Example Use Case:

An altercation resulting in a fight takes place between two members of the public during a stadium event (rugby game), as a result one of the people is arrested. During the altercation the security staff report the incident to the control room, the incident type, time and location are added to the incident log. As a result of the arrest, the police ask the stadium to provide any footage from that event which may be of relevance to the incident, for use later as evidence. To aid this search the police provide a description of the people involved and the rough time the event took place.

As the stadium staff were involved, and the incident logged, an accurate record of when the incident took place is available and can be used to pull up any footage of that area during that time. The footage must then be manually reviewed to ascertain if any useful footage has been captured.

4.3.1 Video Management Systems (VMS)

Given a typical site with hundreds of cameras, it is often necessary to have a system to manage these cameras in a centralised and organised manner. While each camera hosts a web server that facilitates viewing of the video stream, it's not practical to manage a large number of video feeds without some centralised management system. Moreover, it is often desirable to record video for the purposes of review, evidential export (in the case of an incident) and process evaluation (e.g. how can incidents be handled better in future).

Such a system is referred to as a Video Management System (VMS), or sometimes a Networked Video Recorder (NVR). These systems allow the user to rapidly review any camera, organised in a logical way. For example, the cameras may be overlaid on a map view, so it's easy for operators to understand the geospatial relationship between cameras. VMS systems also permit rapid review of recorded video, often by synchronising multiple cameras so an evidential story can be constructed around a specific incident (e.g. a concert-goer entered the stadium on camera 192, where a good face shot was recorded, then proceeded to the stand via cameras 062, 063 and 064, before being observed initiating a disturbance on pan-tilt camera 251).

Since VMSs and NVRs are readily available off-the-shelf, many pilot sites will already have some systems installed. In other cases (e.g. MOVIDA), there may not be a traditional VMS as the recording is conducted locally, within the camera. Some sites may not have any existing infrastructure, and everything will need to be provided as part of the MONICA deployment.

This provides a challenge in that the Application Programming Interfaces (APIs) of VMS and NVR systems are not standardised, so it may be a significant task to integrate to a wide range of existing VMSs. The most

pragmatic approach may be to constrain the video mining activity to only operate with a restricted set of existing systems. For this purpose, it is recommended that any new installations under the MONICA umbrella use a standardised VMS that exposes the same API. One possible recommendation is the Geutebrueck G-Core range of VMS/NVRs, as they have already been deployed for the Rhein-in-Flammen event during early pilot data collection activities in 2017.

The video processing framework within MONICA is currently only able to process live video. Hence, it will be necessary to add additional functionality such that this live video can be stored and used in the Information Retrieval framework, as well as provide some interface to allow specification of the start/end times, and source channel. Due to the amount of work required to integrate against multiple VMSs (due to the reasons outlined above), it is recommended that an example integration is performed against the standardised VMS selected for the MONICA project (e.g. Geutebrueck G-Core).

4.4 Image/Video Mining

As previously defined the action of Video Mining allows the post-event extraction of further insight into an event with the possibility of generating additional meta-data from which relevant video data can be retrieved by a user. Video mining is a strictly *post-event* or *not real time* process and allows for the use of algorithms that otherwise would be unable to be run on the incoming video streams directly. Importantly video mining is not required for Information Retrieval in its basic form and serves only to augment already collected information

Extracting explicit semantic information has been extensively investigated such as object detection, structure analysis and event detection. Video is a media embedding visual, motion, audio, and textual information. Video mining is extracting information from video data using image and video processing techniques, like detecting special events, or finding similar video clips. Transforming low-level features of video objects into high-level semantic information and video patterns from massive amounts of video data are goals of video mining. Implicit, previously unknown, and potentially useful patterns and knowledge can assist pilots in knowledge acquiring, decision making, and security management.

The overall goal of the data mining process here is to extract information from the video data and transform it into an understandable structure which can be stored in 'meta-data storage'. Aside from the raw analysis step, it involves database and data management aspects which are handled separately by the Video Management System module.

One example technique, explored in detail below, which can be applied post event is saliency. This aim being to detect distinctive regions in an image or video frame which draw human attention. Using this property of a video allows for the retrieval of similar images from large data, or abnormality detection using the most/least salient areas.

4.5 Saliency

4.5.1 Introduction

Visual saliency is a broad term that refers to the idea that certain parts of a scene are pre-attentively distinctive and create some form of immediate significant visual arousal. Saliency actually tries to imitate how a human eye identifies important objects in the scene and is typically based on a simple fundamental: the contrast between an object and its neighbour. Therefore, saliency is the technique of finding salient regions of the image which are those regions which seem visually important to us.

