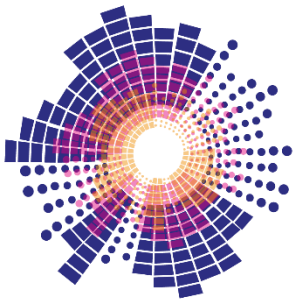


MONICA

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

Information Retrieval for Post-event Analysis

Date:

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: D5.4_Information_Retrieval_for_Post_event_Analysis_Rev1.0
Document version: 1.0
Document owner: KU

Work package: WP5 – Security Closed Loop Systems
Task: T5.4 – 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	Mahdi Maktabdar (KU)	2019-06-20	Initial Draft
0.1	Hamideh Kerdegari (KU)	2019-06-20	Initial Draft
0.1	Hajar Sadeghi (KU)	2019-06-20	Initial Draft

Internal review history:

Reviewed by	Date	Summary of comments
Sebastien Carra (Acoucite)	2019-07-25	• General purpose and summary: I understood that the purpose is to identify anomalous crowd behaviour and that the accuracy of the detection is about 86%. However it is difficult to know if the process developed will be applied for the next demonstrations and how it will be displayed for the technicians: rowdiness? Alerts for the four classes? Is there a basic name for each class? • Indicators: usually the event manager works with the density (people/ square meter) or number of people on the area of interest. Which is the global density for a rowdiness of 100 (to have an idea)? • p16: annotation is always good in a piece of code

Index:

1 Executive Summary

This deliverable documents the current progress relating to tasks T5.4 Information Retrieval for Post-event Analysis of the MONICA project pilots. Below describes the task T5.4, its aim, expected outcomes and the steps taken to develop and deploy this deliverable.

Information retrieval in video signal is often required for an in-depth post-event analysis of incidents and accidents. Analysis of crowd behaviour in public places is a critical objective for video surveillance. We are often reminded how time consuming it is for police officers to search through hours of video looking for an incident. In WP5, we are aimed to automate this task with the help of artificial intelligence, computer vision and machine learning in particular deep learning techniques. Automated detection of anomalous crowd behaviour is a challenging task. Numerous types of activities such as a sudden rush, overcrowding, loitering and stampede, can be categorized as anomalous crowd behaviour. Detection of such activities in video signal, demands sophisticated yet efficient computer vision models, capable to extract the underlying spatial and temporal features that correlate to anomalous crowd behaviour.

In this regard, the aim of this deliverable is to design, develop and deploy a video analytic and information retrieval system for post-event (in this case a MONICA pilot site) anomalous crowd behaviour analysis. Work Package 5 is intended to use a combination of existing and newly-developed computer vision and machine learning algorithms to carry out this deliverable and to achieve the goals set out by the generated user requirements. The proposed system exploits state of the art deep recurrent neural network in its core, continuously monitors and analyses the crowd behaviour in both spatial and temporal spaces for abnormalities such as sudden rush, overcrowding, loitering and stampede. The system is capable to flag these abnormalities in near real-time fashion.

The proposed system can be extremely beneficial for police, organizers and incident response team in the management of crowds in very large-scale events. Information retrieval for anomalous crowd behaviour (D5.4) complements other deliverables (D5.1 Sensor Analytics, D5.2 Information Fusion and Node and Node Cluster, D5.3 Modelling of Complex Dynamics) in WP5 and offers a comprehensive solution for crowd monitoring by means of algorithms such as gate counting, crowd counting, density estimation and localization, fight detection, object detection and anomalous behaviour detection and fulfils WP5 (security closed loop systems) aims and objectives.

This document describes the steps taken to design, develop and integrate video analytic and information retrieval system into the MONICA platform.

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. T5.4 in particular aimed to design, develop and deploy a video analytic and information retrieval system for anomalous crowd behaviour analysis in large-scale outdoor events (in this case a MONICA pilot site).

This document describes how T5.4 fits into WP5 as well as wider MONICA architecture. It also points the MONICA use cases that can potentially benefit this deliverable. Furthermore, a detailed description on data preparation and annotation, model development and integration is given in this document.

2.2 Deliverable contents and structure

This document is split into four sections including modelling the crowd dynamic, data preparation and annotation, video retrieval model deployment and integration of information retrieval model into the MONICA platform. An overview of MONICA and WP5 architecture as well as, the impact and implication of this task to 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 the Crowd Dynamics for Information Retrieval

3.1 Introduction

Information retrieval in video signal is an extremely challenging task and often required for an in-depth post-event analysis of incidents and accidents. Analysis of crowd behaviour in public places is a critical objective for video surveillance. We are often reminded how time consuming it is for police officers or forensic investigator to search through hours of video looking for an incident.

Although crowds are made up of independent individuals, each with their own objectives and behaviour patterns, the behaviour of crowds is widely understood to have collective characteristics which can be described in general terms. The management of crowds' dynamics and behaviour in large-scale outdoor events like football matches, pop concerts, festivals is a substantial problem with serious consequences for human life and safety and for public order if it is not managed successfully. When crowd density exceeds a certain level (this level being dependent upon the collective objective of the crowd and the environment), danger may occur for a variety of reasons. Physical pressure may result directly in injury to individuals, stampede or physical damage to the environment (for example the collapse of barriers). Aside from the crowd density, the mood of the crowd may pose serious threat to the crowd safety. For example an angry crowd of two opposing football supporter or frightened crowd of a terrorist attack or a natural disaster.

These security requirements have led to a continuing increase in the number of deployed closed circuit video cameras, so that more and more public areas are subject to surveillance. However, the successful use of these facilities presents some problems. In traditional CCTV systems, human observers are often employed to watch the TV monitors, and it is often not considered cost-effective to have one monitor per camera and it is even less likely to have one observer per monitor. This issue significantly decreases the effectiveness of the surveillance systems.

Therefore, it is clear that there would be a substantial advantage if some form of image processing system could be applied to the video signal, to automatically spot the anomalous crowd behaviour as they arise, and to trigger suitable response in reasonable time. Classic computer vision and machine learning techniques are struggling with overwhelming complexity of crowd behaviour analysis models. These algorithms are designed to detect the presence of a human in an image or video should have some previous definition of what constitutes a person, this definition is statistically (In the form of training data) or mathematically described within the model. 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. In large crowd, complexity of such models exponentially increases, leads to highly complex multi-dimensional model. Moreover, traditional mainstream hardware solutions were unable to keep up with the complexity of the crowd behaviour analysis models and further crippled the effectiveness of these solutions.

Unlike classic computer vision and machine learning techniques, deep learning approaches are capable to generalize the highly complex crowd behaviour and tackle many environmental challenges such as occlusions, complex backgrounds, non-uniform distributions and variations in scale and perspective, variation in lighting condition and undesirable resolution. In spite of the advantages of these techniques, they demand for extremely large annotated ground truth data to deliver their potential.

Within WP5, we have employed state of the art deep learning techniques to create a model capable to handle overwhelming complexity of crowds and detect anomalous crowd behaviour. To create such a model, we have manually annotated over 20 hours of CCTV footage, recorded from MONICA's FDL and MOVIDA pilots. To increase the integrity and reliability of the annotations, each video has been annotated by 3 folds. The annotation process, qualitatively monitors and measures the crowd movement, behaviour for anomalies such as sudden rush, overcrowding and loitering. These annotations have been used to train a two stage deep convolutional neural network to extract the temporal and spatial features from the crowd and classify the anomalous crowd behaviour. The proposed model has been integrated with other crowd monitoring algorithms and components in WP5 to deliver a complete end to end package for crowd monitoring through CCTV cameras.

To understand the impact of task 5.4 within the MONICA project, Table 1 outlines the use of video-based information retrieval model using crowd anomalies within the current MONICA solutions, the linked MONICA use cases, along with a brief explanation of how the proposed model will be utilised.

Table 1: Task 5.4 Information Retrieval model for event Analysis within MONICA

MONICA Solution	MONICA Use Case	Use of Task in context
Video Retrieval based on Crowd Dynamics Design and development of a model, capable to identify the anomalies such as a sudden rush, overcrowding, loitering and stampede in the crowd using the video signal.	• UC 3.1 Detect high risk queues • UC 3.2 Redirect high risk queues • UC 3.3 Monitor crowd based on capacity • Manage crowd based on capacity • UC 7.1 Detecting an incident • UC 11.1 Initiate full evacuation	The crowd behaviour analysis model can detect abnormal crowd behaviour e.g. rush, overcrowding, loitering and stampede using human activity analysis algorithms. The task can be used to monitor the crowd density in various areas of an event. Capable to predict potentially high-risk and unsafe crowd situations. This task could potentially enable the involved actors to successfully detect/report an incident.

3.2 Architecture

This section presents an overview of the MONICA architecture with focus on the components relating to T5.4 information retrieval model for event analysis. 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 T5.4.

