Video Summarization And Video Highlight Selection Tools To Facilitate Fire Incident Management

Florian Vandecasteele
Ghent University - IDLAB
Florian.Vandecasteele@ugent.be

Krishna Kumar
Ghent University-IDLAB
krishnakumar.tc@ugent.be

Kenzo Milleville
Ghent University - IDLAB
Kenzo.Milleville@ugent.be

Steven Verstockt
Ghent University - IDLAB
Steven.Verstockt@ugent.be

ABSTRACT

This paper reports on the added value of combining different types of sensor data and geographic information for fire incident management. A survey was launched within the Belgian fire community to explore the need of added value and the use of new types of sensor data during a fire incident. This evaluation revealed that people are visually-oriented and that video footages and images are of great value to gain insights in a particular problem. However, due to the limited available time (i.e., fast decisions need to be taken) and the large amount of cameras it is not feasible to analyze all video footages sequentially. To solve this problem we propose a video summarization mechanism and a video highlight selection tool based on the automatic generated image and video tags.

Keywords: machine learning, video data, filtering, crisis management, user-evaluation

1 INTRODUCTION

During the first minutes of an intervention, the incident officer needs to take fast and appropriate decisions with very limited information. However, the scene layout, the structure of the building, the fire dimensions and position, the hazardous materials present, and the status and the position of victims, are factors that affect the initial strategy and tactics.

Currently, there are two main communication tools for fire incident management: a broadcast channel for voice communication and a second textual channel for status updating. However, from a psychological perspective it is proven that our brain is better at processing visuals than text [1]. Our brain can process an image 60 000 times faster than a piece of text. On average, when a person sees a particular image for the first time, it takes 113 milliseconds to process the information. In that sense it is advisable to focus on the development of image related tools to facilitate the incident officers instead of improving the existing auditive or textual communication tools that have problems with reliability (e.g., connection-loss in concrete buildings) and interoperability.

To generate the needed visual information we can exploit the increased performance and quality of fixed security cameras and handheld devices such as smartphones and tablets. It is important, however, that the video footage is captured in the most appropriate way for visualization to a fire incident manager. Furthermore, the sensor input of the fixed and handheld cameras can be streamed to the incident commander to have an overview of the incident scene. However, it is not possible for a person to analyze all the video streams in a sequential manner due to the cognitive disability. There is a need for a video summarization mechanism that filters out the redundant and unnecessary frames, while preserving the distinctive frames.

Besides the visual camera information there are other sensor types (e.g., temperature indications, CO levels) with valuable information for the incident commander. Visualizing too many sensors will lead to enacted sensemaking where an individual will select unconsciously some parts of the data to make their decision. Especially in critical or stress situations this can lead to overshooting or undershooting of the situation. To solve these issues the firefighting community should get aware of the opportunities of web dashboards to quickly and easily create reports and data visualizations [2]. The core idea of dynamic dashboards is to efficiently display data in such a way that it can enhance user insight into the data. An example is the fire probability calculation and visualization for residential regions by Netage. Similarly, the Moses project explored a dashboard system for seamless localization, employability and health status monitoring (e.g., heart rate, air consumption) of firefighters.

The remainder of this paper is organized as follows. Section 2 presents the specific information needs of first responders. Furthermore in Section 2 the data insights from the questionnaire are discussed more thoroughly. Section 3 goes more in detail into the video summarization process. Subsequently, Section 4 proposes the possibilities for video and frame retrieval. Next, Section 5 discusses the results of our proposed methodology and Finally Section 6 gives some conclusions and suggestions for future work.

2 FIRST RESPONDERS INFORMATION NEEDS

2.1 Visible and cognitive disability

Biologically, human beings have psychological restrictions in terms of observing and processing information. Furthermore, an individual can only process a limited amount of simultaneous stimuli. Ungerer et al. [3] stated that a person can only process a fraction of the information in a conscious way. According to his theory, each second 3 up to 5 visual and 3 auditive or tactile inputs can be processed simultaneously. Schaub et al. [4] on the other hand denoted that a human is only capable of processing 7 concurrent signals.

The head of operations and firefighters in general have to work in difficult, often unknown contexts, where there is an increased stress level. Subsequently, due to the increased level of stress there is a higher chance of reduced cognitive ability. Despite the dynamic and difficult character, decisions need to be made without detailed investigation or analysis. In general the fire scene is a complex situation due to the following factors:

- A comprehensive list of 'unknown' variables (e.g., number of people present, evacuation necessity, fire behavior, fire expansion and building contents);
- Variables that affect each other (e.g., ventilation openings will change the fire growth);
- Variables that change in time and space without being transparent or predictable.