The main goal of this deliverable part is to provide a technique to discover a high-quality visual saliency model which can be trained with a deep learning framework, which has had many successes in visual recognition tasks. Deep learning techniques usually obtain high-level features to detect salient regions in a scene. Here, we use both high-level and hand-crafted features under a unified deep learning framework which the low-level features can provide complementary information to enhance the performance of saliency detection that utilizes only high-level features.

4.5.2 Algorithm Description

The overall pipeline of the method is illustrated in Figure 9. Lee et al. (Lee2016) presented a technique for saliency detection in images which is called ELD (Encoded Low Level Distance map). They introduced the encoded low-level distance map (ELD-map) which directly encodes the feature distance between each pair

of superpixels (a group of connected pixels with similar colours or grey levels) in an image. ELD-map encodes feature distance for various low-level features including colours, colour distributions, Gabor filter responses (Li2010), and locations.

This algorithm utilizes a superpixel-based approach for saliency detection. To segment an image into superpixels, the Simple Linear Iterative Clustering (SLIC) (Achanta2012) algorithm is used. In SLIC, segmented superpixels are roughly regular, therefore it provides control on the number of superpixels.

After superpixel segmentation, the initial hand-crafted low-level features of each superpixel are calculated, and the superpixel representation is converted into a regular grid representation according to their occupying area in each cell as illustrated in Figure 10. This regular grid representation is efficient for a Convolutional Neural Network (CNN) architecture because we can convert images with different resolutions and aspect ratios into a fixed size distance map without resizing and cropping.

ELD-map uses deep learning as an auto-encoder to encode these low-level feature distances by multiple convolutional layers with 1 × 1 kernels. The encoded feature distance map has strong discriminative power to evaluate similarities between different parts of an image with precise boundaries among superpixels. The ELD-map and the output of the last convolutional layer from the VGG-net (VGG16) (Simonyan2014) are concatenated to form a new feature vector which is a composite of both high level and low-level information (see Figure 9). The VGG-net is a deep convolutional neural network for object recognition developed and trained by Oxford's renowned Visual Geometry Group (VGG). Using this new feature vector, the saliency of superpixels can be precisely estimated. Without any post-processing, this method generates an accurate saliency map with precise boundaries.

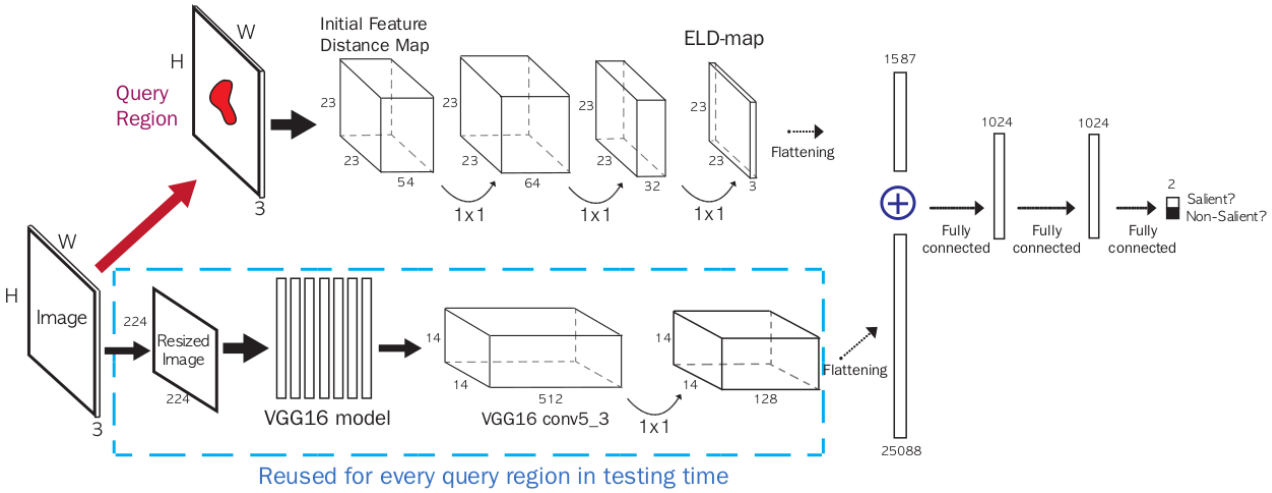

Figure 9: Overall pipeline of the method. The ELD-map is computed from the initial feature distance map for each query region and concatenated with the high level feature from the output of the conv5_3 layer of the VGG16 model.