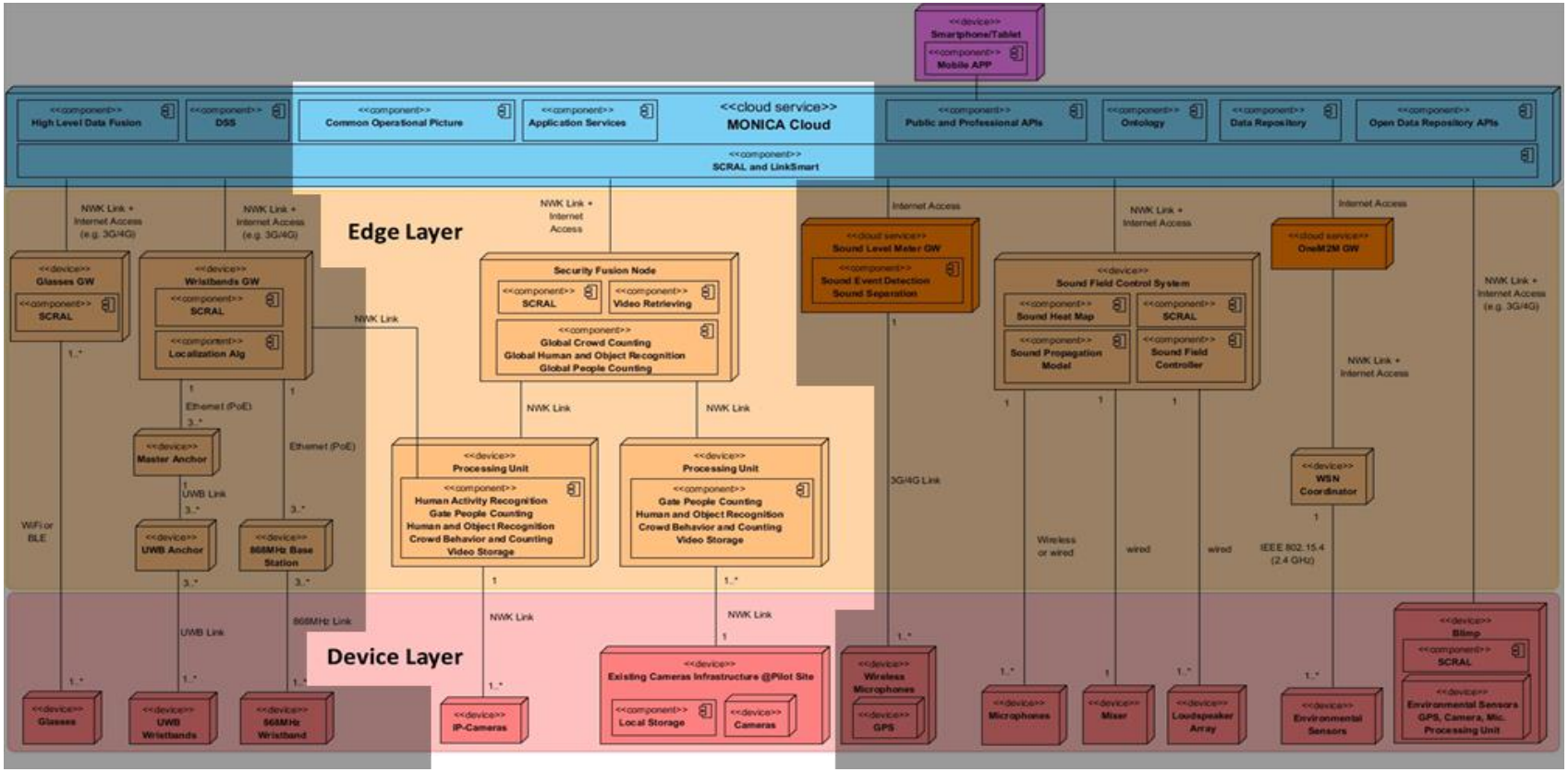


Figure 1: Detailed view of the MONICA Deployment Architecture.

Similar to other crowd monitoring algorithms (crowd counting and density estimation, gate counting, crowd flow analysis, object detection, fight detection) in WP5, the proposed video retrieval algorithm sits in between VCA core and Security Fusion Node (SFN), receives the CCTV video signal from the VCA core, processes the video for possible crowd anomalous behaviour and then forwards the inference and analytics to SFN. SFN then compiles the analytics results into a JSON message format and then forwards it to MONICA cloud. The models reside statically in WP5 architecture. Video retrieval algorithm will be manually updated as more training data provided. Periodical update will gradually improves the accuracy of the algorithm. 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.

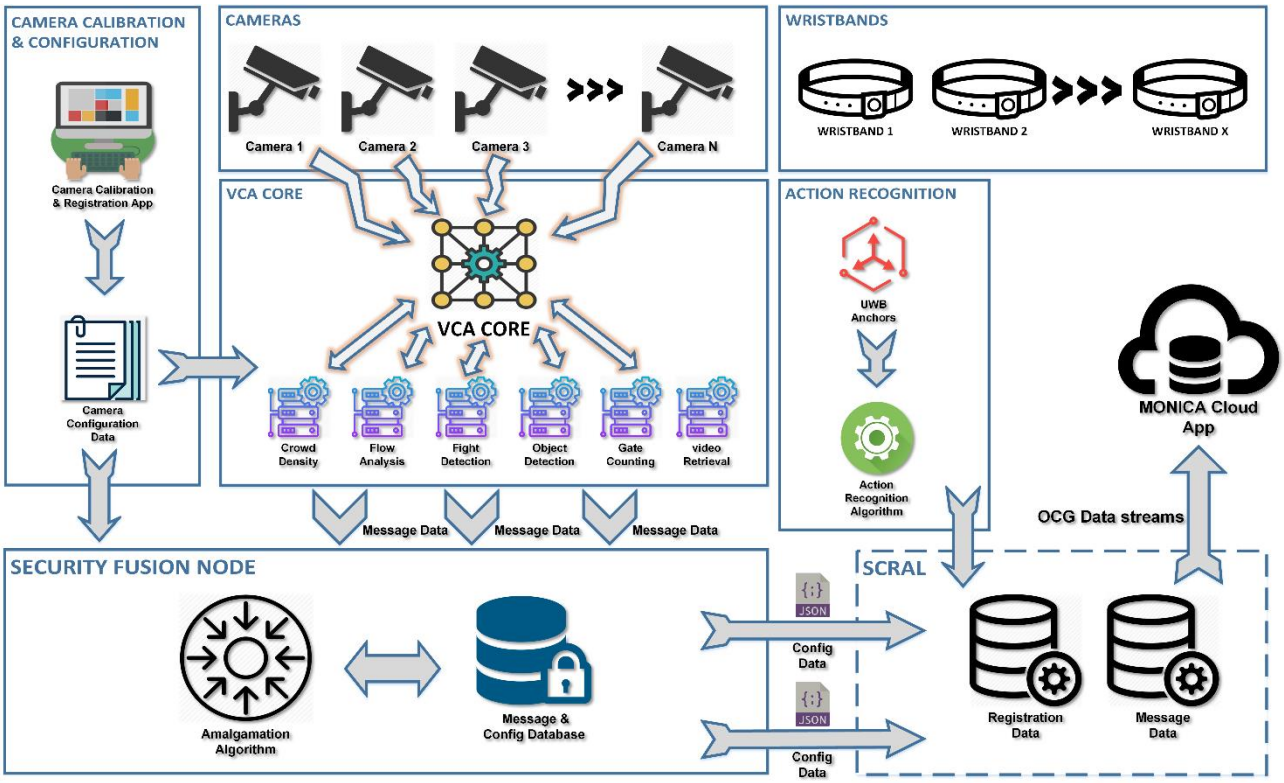


Figure 2: Extrapolated view of the WP5 components with respect to the storage and use of models with the various detection algorithms. The proposed video retrieval algorithm sits in between VCA core and Security Fusion Node (SFN)

Crowd behaviour analysis in images and videos is a fundamental challenge of computer vision. Classic computer vision and machine learning techniques were struggling with overwhelming complexity and unpredictability of the crowd behaviour models. Dramatic progress has been achieved by supervised convolutional neural network (CNN) models on crowd behaviour analysis tasks. The following section describes the data preparation and annotation for anomalous crowd behaviour.

3.3 Data Preparation and Annotation

Video annotation is a time consuming and manually laborious process as it requires an individual to go through each video and identify the crowdness of each video frame. To simplify this, one possible option is to utilize so called ‘crowdsourced’ data annotation services such as Amazon Mechanical Turk, to annotate the available MONICA data for use in training of the algorithm models. However, this is subject to ethical approval, which is often impossible due to the GDPR restrictions for outsourcing video data including humans.

Due to the above mentioned limitations, we have developed our own annotation tool to be able to annotate video data from two MONICA pilot sites i.e. FDL and MOVIDA (details in Table 2) and prepared them for training the algorithm.

Table 2: Data specification from two MONICA pilot sites.

	Description	Format	Annotations
FDL	Footage from 4 cameras taken during the Fete des Lumieres event.	1080x1920 25fps H.265	Manual annotation was performed.
MOVIDA	Footage from 2 cameras taken during the MOVIDA event.	1080x1920 25fps H.265	Manual annotation was performed.

A key aspect of implementation of the proposed algorithm is availability of labelled data for training the algorithm to learn and perform the required tasks. Labelled data can include bounding boxes highlighting specific areas of a scene such as crowd or summarizing a scene through the global count. In WP5 deliverable, labelled data are annotated videos where the presence of the target object, which is crowdness of the scene, is annotated with numbers in a range of 0-100 where 0 presents no-crowd and 100 presents fully-crowded scenes. Then, these annotated data are used as training data by the proposed algorithm to learn features in the videos and identify anomalous crowd behaviour. Next section explains details of video annotation tool developed by KU.