Due to the risk of reduced cognitive ability and the psychological restrictions there is a need for a smart, adaptive visualization application. The time first-responders spend on gathering static and volatile situational information from affected people as well as from responsible persons at the emergency site can be reduced if they have access to the needed information. The following subsections will discuss the related work for fire incident visualization and the subjective criteria for data importance rating.

2.2 Related work

Nunavath et al. identified the information needs for first responders from a literature review, fire drills and interviews. Prioritizing of information items is needed for different stages of building fire emergency response operations and to ensure a maximum of 7 concurrent decision signals. The most commonly mentioned information item in the work of Nunavath et al. is the building related information, which includes important information about the building layout plans, hazardous material location, resources location, floors, and rooms. Furthermore, the other most commonly mentioned information item was fire related information such as color, location and condition of the fire which can be derived with the techniques of Beji et al. [5] and Vandecasteele et al. [6]. Subsequently, Li et al. [7] investigated the information sources used in current practice (year 2013) and those desired to be used in the future. Furthermore, Li et al. considered the information items needed by first responders and the availability of technological solutions to obtain them in the United Stated. Furthermore, Li et al. mention that there are three timings where data should be delivered: before arrival, when arriving at the emergency scene and during the attack and the mitigation. In an automatic framework these three states should also be considered and evaluated. Hamins et al. [7] created the smart-firefighting roadmap and although, the roadmap gives some directions for further research, currently only limited implementations are available. This paper proposes some technological building blocks to start with 'smart-firefighting'. Finally, emergency

responders have different roles to perform, different tasks to handle, and different modes to communicate, which lead to an insufficient view of the complete emergency situation at the emergency site. In that respect Nunavath et al. [8] presented a conceptual domain model. Their model consists of four components: event component, actor component, objective component and building component. Each component contains several information resources and all components capture the complete building fire emergency response. However, as stated before it is not feasible to process all components sequentially and filtering and summarization mechanisms will be necessary.

2.3 Subjective criteria and discussion

Within the Belgian fire community, a survey was launched to explore the necessity and the use of new types of sensor data during a fire incident. This survey can be seen as an updated version of to the work of Li et al. The questionnaire was sent to all Flemish firefighting communities and input was asked from all hierarchical levels (i.e., strategic, tactical and operational). Furthermore, we targeted a balance in the ratio of volunteer and professional firefighters that filled in the inquiry. 21 different questions were asked and the first question was: ”What kind of data should be available during an incident?”. From the survey results it was remarked that the thermal imaging device (95,9% of respondents) is the most informative device used by firefighters, followed by the explosion warning device (65,8%) and the visual camera (65,7%). In the questionnaire it was pointed out that the more information available during an incident, the higher the fire crew score their situational awareness. However, as stated earlier in this section, there is a maximum of 7 simultaneous inputs that can be used. The most important features are the location of the fire source, the amount and the position of the victims and firefighters, and the structure and lay-out of the building.

Subsequently, in the questionnaire, some questions were related to the priority of the data. The camera footages, the thermal camera footages and the surrounding temperature were pointed out as prior. It is important to remark that this should be re-evaluated within a couple of years as most of the other sensor and data values are currently not available during an intervention and the decision makers are not trained to use certain variables. Furthermore, the questionnaires indicated that they allow a maximum processing time (i.e., delay to receive the first information) for the video streaming of 1 minute.

3 VIDEO SUMMARIZATION

As indicated in the questionnaire, decision makers want to have access to visual and thermal video streams. As it is not possible to analyze the complete history of a CCTV camera and it is not possible to investigate several cameras simultaneously, there is a need for video summarization. Video summarization can be classified into two types, namely static video summarization and dynamic video summarization. Static summarization results in a set of keyframes that best conveys the overall idea of the video. In most cases, keyframes are representative frames that could identify either the beginning or end of a scene transition in a video sequence. The number of resulting keyframes can vary as per the setting. Contrary, dynamic summarization collects small chosen fragments of the original video and arranges these fragments to obtain a new version of the video (i.e., logical story units).