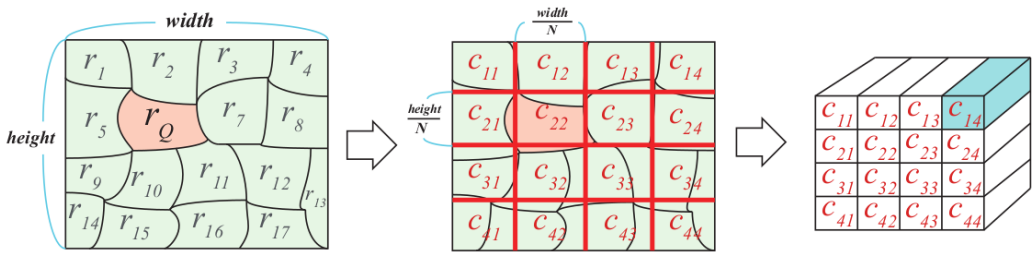

Figure 10: Visualization of the construction process for the initial low level feature distance map. Each grid cell, which represents uniformly divided area of an image, is described by the features of the superpixel that occupies the largest area of the grid cell.

4.5.3 Experiment Results

We conducted experiments to identify regions of saliency within some crowded images. Such salient regions are used to extract descriptions which could then be used to solve vision problems necessitating matching or correspondence. In this section, we illustrate the results of applying the described technique to a range of different types of images (see Figure 11). These results show the impressive performance on images with salient objects and crowded backgrounds.