3.3.1 Video Annotation Tool

This section describes developed annotation tool, its interface and video data pre-processing of FDL and MOVIDA pilot sites where MONICA events happened. Figure 3 and Figure 4 presents example frames from FDL and MOVIDA that would be annotated by the annotation tool and finally evaluated by the algorithm.

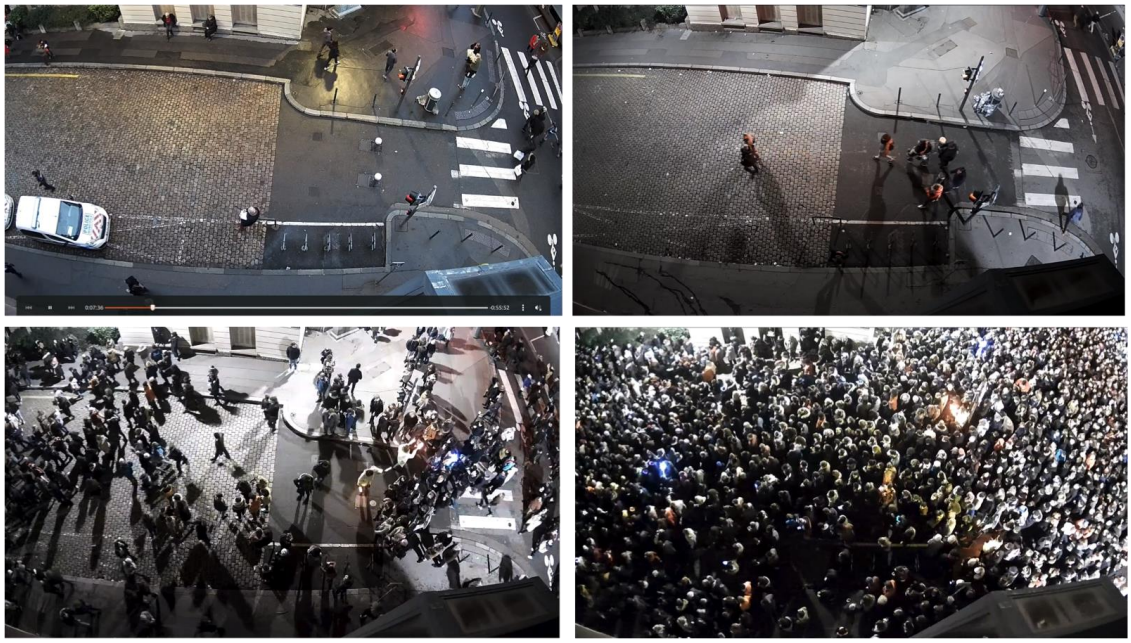


Figure 3: Example images extracted from FDL pilot video data with different crowd density.

Figure 4: Example images extracted from MOVIDA pilot video data with different crowd density.

The FDL pilot contains 20 hours of videos recorded from installed CCTV cameras in the venue and the MOVIDA pilot contains 10 hours video data, totally give us 30 hours of data for annotation.

Over the past years, several annotation tools for image and video data have been proposed that are tailored to address one specific computer vision problem such as segmentation, classification and object detection. In order to recognize anomalous crowd behaviour, we developed an annotation tool to read each video data and label the crowdness of the scene.

The annotation tool was designed using Qt Creator software (version 3.5.1) due to its friendly interface, extensive documentation available and portability between Windows and Linux operating systems. The main screen of the developed annotation tool is shown in Figure 5. By running the main source .CPP file written for this task, a window is popping up including four buttons on the up-left side and two sliders on the down-right side.

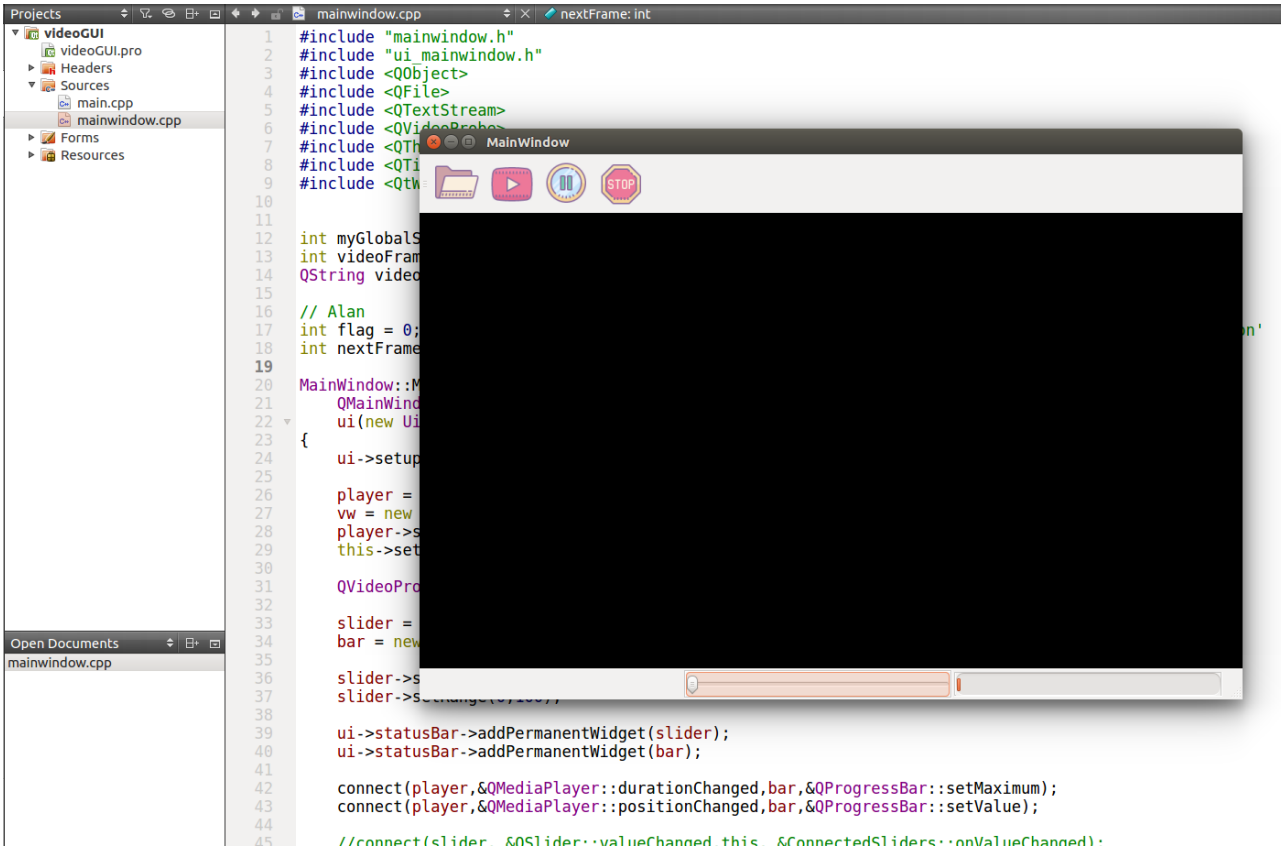


Figure 5: A Graphical User Interface (GUI) developed for video annotation.

The four top buttons in the GUI are used for video uploading, playing, pausing and stopping. The first slider from the left bottom side of the GUI is employed for determining the crowdness of the video by moving the slider toward left/right and the second slider shows the video progress in percentage.

By uploading the intended video file in the GUI and playing the video, annotation process is started. Then, the annotator can label each frame of video by moving the slider and assigning a value from 0-100 to each frame where 0 presents no-crowd and 100 presents fully-crowded scene.

Figure 6 shows an example of the GUI used for FDL pilot data annotation. As it is shown, at the bottom of the figure, for each video frame the saliency, which is the crowdness of the scene, is shown that has been assigned by the annotator. Here, the crowdness of the scene was annotated as value 71 that shows a crowded scene.