Depending on the use case (e.g., fire incident, hazard materials incident) and the commanders function (e.g., evaluating the evacuation, investigating the structural damage or making strategic decisions for the intervention crews) different static or dynamic fragments will be indicated as important. Furthermore, personalized interaction will be necessary to learn the visualization needs based on the incident type, the operational function and the geographical location. The main focus of the following sections is on the static summary generation as it is not feasible to analyze several logical story units sequentially in a fast manner.

A schematic overview of our proposed indexing strategy is given in Figure 1. First the different video sensors are captured. Secondly, the video summarization selects a set of representative keyframes. Thirdly, for each keyframe the scene content understanding is performed and indexed. Finally, the user can visualize and query the set of keyframes on a semantic and visual level.

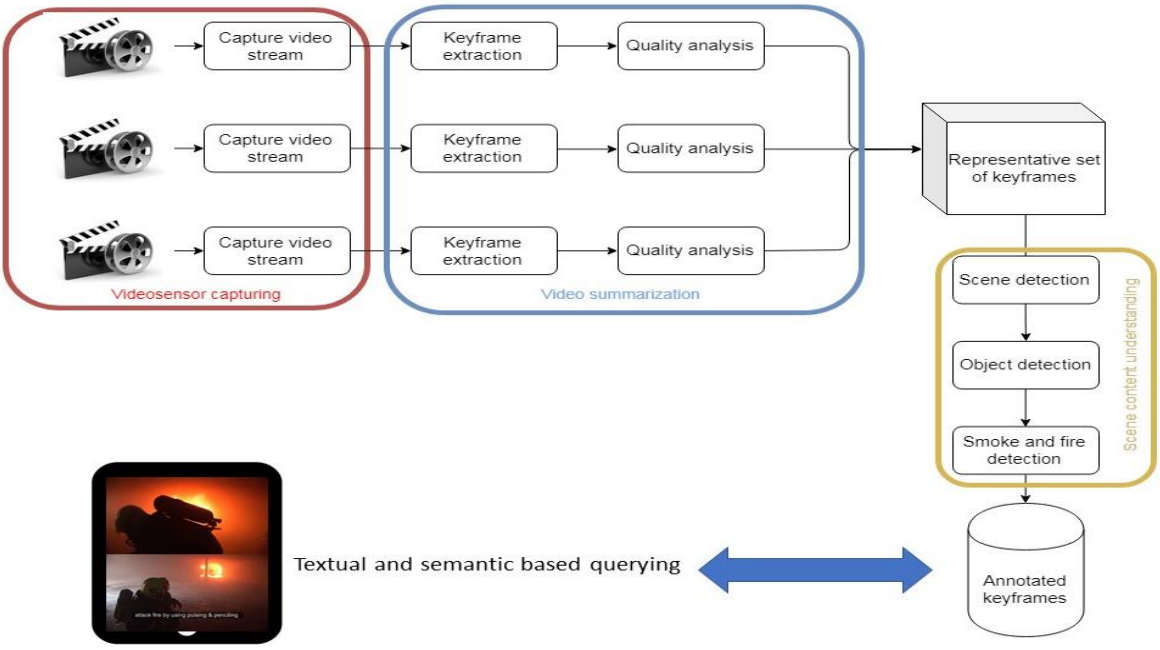

Figure 1. Schematic overview of the video capturing, summarization and content understanding process.

In order to visualize the relevant keyframes in a dashboard system and to decrease the computational cost of subsequent video processing tasks it is important to reduce the amount of video data by filtering out redundant and unnecessary frames, while preserving only those frames, distinctive and essential, to capture the entire video content. Furthermore, presenting the end-user with a limited list of representative keyframes instead of dynamic video fragments, improves their exploration, understanding and search process. Therefore the main focus of the following sections is on the static summary generation. The automated summarization of video content into representative keyframes, however, is a challenging problem due to the rapid change in lightning, viewpoint, and scene. In order to cope with these issues, we present a novel solution to extract, cluster, and filter meaningful keyframes. This summarization process consists of three major steps:

- The video summarization converts the video into a set of representative keyframes. All the keyframes represent one single video shot. These shots are detected using a grid-based histogram detection method and within each shot no-reference keyframe quality analysis is used to select the best frame within the shot as proposed by Vandecasteele et al. [9].

- The number of keyframes is reduced by clustering visually similar frames, while preserving as much as possible of the entire content.

