An Approach for CCTV Contents Filtering Based on Contextual Enrichment via Spatial and Temporal Metadata

Franck Jeveme Panta, André Péninou, Florence Sèdes

An Approach for CCTV Contents Filtering Based on Contextual Enrichment via Spatial and Temporal Metadata

Relevant Video Segments Recommended for CCTV Operators

Franck Jeveme Panta
IRIT
Université Paul Sabatier
Toulouse, France
franck.panta@irit.fr

André Péninou
IRIT
Université Paul Sabatier
Toulouse, France
andre.peninou@irit.fr

Florence Sèdes
IRIT
Université Paul Sabatier
Toulouse, France
florence.sedes@irit.fr

ABSTRACT

With the constant evolution of CCTV cameras deployed in major cities to ensure the citizens' security, CCTV operators have to watch a huge amount of video when they are searching for scenes, objects, or target persons. Watching or processing some video sequences can be useless for several reasons: content is unsuitable for operators' needs, unusable shooting conditions, etc. Filtering useless content can be an efficient way for operators to save time. In this paper we propose an approach for CCTV contents filtering based on contextual information in order to provide CCTV operators with video sequences of interest. The proposed approach takes into account many sources of contextual information such as: open data, social media, mobility, geolocation, and crowdsourcing. We provide an analysis of contextual information relevant for this approach. Since interoperability is one of the main problems of context-based approaches, we propose a generic data model of contextual information used in our approach in order to tackle this issue. Based on this data, we propose a framework architecture for relevant video segments recommendation.

KEYWORDS

Context; Contextual Enrichment; Spatial an Temporal Metadata; CCTV Systems; Filtering; Querying.

1 INTRODUCTION

Context or contextual information is defeined as any information (implicit or explicit) that can be used to describe an entity (person, physical/informatic object, task, concept, etc.) in a given situation (spatial and temporal). In this paper, we introduce the concept of **"contextual enrichment"** (via spatial and temporal metadata), which is the process of identifying and modelling relevant contextual information for an entity in a system in order to address some operational or functional requirements for this system. Our work applies to video surveillance systems (CCTV systems) and can be extended to other applications that process huge amount of data such as teledetection, medical imaging, etc. Our goal is to provide CCTV operators with a recommended system for relevant video segments or video sequences of interest via an approach based on filtering and intelligent querying of huge amount of data (video and metadata) through contextual information.

In a use case such as ours (CCTV systems), which processes a huge amount of data (videos) and metadata, **contextual enrichment** can improve the filtering process (filtering from other points of interest) and video querying process in order to reduce the volume of video to watch or to analyze. With the constant evolution of CCTV cameras deployed in major cities to ensure the citizens' security, CCTV operators have to watch/analyze a huge amount of video when they are searching for scenes, objects, or target persons. Watching/analyzing some video sequences can be useless for several reasons: content is unsuitable for operators' needs, unusable shooting conditions, etc. For instance, environmental problems such as the presence of atmospheric particles (rain, fog, cloud, dust, pollution, etc.) can alter images quality and prevent visibility. A degraded image can have a negative impact on its interpretation by the human eye or by automatic processing algorithms. Therefore, watching or processing these images is a waste of time and resources. In this case, the approach we are proposing in this paper will reduce the research and watching time, either by filtering out video sequences that contain unusable images or by offering video sequences to CCTV operators with confidence levels of usability; which gives the CCTV operator the operating choice. **Contextual enrichment** can also address some inherent limitations of CCTV data such as their incompleteness, or help in interpretation/understanding of these data.

There are many solutions based on video content analysis algorithms that have been proposed for the search of video sequences of interest. However, these solutions require improvements to be operational in some applications. Our approach is complementary to existing approaches, and aims to significantly reduce the volume of video to watch/analyze and to facilitate the querying of these videos through contextual information.