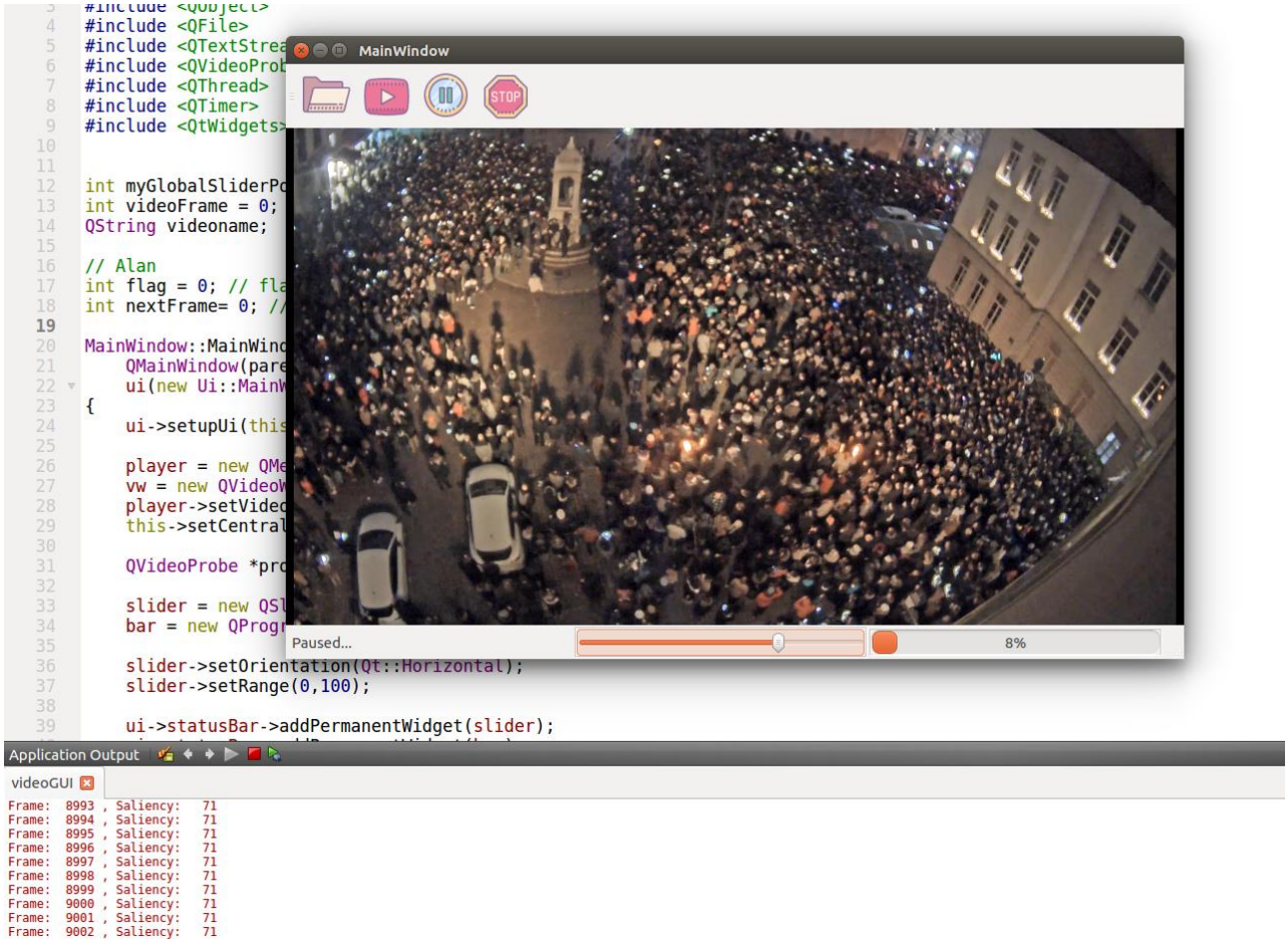
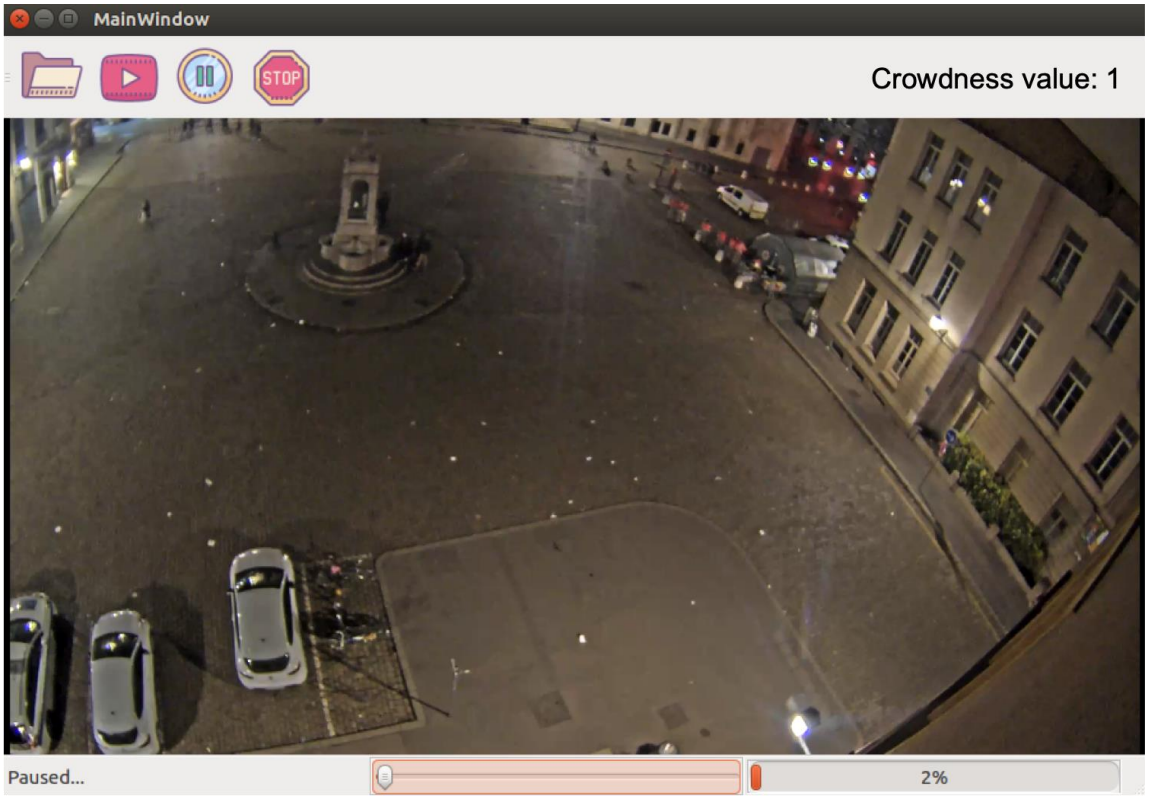
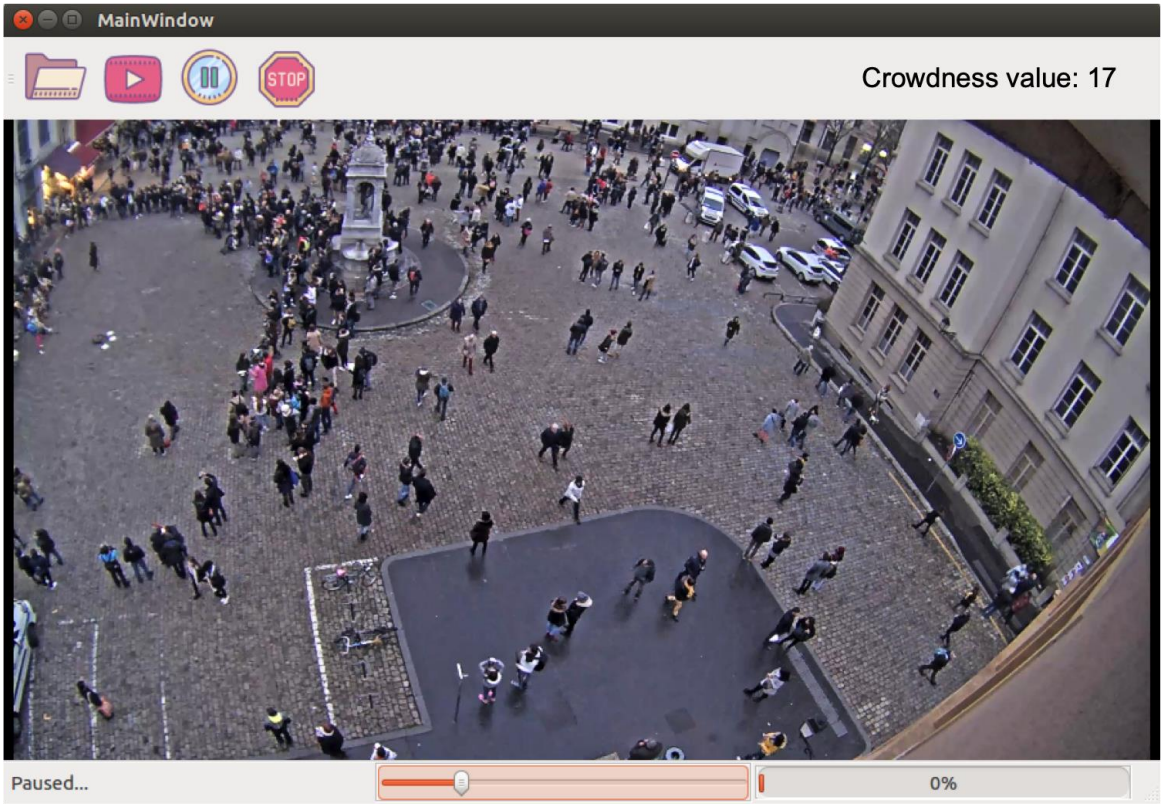