- A limited set of representative keyframes is given to the object and scene retrieval algorithms to generate meaningful tags. This is out-of-scope of this paper, but more details can be found in the work of Vandecasteele et al. [10].

Much research has been done in the area of video summarization (i.e., keyframe retrieval). Ajmal et al. [11] give an overview of the different techniques and classification methods that are commonly discussed in literature, i.e., feature classification [12], clustering [13], shot selection [14] and trajectory analysis [15]. In the following subsection the grid-based histogram shot detection approach is proposed.

3.1 Shot detection

Compared to the state-of-the-art temporal shot segmentation algorithms [16], the proposed local histogram analysis on a 5-by-5 grid copes better with fast camera movements, zoom gestures, and similar scene discrimination (i.e., problems that arise in handheld video analysis. The proposed algorithm consists of 4 major steps:

- A conversion to gray scale frames;
- A rasterization (splitting) of each frame in a 5-by-5 grid and a calculation of the histogram for each cell in the grid;
- A correlation based temporal analysis of each histogram;
- A calculation of the amount of changed cells. If this is larger than an experimentally defined threshold the transition is seen as a shot.

The histogram based shot detection has an outstanding performance on detecting hard and gradual shots, which makes it suitable to process all kinds of video content (i.e., handheld and CCTV data streams). The only limitation is that it does not handle similar scenes very well, but this is solved in Section 3.3. In order to cope with gradual shots, like blends and fades, we also count the amount of frames that have been passed since the last detected transition. If that amount is too low and a new transition is found, we consider it to be the same transition. This way we successfully manage to detect both gradual- and abrupt scene transitions.

3.2 No-reference keyframe quality analysis

Currently, multiple image quality metrics are available [17]. However, most of them need a reference image or they are not suitable for real-time quality measurement. In our methodology, based on the detected shot boundaries, the frame with the highest quality within each shot is chosen as a representative keyframe. The quality is estimated using a weighted set of three no-reference quality metrics that are suitable for real-time images as proposed by Vandecasteele et al. [9]. The proposed no-reference exposure, contrast and sharpness metrics are evaluated on a variety of video footages.

3.3 Similarity clustering

In general, a video scene consists of many shots that are visually similar (i.e., taken in the same scene setting). On the one hand, removing these redundant and unnecessary frames improves the summarization visualization on a small screen or on a tablet. On the other hand, clustering the similar keyframes will decrease the time for tag generation i.e., an automatic annotation can be performed on all object representations within the same cluster.

The clustering process is done in three steps. First, we perform a global feature extraction with CNN based learned features as proposed by Zagoruyko et al. [18]. Subsequently, we do a feature reduction by using the principal component analysis (PCA) technique. Finally, we use k-means clustering with the L2-distance of the reduced features of the keyframe. Furthermore, it is important to remark that if the amount of clusters is too high, redundant keyframes will appear. On the other hand, if the amount of clusters is too low, outliers (i.e., very unique keyframes or single scene shots) will not be shown.

4 VIDEO AND FRAME RETRIEVAL

The set of keyframes (from different camera viewpoints and camera sources) could still be too versatile and to improve the fast exploration two additional mechanisms are proposed: (I) an automated importance scoring based on the scene content and (II) a semantic web-based querying tool as proposed by Verstockt et al. [19]. Furthermore, the tags that are generated from the scene understanding process are used as input for both mechanisms. Figure 2 gives a schematic overview of the scene content understanding framework.

4.1 Automated scene content understanding

4.1.1 Scene type detection

Over the last decade, several literature studies have already shown impressive results regarding the classification of indoor and outdoor scenes [20], and more recent studies have added the ability to predict more detailed attributes, such as the scene type, e.g., kitchen, living room, or bedroom. Zhou et al. [21] constructed a new scene-centric database named Places, containing over 7 million labeled pictures of scenes distributed over 365 classes. Currently, the state-of-the-art systems for scene classification are based on CNNs, mostly trained using a transfer learning approach on the ImageNet models [21]. For the scene classification task the current highest top-5 accuracies achieved on this Places dataset are 85,08 percent with the RESNET architecture [22].