Contextual information can be considered as a set of metadata collected from different sources. These metadata formats,

types and representation can differ from one source to another, making them heterogeneous and limiting their aggregation and interrogation. **Contextual enrichment** is a complex process, since it must take into account the spatio-temporal relationships between videos and context information. Thus, **contextual enrichment** leads to a problem of integration of heterogeneous (meta)data with a spatial-temporal reasoning capacity. In order to tackle interoperability issues, we propose a generic data model allowing efficient management and interoperability of contextual information collected from different sources. The proposed data model is scalable and enables integration of new contextual information. Scalability is one of the main challenges for context-aware applications. In addition to this generic and scalable data model, we propose a framework architecture in order to recommend video sequences of interest to CCTV operators.

The remainder of this paper is organized as following: section 2 reviews the related works; Section 3 presents a description of our approach; Section 4 shows the framework architecture for relevant video segments recommendation. Finally, in section 5, we discuss and conclude with suggestions for future research.

2 RELATED WORK

We define the recommendation of interesting video segments or video sequences of interest here as a kind of filtering and querying large volumes of video. Therefore, it is an issue of processing huge amounts of data and metadata, and to address many limitations such as the heterogeneity of these meta(data) and time constraints related to their processing. Many works on automatic analysis of video content has been done in order to propose solutions for retrieving video sequences of interest. Our goal is not to provide an exhaustive overview of these works, nor to make a direct contribution to video content analysis itself : the main goal is to propose a solution based on contextual information, which are currently not or a little used in this field.

Most of the proposed studies focus on the development of video content analysis tools in oder to detect and/or track objects [2], to recognize actions [10], events [4] or scenes [6], to analyze the behavior of human crowds [14], etc. Generally, approaches used in these works are based on shot boundary detection [7], key frame extraction [9] and scene segmentation [5]. The limitations with these approaches are the following : (i) the huge amount of data that processing is time consuming, (ii) the difficulty to develop generic video content analysis algorithms (most of the video analysis algorithms have been developed for specific applications and trained with well selected data, therefore it remains difficult to use these algorithms for applications that they have not been trained for), and (iii) the lack of robustness to environmental variations and the complexity of the scenes.

During the last few years, many approaches for video metadata processing have been proposed. These approaches usually aim to describe and facilitate the extraction of relevant metadata related to videos, and use these metadata to simplify the videos management processing. An effective and efficient method for constructing and extracting descriptive metadata for Web videos is presented in [1]. The authors proposed a metadata model that can help the users to search and personalize video data in an efficient ways. Proposed model is only appropriate for the classification of web videos. Metadata modeling for distributed multimedia document management and interoperability of video surveillance systems has been addressed in [8]. Authors proposed a metadata format that enables multiple systems to exchange data. Data exchange is not highly valued in our approach, but it's used in the background.

Several other works have focused on metadata to search for video sequences of interest. [3], [12] proposed approaches that aims to help human video surveillance operators in the research of relevant video sequences by offering them a set of cameras that could have filmed a desired scene. [11] proposed a solution for video content filtering using metadata from automatic image processing algorithms. These works do not take into account contextual information that can enrich existing metadata and improve results.

From a contextual perspective, several researchers used the context to develop intelligent applications in fields ranging from ubiquitous and mobile computing to artificial intelligence [13]. Context has been introduced by several researchers in the literature. Generally, context is defined in the literature as any information that can be used to characterize the situation of an entity (people, objects, places). Although this definition is sufficient for context-aware computing, it does not necessarily define context from the perspective of CCTV systems or other large-scale and multi-application environments.