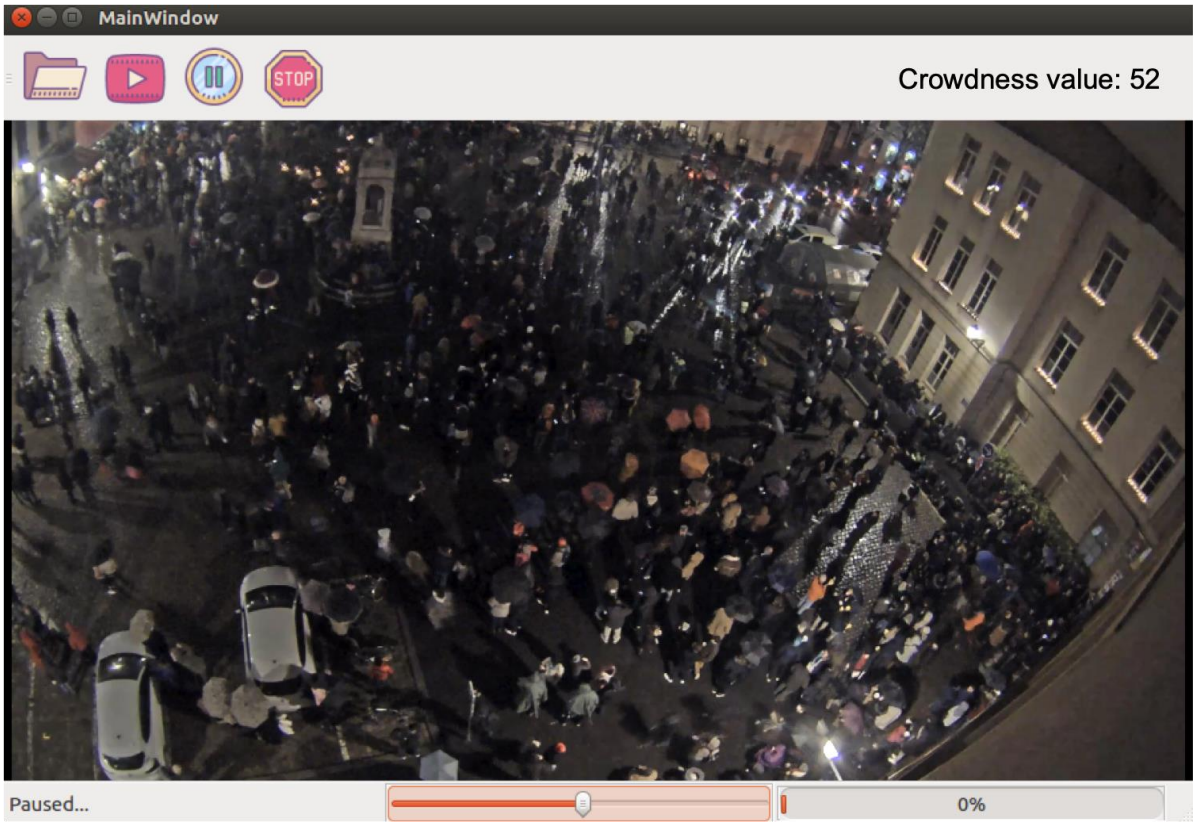
Figure 6: An example of GUI with FDL pilot data and annotated frames.

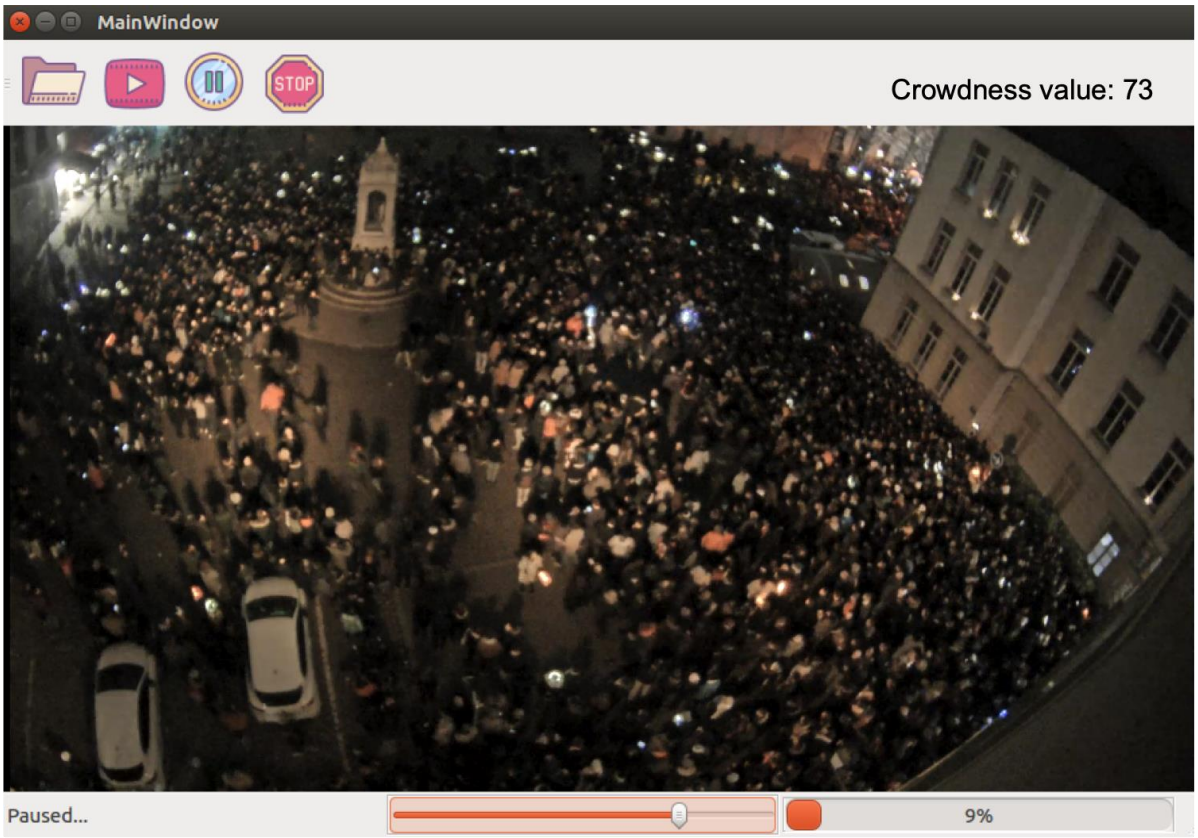
Figure 7 presents some examples of annotated frames with different values for the crowdness of the scene such as crowdness level of 1, 17, 52, 73 and 92.

Three annotators were hired to label the FDL and MOVIDA pilot video data. The annotators did not require special background and were briefed about the task requirements and were provided instructions on annotating the videos. Video clips with different durations such as 1 hour and 1:20 hours were played and annotators had the choice of pausing the video and take a break whenever they required without time constraint. Each video clip was annotated three times by three annotators and then the annotation values were averaged to ensure the annotations returned are representative. After each frame annotation, the annotation values which are numbers between 0 and 100 were saved in a text file.


MainWindow
STOP
Crowdness value: 1
Paused...
2%


MainWindow
STOP
Crowdness value: 17
Paused...
0%


MainWindow
STOP
Crowdness value: 52
Paused...
0%


MainWindow
STOP
Crowdness value: 73
Paused...
9%

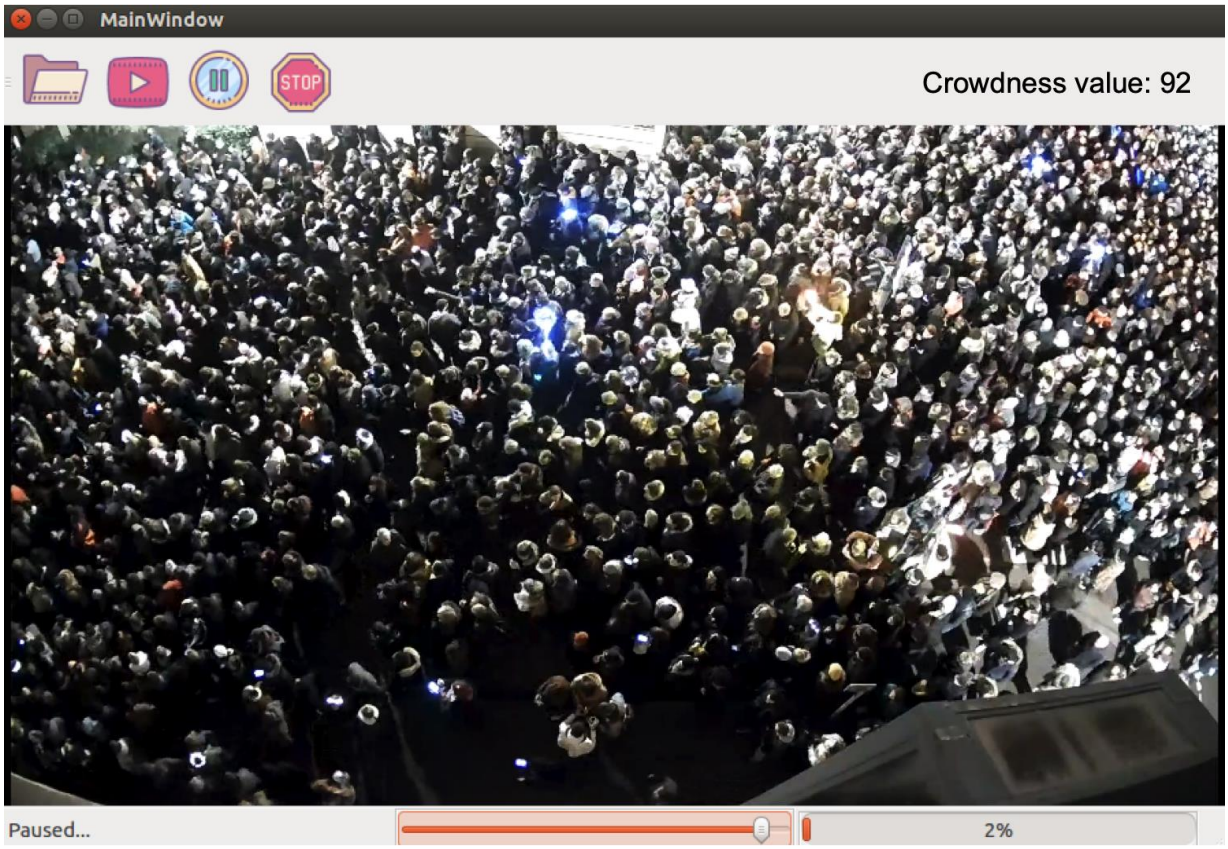


Figure 7: Some sample annotated images from FDL MONICA pilot with different level of crowdness such as 1, 17, 52, 73 and 92, respectively.