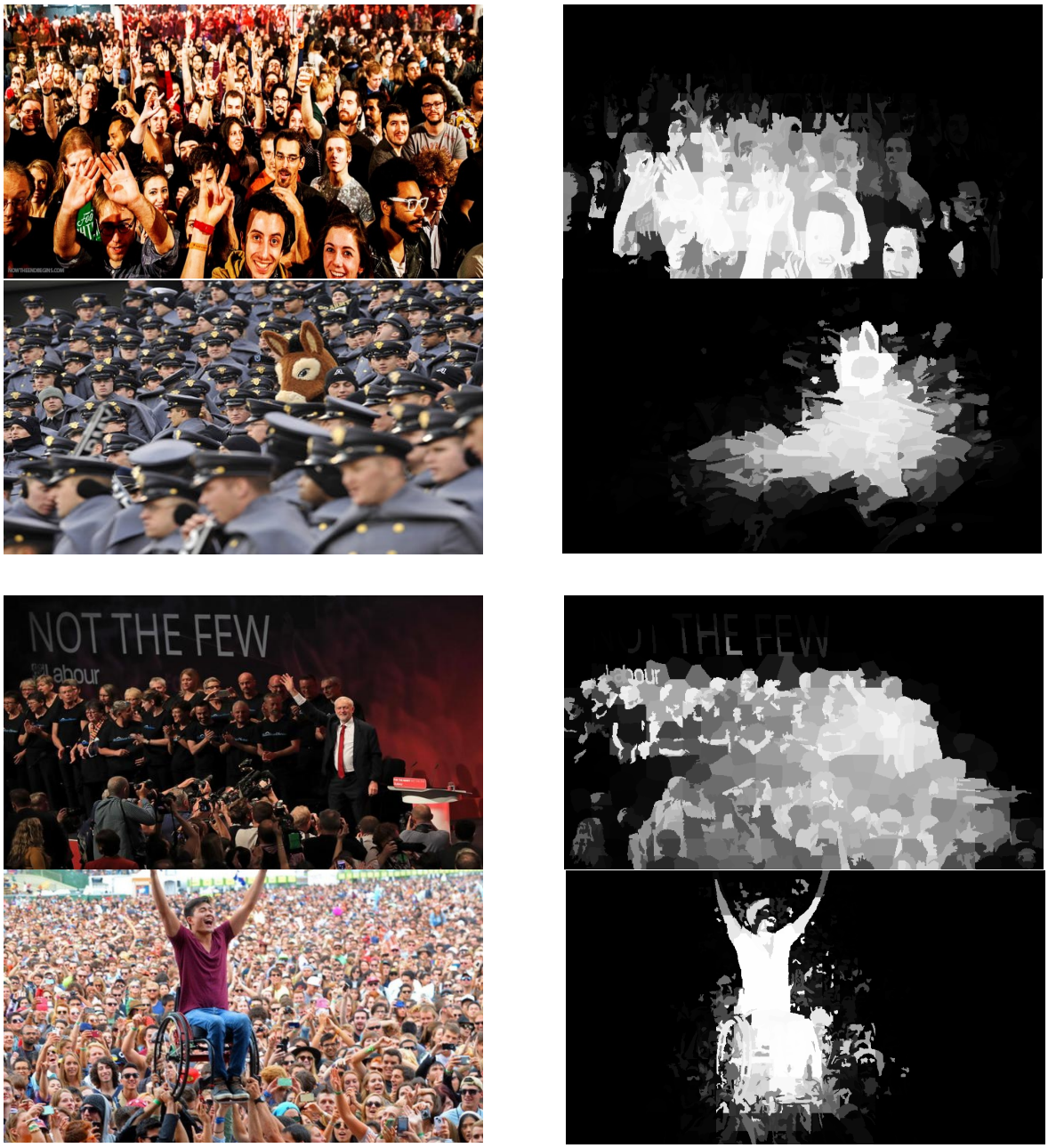

Figure 11: Example results of images with saliency applied.

5 Conclusion

In summary this document provides an overview of what Tasks 5.3 and 5.4 constitute as well as to demonstrate the current state of work within these areas. The context of where the tasks themselves fit into MONICA has been provided by demonstrating the various use case and solutions in which these tasks play a role. Additionally, an overview is given of the architectures and therefore how they fit with the upcoming deployments.

As the task Modelling Complex Dynamics is mostly a conceptual task and has much overlap with T5.1 and T5.2, as well as general description of what a modelling process entails, an in-depth example is also given utilising crowd behaviour analysis as a subject. This also serves to provide information on additional data that can be extracted from the infrastructure attached to WP5 whilst addressing some of the use cases raised in MONICA.

The concept of Information Retrieval is explored and divided into its two principal components. The relatively well-established problem of retrieving specific video footage based on a set of criteria is a process already addressed with a number of off-the-shelf solutions. The integration of MONICA with these solutions is looked at and a rudimentary architecture and development road map provided. The concept and benefits of a video mining system is also evaluated, with an example algorithm (Saliency) that could be run in a post process to provide additional meta data given in detail.

It is important to stress at this point that the use cases for task 5.4 are still to be defined and will require further discussion before any key decisions are taken, and to note that this document only suggests ideas on how to proceed.

6 List of Figures and Tables

6.1 Figures

6.2 Tables

7 References

(Li 2015) Teng Li, Huan Chang, Meng Wang, Bingbing Ni, Richang Hong, and Shuicheng Yan, “Crowded scene analysis: A survey,” IEEE Trans. Circuits Syst. Video Techn., vol. 25, no. 3, pp. 367–386, 2015.

(Julio 2010) Julio Cezar Silveira Jacques, Soraia Raupp Musse, and Cl´audio Rosito Jung, Crowd analysis using computer vision techniques,” IEEE Signal Process. Mag., vol. 27, no. 5, pp. 66–77, 2010.

(Shao2015) Jing Shao, Kai Kang, Chen Change Loy, and Xiaogang Wang, “Deeply learned attributes for crowded scene understanding,” in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, pp. 4657–4666.

(Mandal2015) B. Mandal, Liyuan Li, V. Chandrasekhar and Joo Hwee Lim, "Whole Space Subclass Discriminant Analysis for Face Recognition", IEEE International Conference on Image Processing (ICIP), pp. 329-333, Quebec city, Canada, Sep 2015.

(Solmaz2012) Berkan Solmaz, Brian E. Moore, and Mubarak Shah, “Identifying behaviors in crowd scenes using stability analysis for dynamical systems,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 10, pp. 2064–2070, 2012.

(Zhou2012) Bolei Zhou, XiaogangWang, and Xiaoou Tang, “Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents,” in 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012, 2012, pp. 2871–2878.

(Shao2014) Jing Shao, Chen Change Loy, and Xiaogang Wang, “Scene-independent group profiling in crowd,” in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014.

(Simonyan2014) K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR,abs/1409.1556, 2014.

(Achanta2012) R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI), 34(11):2274–2282, 2012

(Lee2016) Lee, Gayoung, Yu-Wing Tai, and Junmo Kim. "Deep saliency with encoded low level distance map and high level features." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 660-668. 2016

(Li2010) Weitao Li, KeZhi Mao, Hong Zhang. Selection of Gabor filters for improved texture feature extraction. Image processing (ICIP). 2010