3 PROPOSED APPROACH

This section develops our contextual enrichment approach for filtering and intelligent querying of large volumes of video in order to provide CCTV operators with video sequences of interest. An overview of the proposed approach is presented in the following steps: *Step 1:* analysis of relevant contextual information sources for filtering and querying video content; *Step 2:* modeling of different contextual information sources (Open data information, social media information, mobility and geolocation information, crowdsourcing information and sensors information); *Step 3:* definition of filtering and querying mechanism based on contextual information;

3.1 Contextual information sources

The use of contextual information depends on the objectives and the area of work. In this study, the target area is video surveillance and the objective is to offer CCTV operators relevant video segments or video sequences of interest. For such purposes, contextual information can be used in two ways, namely: (i) for videos filtering, and (ii) to address some inherent limitations of CCTV data such as their incompleteness, or help in interpretation of these data. In the following, we present several contextual information sources and describe how these contextual information can be used.

3.1.1 Open data.

Open Data are data for which access is totally public and free of charge, as well as exploitation and reuse. Open Data

offer many opportunities to extend human knowledge and create new products and quality services. Open Data can be relevant in many sectors, for various groups of people and organizations. For each category of data, the benefits can be specific and can vary according to the fields and applications. The contribution of Open Data can vary according to the needs of the video surveillance operator and the different open data sources (geographical data, weather, environment, finance, etc.). For example, if the operator's need is to watch or process high-quality videos as a priority or only, Open Data such as weather can be used to filter videos, since weather conditions affect the quality of images when they are captured by CCTV cameras. A non-exhaustive list of contextual information used for video filtering based on image quality is shown below : (i) Presence of atmospheric elements (rain, fog, pollution, dust, etc.); (ii) low lighting of the scene (darkness, night); (iii) occultation of the scene.

3.1.2 *Social media.*

Social media are supports for the massive diffusion of information using all existing formats (text, image, video, audio...) and enabling social interaction. Social networks (Facebook, Twitter, Instagram, LinkedIn) and blogs are some examples of social media platforms. Social media data are additional sources of information that can be used to facilitate or improve filtering and querying of large volumes of video. For example, videos coming from CCTV cameras do not always contain enough information to analyze a situation or event/incident. Then, using other sources of information becomes a requirement for CCTV operators. Social media is therefore an enriching source of information that can facilitate detection, identification, localization in CCTV videos and also interpretation or deduction of new knowledge through the analysis of user profiles, links (subscribers, friends, groups) and shared content (texts, images, videos).

3.1.3 *Mobility .*

Mobility, now called intelligent/smart mobility refers to all transportation means, technologies and applications of transportation and communication implemented to ensure the movement (reservation, route calculation,...) of users by guarantying efficiency (e.g. speed control), safety (e.g. personal safety) and comfort (e.g. traffic information). Mobility systems services (e.g. route planning, navigation, ticketing) are consumed by users or costumers and produce a set of data that can be used for other purposes or in other areas. Data from mobility services such as location information, trajectories, spatial and temporal information, movements of entities (users, vehicles, etc.), connectivity and access data to different services can be used by CCTV operators in order to facilitate videos interpretation, which do not always contain the relevant information needed to identify or track an entity.

3.1.4 *Geolocation (devices location).*

Geolocation refers to a set of ways to locate people or objects (e.g. vehicles, equipment) on a map or a plan using their spatial coordinates. Geolocation applications are under development and offer many valuable services such as real-time and historical tracking, location (incident location), navigation, detailed trajectories on maps, location of objects in a specific area or nearby, etc. Location information generated by geolocation systems and applications can be an external source of contextual information for tracking people and objects (vehicles, equipment, etc.) in CCTV videos.

3.1.5 *Crowdsourcing.*

Crowdsourcing is a type of online collaborative activity that involves a large number of people in order to work on a specific task or set of tasks. Crowdsourcing is also used to collect information from the general public and use these information to perform some tasks. For CCTV metadata and video content, data collected through crowdsourcing can be considered as a valuable source of information. For example, let us consider a crowdsourcing platform set up by the police and that allows the public to publish information (audio, video, images, etc.) about an incident that occurred. This platform will be able to create value by aggregating or crossing its content with CCTV video, and provide investigative opportunities. All these described contextual information sources generate a set of contextual metadata that can be used in video content filtering and querying algorithms.In the following, we propose a modeling of this contextual information in order to address interoperability issues.