3.3.2 Data Preparation for Training the Algorithm

Before feeding the labelled data to the algorithm, we need to do some data pre-processing and prepare them for training/testing the algorithm. In order to create the training / testing data, we have split the raw input CCTV videos into short clips of 30 seconds. We have identified that 30 seconds is the sweet spot to identify anomalous behaviour in the crowd. Further, this allows us to increase the number of training samples.

We have 100 classes (values between 0-100) of labeled video data. The overall aim of the proposed algorithm is to identify four types of anomalous behaviours i.e. high-anomalous, medium-anomalous, low-anomalous and no-anomalous behaviour. Therefore, we assigned one of four mentioned classes as a label for each 30 seconds video clip. Using the following code we transformed 101 classes into four anomalous behaviour classes:

As it is shown in the above code, from each annotated video file 'math.floor' function is calculated that shows the anomaly level of that 30 seconds video clip and finally will be used as a label for that video file.

This section presented data annotation process of post-event (in this case a MONICA pilot site) anomalous crowd behaviour analysis. FDL and MOVIDA pilot sites data were annotated using an annotation tool that was developed by KU. After annotation, data were divided into four classes of no-anomalous behaviour, low-anomalous behaviour, medium-anomalous behaviour and high-anomalous behaviour. We will use these

annotated data to detect anomalous crowd behaviour using our proposed algorithm. The following diagram describes the proposed video retrieval algorithm workflow for anomalous crowd behaviour.

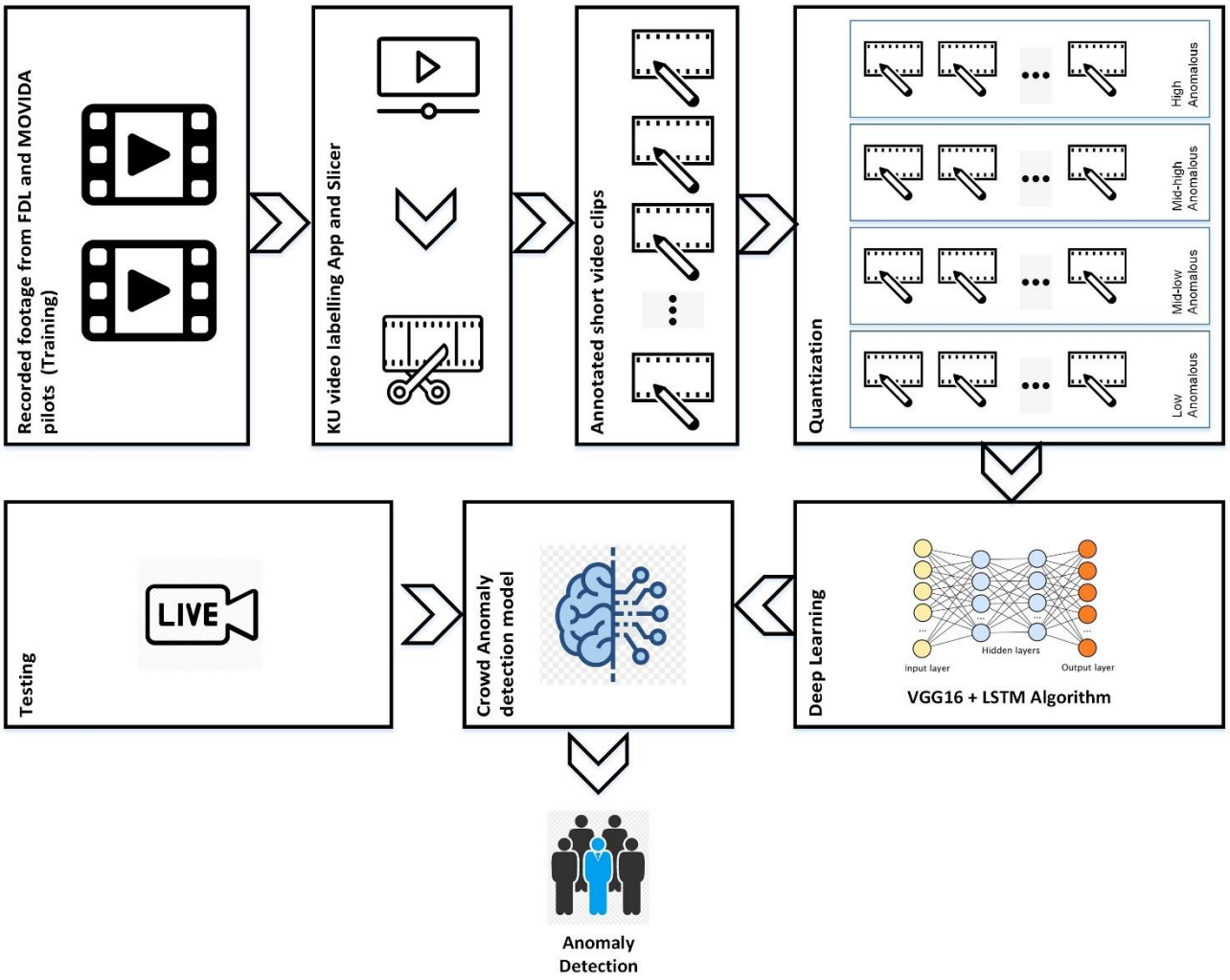


Figure 8: Proposed video retrieval algorithm workflow for anomalous crowd behaviour

3.4 Video Retrieval Algorithm for Anomalous Crowd Behaviour

Modelling crowd dynamics and behaviour is a very challenging problem. Currently, deep learning-based models are being developed to automatically generate rich pattern representations and statistics of the crowd videos involving anomalous crowd behaviour. As illustrated in the architecture overview, these statistics are then fed up to the MONICA cloud for further high-level processing in HLDF 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, and public gatherings at train stations or bus stops. Numerous studies have been devoted to track, recognize and understand various behaviours in videos. These studies have primarily focused on low density crowds, nonetheless, relatively little effort has been made on reliable classification and understanding of crowd behaviour in real world crowded scenes like MONICA pilots. In dense crowd scenes, individuals are often relatively small and take up only few pixels across the frame, making the recognition and analysis of the anomalous crowd behaviour very challenging.

In general, crowd behaviour, counting and simulation studies can be categorized into two major approaches, Holistic and Atomistic. Holistic crowd counting and behaviour analysis approaches utilize global crowd features and characteristics to estimate the crowd size, density and behaviour. In other words, in holistic approaches 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. It is often advantageous to understand the crowd behaviour without knowing the actions of the individuals. Hence, these approaches are suitable for large and dense crowds.
In contrast, Atomistic crowd counting and behaviour analysis approaches exploit local feature (individuals shape, colour and characteristics) to estimate the crowd size, density and behaviour. In these approaches, individuals (human and objects) are detected and segmented to perform behaviour analysis. This kind of complex segmenting and tracking of individuals in crowded videos is a very challenging task. Atomistic approaches are suited better with sparse crowds. To form the video retrieval model for anomalous crowd behaviour in T5.4, we use the former approach, where individuals are not segmented or tracked, but the group of people are perceived holistically so as to recognize their behaviour.
The proposed model uses a form of long-term recurrent deep convolutional neural network, a class of architectures for visual recognition and description in video sequence which combines convolutional layers and long-range temporal recursion in an end-to-end trainable fashion. This architecture, which has been used for various applications such as activity recognition, can be adopted to discriminate anomalous crowd behaviour once relevant training ground truth data provided. Figure 9 shows the proposed long-term recurrent deep convolutional neural network architecture.

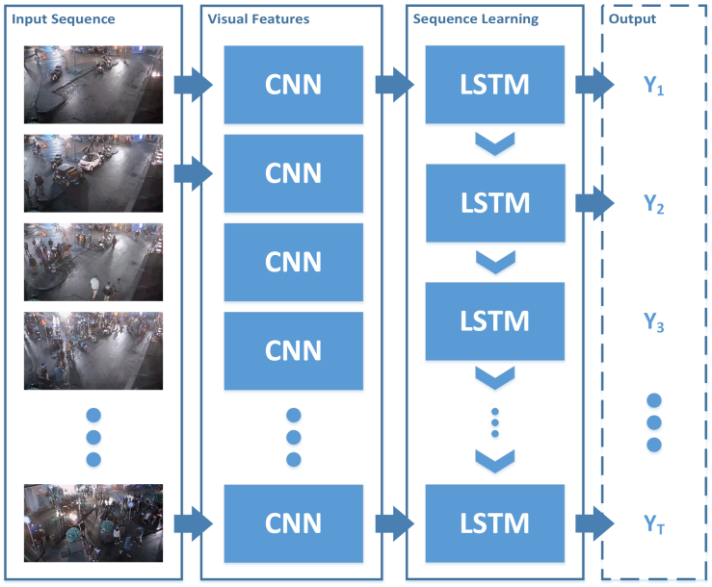


Figure 8: Long-term Recurrent Convolutional Networks (LRCNs) architecture, leverages the strengths of rapid progress in CNNs for visual recognition is video sequence with time-varying inputs.