4.1.2 Visual and thermal object localization

Despite the broad application of handheld thermal imaging cameras in firefighting, its usage is mostly limited to subjective interpretation by the person carrying the device. As remedies to overcome this limitation, object localization and classification mechanisms [23, 24] could assist the fireground understanding and help with the automated localization, characterization and spatial-temporal (spreading) analysis of the fire. An automated understanding of thermal images can transform the conventional knowledge-based firefighting into sensor-driven based firefighting. In this paper, transfer learning is applied to multi-labeling convolution neural network architectures for object localization and recognition in monocular visual, infrared images. More details are given in the work of Vandecasteele et al. [25].

4.1.3 Room configuration estimation optimization

Intuitively, individual predictions of objects and relationships can benefit from their surrounding context (i.e., the probability to predict a bed, given that he room is a sleeping room is much higher compared to a kitchen). In our work we optimized the scene classification results by exploiting the object detections. First we learned the object co-occurrence from the ObjectNet3D dataset. Some preprocessing was necessary to remove the objects that appear only once in the dataset. Furthermore the classes *'floor', 'ceiling'* and *'wall'* were removed from the object dataset as they are not descriptive for a specific room type. The object co-occurrence will assist the thermal object detection by indicating the objects that are related to the detected classes. For example if a chair is detected, there is a 60 percent chance that there will be a table in the same scene. This can then be incorporated in the fire load calculation.

4.1.4 Fire and smoke characteristics

Video based fire detection and analysis with cameras has been discussed several times in literature over the past years [26]. Different features, such as color, pixel disorder, wavelet analysis are combined with simple linear classifiers or with more advanced deep learning mechanisms [27]. However, the focus in literature was mainly on the detection, the propagation of the smoke, the height of the smoke layer or the visibility for example, were investigated less. Within our framework, the smoke and fire probability are calculated with the techniques of Frizzi et al. [27]. Finally, the visibility on each frame is calculated with the techniques of Vandecasteele et al. [6], the visibility is thereby scored between High-visibility (5) up to No-visibility (0).

4.2 Importance scoring

The content of the scene is important because it includes useful information that attracts the viewer (e.g., the decision maker). Rationally, the objects and locations that are shown for a longer period of time in the handheld video have a higher chance to be of importance. It is important to remark that further user-driven evaluations will be necessary to prove this statement. Furthermore, since there are tags pertaining to all the frames it is necessary to alter them and to preserve only the most informative words that define the main semantic components. This is done as follows. First the tags are preprocessed by tokenization and part-of-speech tagging. After the extraction, the saliency score for each subset s is calculated according to the words they contain. In combination with the similarity score (e.g., two visualized keyframes should be sufficiently dissimilar), the importance scoring is used to automatically present a couple of keyframes to the decision makers. Still, user evaluations and annotations could help improve and validate the automated video selecting. Besides the automated selecting of the keyframes, we propose a semantic querying system that allows to manually find the most related frames.

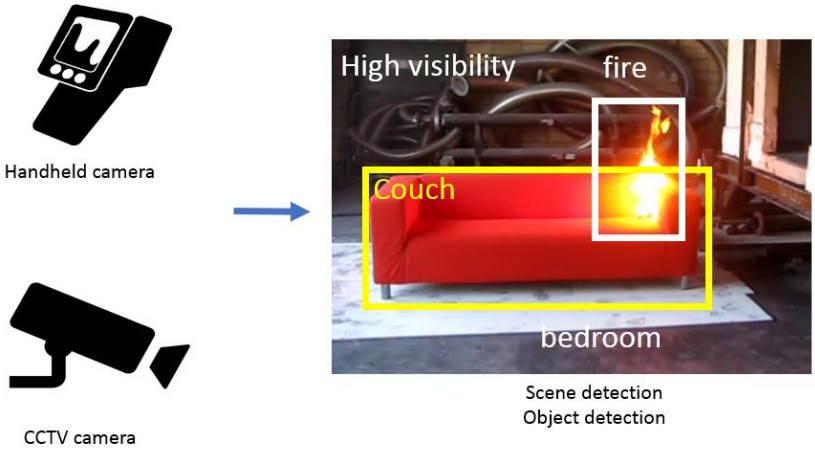

Figure 2. Automated scene content understanding from video footages (e.g., handheld and CCTV cameras).

4.3 Semantic based querying

The web-based tool that could be used to explore the enriched (i.e., the annotated) keyframes is a metadata filtering and clustering service. The querying mechanism can be activated in three different ways:

- A textual input where the user can type the tag that he wants. In case the input is not in the list of predefined tags, the tag with the closest path distance (i.e., the highest similarity) according to Wordnet [28] is selected. However, other semantic distance metrics can easily be integrated.

- A drop-down list with predefined tags (e.g., places, fire-related topics, evacuation keywords), where the user can select the necessary tags (e.g., show all the keyframes annotated with kitchen and fire).

- An interactive, hierarchical ontology visualization where the user can select and easily search for the best matching tags. Furthermore, if a tag on a higher semantic level in the hierarchy is selected, all the underlaying and related tags are selected (e.g., if the tag opening is selected, the tags door and window are likewise selected).

> *An ontology is a representation, formal naming, and definition of the categories, properties, and the relations between the concepts, data, and entities over one or more domains.*

The semantic similarity is currently calculated based on the ontology of Zhang et al. [29], but new ontologies can easily be integrated. Zhang et al. proposed a street scene ontology for qualitative understanding of outdoor scenes. This is valuable for large multi-disciplinary fire incidents as it contains building elements (e.g., floor, window, wall, column), but also construction, land and terrain elements. Some alternative ontologies that can be used are, for example, the work of Kadar et al. [30] who created a database of 100 scene categories (e.g., classroom, bathroom, bedroom, alley) derived from human vision. This ontology could be valuable in indoor firefighting scenarios. Jaoa et al. [31] created an ontology on how scene situations progress in time. The FIRE ontology was created in order to represent the set of concepts about the fire occurring in natural vegetation, its characteristics, causes and effects. Similar concepts and effects are found in indoor fire situations. Poveda et al. [32] created an ontology for designing and validating emergency plans and the sensor, users and furniture connections are highly valuable for our framework.

5 RESULTS AND DISCUSSION

Murugan et al. [33] proposed various methods for video summarization for surveillance applications still latter discussion was given on the exportability of the video footages. Within this Section a subjective evaluation is performed on our proposed mechanisms. Video footages from multidisciplinary incidents are taken as input for our evaluation. Figure 3 gives a schematic overview. First the video footages are selected (currently, this is a manual task, but the integration of online IP-cameras should be easily possible). Secondly, the keyframe extraction, similarity removal and no-reference analysis is used to select the most representative keyframes. Thirdly, the semantic tag understanding process is elaborated to automatically generate tags for each frame. Finally, the exploration and selection tools are used to get a fast overview of the current state of the incident and the actions.

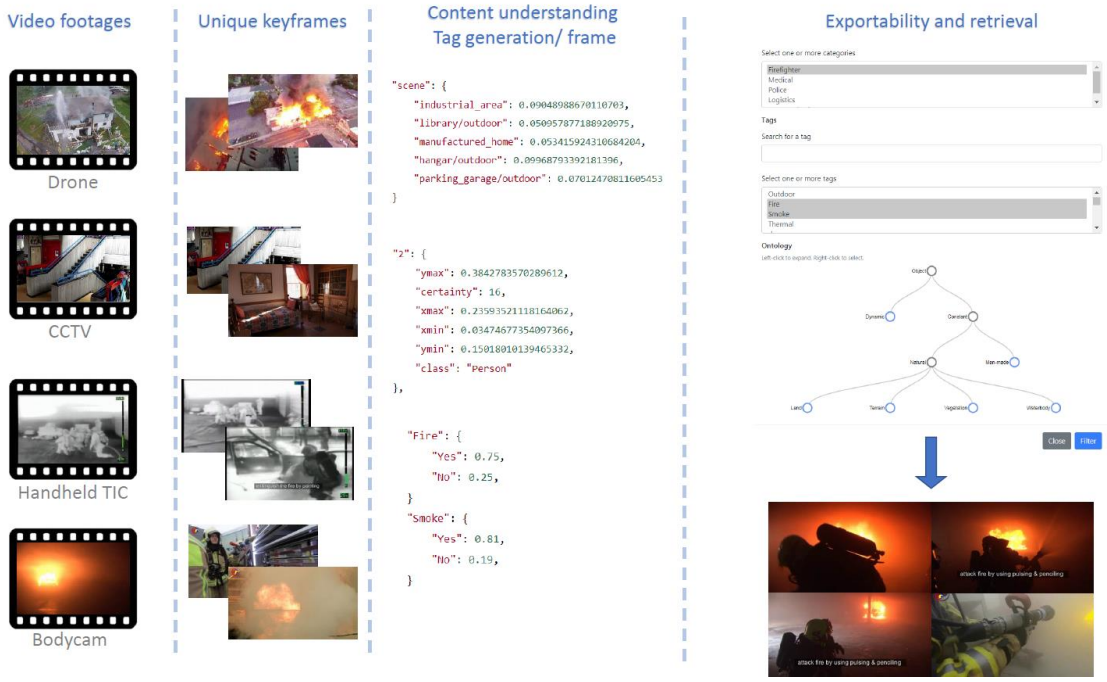

Figure 3. Schematic overview of the video footage analysis framework, first the footage linking, secondly the keyframe generation, thirdly the content understanding tool and finally the frame retrieval tool.