In order to identify crowd anomalous behaviour, deep models which are also deep over temporal dimensions; i.e., have temporal recurrence of latent variables are needed to process a sequence of frames long enough to portray a certain crowd behavioural activity. RNN models are “deep in time” explicitly so when unrolled and form implicit compositional representations in the time domain. The proposed model uses a Long-term Recurrent Convolutional Network (LRCN) model (Jeffery 2015) combining a deep hierarchical visual feature extractor (such as CNN) with a model that can learn to recognize and synthesize temporal dynamics for tasks involving sequential data (such as anomalous crowd behaviour).

Long Short Term Memory networks, which usually just called “LSTM”, are a special kind of RNN, capable of learning long-term inter-frame dependencies. LSTM were introduced by Hochreiter & Schmidhuber in 1997, and were refined by many others afterward. They work tremendously well on a large variety of problems where temporal features are key discriminant factor. LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behaviour.

All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single $tanh$ layer. Similar to other RNN networks, LSTM have this chain like structure, but the repeating module has a different structure. Instead of having a single neural network layer, there are four, interacting in a very special way. Figure 10 shows the basic architecture of LSTM network.

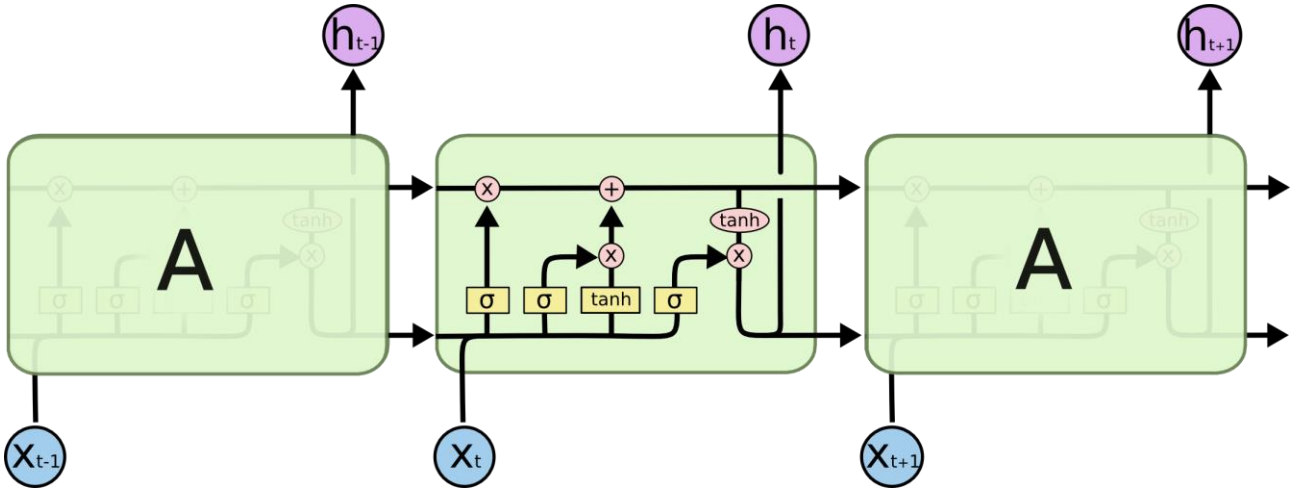


Figure 10: The repeating module in an LSTM contains four interacting layers

Figure 10 depicts the core of video retrieval approach. LRCN works by passing each visual input X_t(an image in isolation, or a frame from a video) through a feature transformation $\varphi v(.)$ with parameters , usually a CNN, to produce a fixed-length vector representation $\varphi v(X_t)$. The outputs of φv are then passed into a recurrent sequence learning module. In its most general form, a recurrent model has parameters$W$, and maps an input x_t and a previous time step hidden state h_{t-1} to an output $z_t$and updated hidden stateh_t. Therefore, inference must be run sequentially (i.e., from top to bottom, in the Sequence Learning box of Figure 9), by computing in order: $h_1 = f_w(x_1, h_0) = f_w(x_1, 0)$, then $h_2 = f_w(x_2, h_1)$etc., up to h_t.

Each frame in a length T sequence is the input to a single convolutional network (i.e., the convnet weights are tied across time). We consider both RGB and flow as inputs to our crowd anomalous behaviour recognition system. Flow is computed and transformed into a “flow image” by scaling and shifting x and y flow values to a range of $[-128, +128]$. A third channel for the flow image is created by calculating the flow magnitude.

During training stage, input videos are resized to 240 × 320 and augmented to 227 × 227 crops and mirrored. Augmentation helps us to increase the number of training samples and improve the overall accuracy of the proposed model. The LRCN model is trained with overall 120 hours of crowd videos (after augmentation) captured from FDL and MOVIDA MONICA pilots (on the order of 100 frames when extracting frames at 25 FPS). LRCN is trained to predict the anomalous behaviour in crowd videos based on the provided class labels in ground truth annotation. To produce a single label prediction for an entire input video, we average the label probabilities of the outputs of the network’s softmax layer across all frames and choose the most probable label. At test time, we extract 16 frame clips with a stride of 8 frames from each video and average across all clips from a single video.

The CNN base of LRCN in crowd anomalous behaviour detection model is a hybrid of the CaffeNet reference model (a minor variant of AlexNet) and the network used is pre-trained on the 1.2M image ILSVRC-2012, classification training subset of the ImageNet dataset, giving the network a strong initialization to facilitate faster training and avoid overfitting to our relatively small MONICA annotated dataset.

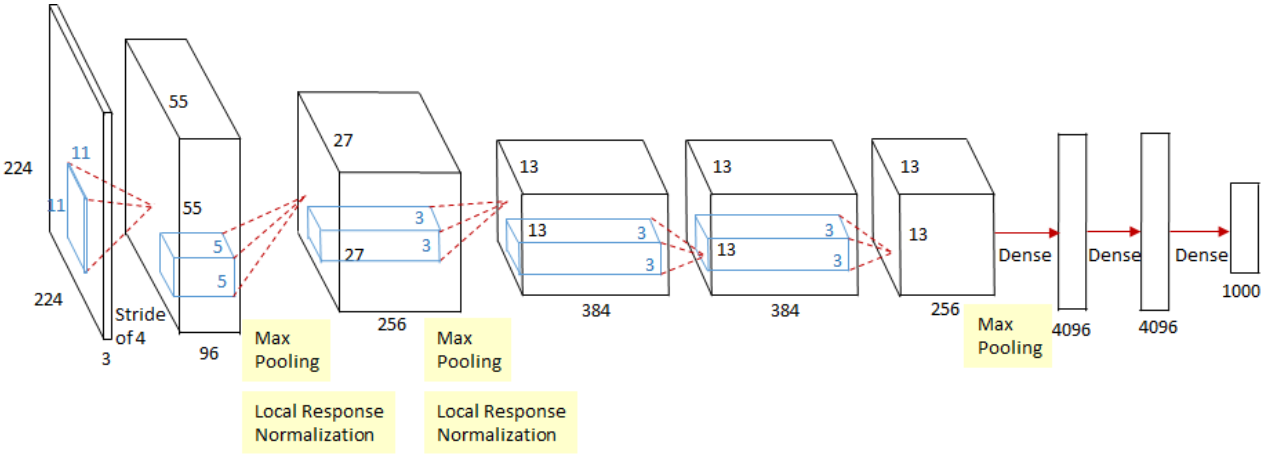


Figure 11: The CaffeNet CNN model is used to extract spatial features from each input frame.

Aside from CaffeNet, we also investigate the popular VGG16 model, pertained on 1.2M image from ILSVRC-2012 dataset in order to extract the spatial features from the individual video frames. VGG16 model hits the sweet spot between complexity and performance. VGG16 has significantly lesser parameters than models like GoogleNet/Inception, yet delivers competitive results in majority of the real-world scenarios. Figure 12 shows the VGG16 architecture.

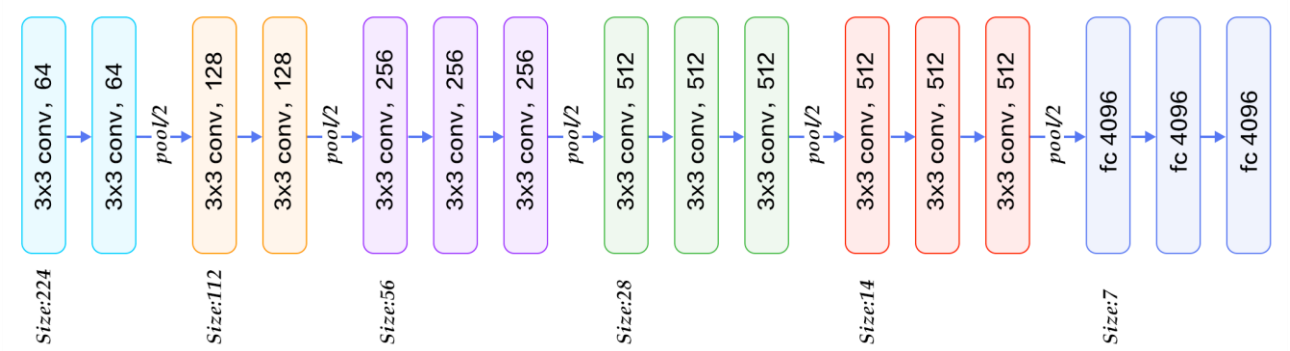


Figure 12: The VGG16 CNN model is used to extract spatial features from each input frame.

3.4.1 Experimental Results

The proposed anomalous crowd behaviour detector has been trained and tested using our in-house MONICA dataset. Dataset annotation and preparation has been explained earlier in section 3.3 of this report. We have investigated three different architecture including VGG16+LSTM which uses VGG16 to extract features from individual frame of the video and the sequence of frame features are then taken into LSTM recurrent networks for classification. Another architecture uses VGG16 + Bidirectional LSTM approach where VGG16 used to extract features from individual frames of the video, the sequence of frame features are then taken into bidirectional LSTM recurrent networks for classifier. The CaffeNet + LSTM setup also being investigated in this research.

In all setups, the network has been trained for 20 epochs across all four classes including Low, Midlow, Midhigh and High. The network has been trained with 2/3 of the entire dataset (~2400 short clips) and been tested over the remaining video clips (~1200 short clips). We have measured the performance using Accuracy and Loss metrics in both training and testing phases. Table 4 shows the training and testing subsets statistics.

Figure 13 shows the evaluate results using VGG16 + LSTM over 20 epochs. It can be observed that the network converged to global minima in less than 15 epochs and behave as expected. Training and testing loss values are comparable which indicates overfitting is negligible. The VGG16 + LSTM architecture achieved average accuracy and Loss of 80.46 and 0.62 respectively in testing phase. This algorithm achieved average recall value of 83.2 across four designated classes in the testing set; however, the recall drops to 74.9 in high

anomaly class only. We also put VGG16 + Bidirectional LSTM architecture into the test. This model has significantly higher computational cost which significantly slows down the process. However, in MONICA D5.4 objective we mainly focus on the overall system accuracy and optimal computational latency is less of a priority.

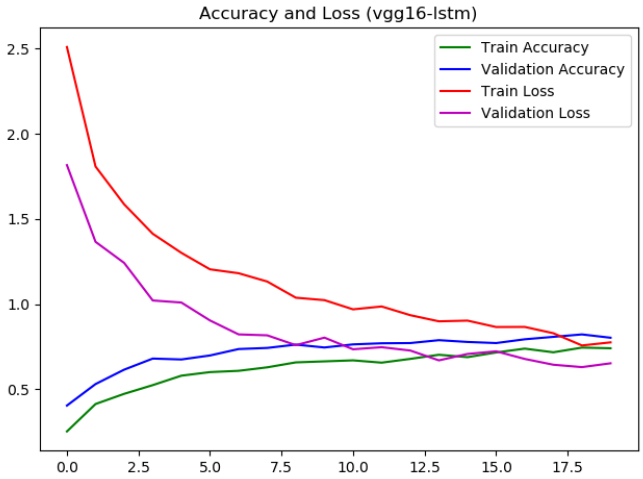


Figure 13: Accuracy and Loss metrics of VGG16 + unidirectional LSTM over 20 epochs

Similar to VGG16 + LSTM architecture, the VGG16 + bidirectional LSTM architecture has been ran for 20 epochs across 2/3 of the entire dataset and been tested over the complement. The results which presented in Figure 14 shows that the VGG16 + Bidirectional LSTM architecture with respective accuracy and loss value of 86.07 and 0.88 in testing phase, outperformed the unidirectional LSTM in every respect. This algorithm achieved average recall value of 88.9 across four designated classes in the testing set; however, the recall drops to 80.3 in high anomaly class only. With the help of dropout layers and normalizations, bidirectional LSTM architecture does not show significant amount of overfitting. Table 3 summarizes the experiment results in this study.

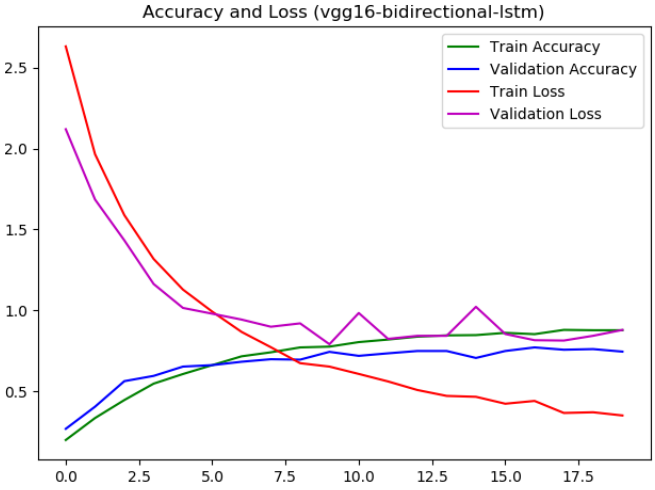


Figure 14: Accuracy and Loss metrics of VGG16 + bidirectional LSTM over 20 epochs

Table 3: Summary of the results

Architecture	Loss		Accuracy		Recall (Average)		Recall (High Anomaly Class)
	Training	Testing	Training	Testing	Training	Testing	Testing
VGG16 + Unidirectional LSTM	0.74	0.62	91.21	80.46	-NA-	83.2	74.9
VGG16 + Bidirectional LSTM	0.39	0.88	94.71	**86.07**	-NA-	**88.9**	**80.3**
Top less VGG16 + Unidirectional LSTM	0.43	0.38	90.01	83.14	-NA-	85.4	77.5

Table 4: Training and testing sets statistics

	Low Anomaly	Mid-low Anomaly	Mid-high Anomaly	High Anomaly	TOTAL
Training set	**1240**	**703**	**334**	**123**	**2400**
Testing set	**515**	**341**	**259**	**85**	**1200**

4 Summary

This deliverable presents the current progress relating to task T5.4 Information Retrieval for Post-event Analysis of the MONICA project pilots. Automated detection of anomalous crowd behaviour such as sudden rush, overcrowding, loitering and stampede is a challenging task. Detection of such activities in video signal, demands sophisticated yet efficient computer vision models, capable to extract the underlying spatial and temporal features that correlate to anomalous crowd behaviour.

In this regard, we have employed state of the art deep learning techniques to create a model capable of handling overwhelming complexity of crowds and detect anomalous crowd behaviour. The proposed model uses a form of long-term recurrent deep convolutional neural network (LRCN), which combines a deep hierarchical visual feature extractor (such as CNN) with a model that can learn to recognize and synthesize temporal dynamics for tasks involving sequential data (such as anomalous crowd behaviour). This architecture, which has been used for various applications such as activity recognition, can be adopted to discriminate anomalous crowd behaviour once relevant training ground truth data provided.

The performance of the model was tested on two MONICA pilot video data such as FDL and MOVIDA. These data were annotated manually using an annotation tool that was developed by KU. To increase the integrity and reliability of the annotations, each video has been annotated by 3 folds. During the annotation, crowdness of the scene was annotated using numbers in a range of 0-100 where 0 presents no-crowd and 100 presents fully-crowded scenes. To prepare the annotated data for training the algorithm, they were split into short 30 seconds video clips and for each clip one label (from no-anomalous behaviour, low-anomalous behaviour, medium-anomalous behaviour and high-anomalous behaviour) was assigned. Experimental results on in-house MONICA crowd behaviour dataset shows that on average the proposed VGG16 + Bidirectional LSTM algorithm achieved 86.07 % accuracy and 88.9 % of recall across four classes.

5 List of Figures and Tables

a. Figures

b. Tables

6 References

(Jeffery 2015) Donahue, Jeffrey, et al. "Long-term recurrent convolutional networks for visual recognition and description." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

(Hochreiter 1997) Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.

(Mahadevan 2010) Mahadevan, Vijay, et al. "Anomaly detection in crowded scenes." *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010.

(Wu 2010) Wu, Shandong, Brian E. Moore, and Mubarak Shah. "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes." *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010.

(Yuan 2014) Yuan, Yuan, Jianwu Fang, and Qi Wang. "Online anomaly detection in crowd scenes via structure analysis." *IEEE transactions on cybernetics* 45.3 (2014): 548-561.

(Feng 2017) Feng, Yachuang, Yuan Yuan, and Xiaoqiang Lu. "Learning deep event models for crowd anomaly detection." *Neurocomputing* 219 (2017): 548-556.

(Saligrama 2012) Saligrama, Venkatesh, and Zhu Chen. "Video anomaly detection based on local statistical aggregates." 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012.

(Li 2013) Li, Weixin, Vijay Mahadevan, and Nuno Vasconcelos. "Anomaly detection and localization in crowded scenes." IEEE transactions on pattern analysis and machine intelligence 36.1 (2013): 18-32.

(Jiang 2009) Jiang, Fan, Ying Wu, and Aggelos K. Katsaggelos. "Detecting contextual anomalies of crowd motion in surveillance video." 2009 16th IEEE International Conference on Image Processing (ICIP). IEEE, 2009.

(Wang 2010) Wang, Shu, and Zhenjiang Miao. "Anomaly detection in crowd scene." IEEE 10th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING PROCEEDINGS. IEEE, 2010.