Some improvements for the tag generation are the classification into categories, for example an additional classifier into medical, police, fire service. Furthermore, it is valuable to add the source of the video footage as a tag (e.g., handheld camera first crew, TIC second crew, drone footage). Finally, as indicated earlier in this section, the video summarization and retrieval building blocks require further user-driven evaluations. Still the feasibility of the proposed building blocks is shown in this paper. Real-fire experiments that are recorded from different point-of-views and with different resources are highly valuable for further evaluation and improvement.

6 CONCLUSION

This paper presented the user-needs and data restrictions specific for fire incident management. The insights were gained through the analysis of a questionnaire launched in the firefighting community of Belgium. This evaluation revealed that people are visually-oriented and that video footages are great to gain insights into a problem. Still, people can only process 7 image inputs simultaneous and for that reason, the video summarization framework, consisting of shot detection, frame quality and similarity analysis was proposed. Subsequently, in order to facilitate the video search process, the video and frame retrieval mechanism was clarified and semantic tag based querying on an existing ontology map was initiated. Future work will focus on video analytics to perform person recognition, crowd analysis(e.g., detecting the amount of people and their behavior) and anomaly detection on global scale. Maybe in the future, even before a fire incident will occur the video analysis will notify suspicious behavior and faster interventions could reduce the economic and material damage. Subsequently, more user-evaluations and usability testing will be necessary to make a complete product out of our proposed framework. Therefore the next step is to organize an usability evaluation using the System Usability Scale (SUS)[34]. Finally a closed field test (limited set of users) and a open field test will give insights in the required development steps.

7 ACKNOWLEDGMENTS

The research activities as described in this paper were funded by Ghent University through GOA project BOF16/GOA/004.

8 REFERENCES

[1] M.M. Chun, Contextual cueing of visual attention, Trends in cognitive sciences, 4 (2000) 170-178.
[2] O. Janssens, Data-driven performance monitoring, fault detection and dynamic dashboards for offshore wind farms, in, Ghent University, 2017.
[3] F. Ungerer, H.-J. Schmid, An introduction to cognitive linguistics, Routledge, 2013.
[4] D. Dörner, H. Schaub, Errors in planning and decision-making and the nature of human information processing, Applied psychology, 43 (1994) 433-453.
[5] T. Beji, S. Verstockt, R. Van de Walle, B. Merci, On the use of real-time video to forecast fire growth in enclosures, Fire Technology, 50 (2014) 1021-1040.
[6] F. Vandecasteele, B. Merci, S. Verstockt, Reasoning on multi-sensor geographic smoke spread data for fire development and risk analysis, Fire Safety Journal, 86 (2016) 65-74.
[7] N. Li, Z. Yang, A. Ghahramani, B. Becerik-Gerber, L. Soibelman, Situational awareness for supporting building fire emergency response: Information needs, information sources, and implementation requirements, Fire safety journal, 63 (2014) 17-28.
[8] V. Nunavath, A. Prinz, T. Comes, J. Radianti, Representing fire emergency response knowledge through a domain modelling approach, (2016).
[9] F. Vandecasteele, K. Vandenbroucke, D. Schuurman, S. Verstockt, Spott: On-the-Spot e-Commerce for Television Using Deep Learning-Based Video Analysis Techniques, ACM Trans. Multimedia Comput. Commun. Appl., 13 (2017) 1-16.
[10] F. Vandecasteele, B. Merci, S. Verstockt, Fireground location understanding by semantic linking of visual objects and building information models, Fire Safety Journal, 91 (2017) 1026-1034.
[11] M. Ajmal, M.H. Ashraf, M. Shakir, Y. Abbas, F.A. Shah, Video summarization: techniques and classification, in: International Conference on Computer Vision and Graphics, Springer, 2012, pp. 1-13.
[12] F. Wang, C.-W. Ngo, Summarizing rushes videos by motion, object, and event understanding, IEEE Transactions on Multimedia, 14 (2012) 76-87.
[13] L. dos Santos Belo, C.A. Caetano Jr, Z.K.G. do Patrocínio Jr, S.J.F. Guimarães, Summarizing video sequence using a graph-based hierarchical approach, Neurocomputing, 173 (2016) 1001-1016.
[14] L. Baraldi, C. Grana, R. Cucchiara, Shot and scene detection via hierarchical clustering for re-using broadcast video, in: International Conference on Computer Analysis of Images and Patterns, Springer, 2015, pp. 801-811.
[15] X. Qiu, S. Jiang, H. Liu, Q. Huang, L. Cao, Spatial-temporal attention analysis for home video, in: 2008 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2008, pp. 1517-1520.
[16] M.R. Mishra, S. Singhai, M. Sharma, A Comparative based study of Different Video-Shot Boundary Detection algorithms, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2 (2013) pp: 282-289.
[17] K. Joy, E.G. Sarma, RECENT DEVELOPMENTS IN IMAGE QUALITY ASSESSMENT ALGORITHMS: A REVIEW, Journal of Theoretical & Applied Information Technology, 65 (2014).

[18] S. Zagoruyko, N. Komodakis, Learning to compare image patches via convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4353-4361.
[19] S. Verstockt, S. Nop, F. Vandecasteele, T. Baert, N. Van de Weghe, H. Paulussen, E. Rizza, M. Roeges, UGESCO-A Hybrid Platform for Geo-Temporal Enrichment of Digital Photo Collections Based on Computational and Crowdsourced Metadata Generation, in: Euro-Mediterranean Conference, Springer, 2018, pp. 113-124.
[20] J. Luo, A. Savakis, Indoor vs outdoor classification of consumer photographs using low-level and semantic features, in: Image Processing, 2001. Proceedings. 2001 International Conference on, IEEE, 2001, pp. 745-748.
[21] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Advances in neural information processing systems, 2014, pp. 487-495.
[22] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.
[23] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440-1448.
[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, Springer, 2016, pp. 21-37.
[25] F. Vandecasteele, B. Merci, A. Jalalvand, S. Verstockt, Object localization in handheld thermal images for fireground understanding, in: Thermosense: Thermal Infrared Applications XXXIX, International Society for Optics and Photonics, 2017, pp. 1021405.
[26] A.E. Çetin, K. Dimitropoulos, B. Gouverneur, N. Grammalidis, O. Günay, Y.H. Habiboğlu, B.U. Töreyin, S. Verstockt, Video fire detection–review, Digital Signal Processing, 23 (2013) 1827-1843.
[27] S. Frizzi, R. Kaabi, M. Bouchouicha, J.-M. Ginoux, E. Moreau, F. Fnaiech, Convolutional neural network for video fire and smoke detection, in: Industrial Electronics Society, IECON 2016-42nd Annual Conference of the IEEE, IEEE, 2016, pp. 877-882.
[28] T. Pedersen, S. Patwardhan, J. Michelizzi, WordNet:: Similarity: measuring the relatedness of concepts, in: Demonstration papers at HLT-NAACL 2004, Association for Computational Linguistics, 2004, pp. 38-41.
[29] F. Zhang, D. Zhang, Y. Liu, H. Lin, Representing place locales using scene elements, Computers, Environment and Urban Systems, (2018).
[30] I. Kadar, O. Ben-Shahar, SceneNet: A perceptual ontology for scene understanding, in: European Conference on Computer Vision, Springer, 2014, pp. 385-400.
[31] J.P.A. Almeida, P.D. Costa, G. Guizzardi, Towards an Ontology of Scenes and Situations, in: 2018 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), IEEE, 2018, pp. 29-35.
[32] G. Poveda, E. Serrano, M. Garijo, Designing emergency management services by ontology driven social simulation, IT CoNvergence PRActice, 3 (2015) 17-32.
[33] A.S. Murugan, K.S. Devi, A. Sivaranjani, P. Srinivasan, A study on various methods used for video summarization and moving object detection for video surveillance applications, Multimedia Tools and Applications, (2018) 1-18.
[34] J. Brooke, SUS-A quick and dirty usability scale, Usability evaluation in industry, 189 (1996) 4-7.