3.2 Contextual metadata modeling

A system that considers contextual information for decision-making is known as a context aware system. Thereby, a CCTV context-aware system should accurately perceive data generated by camera (video and camera metadata) and integrate them with contextual information to provide improved services and functionalities. By contextual information we refer to : environmental information (weather, pollution, time, etc.), social media information (contents, profile, events, etc.), mobility information (devices location, trajectories, etc.), crowdsourcing information (shared content), and sensors information (camera location). One of the requirements of such a CCTV context-aware system is to ensure interoperability of multisource and heterogeneous contextual information. In this section, a data model is proposed for each source of contextual information, and then a generic data model integrating all the different sources is proposed at the end. This generic data model addresses the interoperability issue of contextual information.

3.2.1 *Camera information.*

Camera location is a key element in our approach, it enables to set the spatial link with other contextual information. A camera located at a given position, with a specific orientation and installation can capture a given area. Depending on where the camera is deployed, on a fixed place (e.g. in a street, a subway, a room) or on a mobile object (e.g. in a bus), it is called a fixed or mobile camera. Other descriptive metadata of the camera such as field of view and image parameters (resolution, contrast, brightness, etc.) are relevant in some situations. Figure 1 shows the relationships between camera, field of view and location, specifying that a camera can have several positions with variable fields of view at different times. Information about camera parameters are also represented in

the data model. The data model is scalable and can take into account new sensors (inheriting "SENSOR" class) and their specifications.

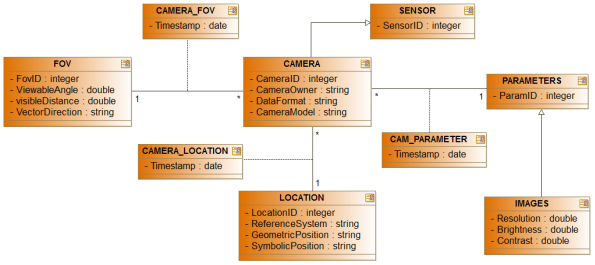

Figure 1: Camera information.

3.2.2 *Open data information.*

We are interested here in environmental information and events that have occurred such as rain, wind, fog, pollution, etc. The data model (Figure 2) represents information such as event location and occurrence time. Data model can take into account events that were not expected when designing (inheriting "ENVIRONMENTAL_DATA" class).

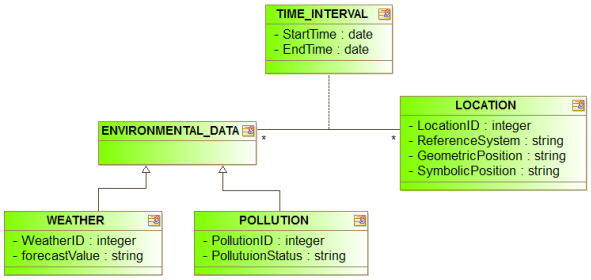

Figure 2: Open Data information.

3.2.3 *Social media information.*

Information about social media can be summarized by content (videos, audio, texts, etc.), relationships (friends, followers, groups, etc.) and events (birthdays, parties, etc.). The links between all this information are represented on the data model shown at Figure 3. Types and dates of content publication, locations and events dates are taken into account in the data model. Information on individuals' participation in an event, their membership in groups, as well as their different publications can be retrieved by querying the data model.

3.2.4 *Mobility information and geolocation.*

Transportation applications and services generate a set of contextual information such as routes (departure and arrival stations) and timetables (departure and arrival times) of vehicles (buses, trains, metro). All these in formations are taken into account in the data model (Figure 4) and can be queried for specific purposes. Information related to devices location are also taken into account in the data model. We can see on Figure 4 that a device is located in different places at different times, and can be attached to a person.

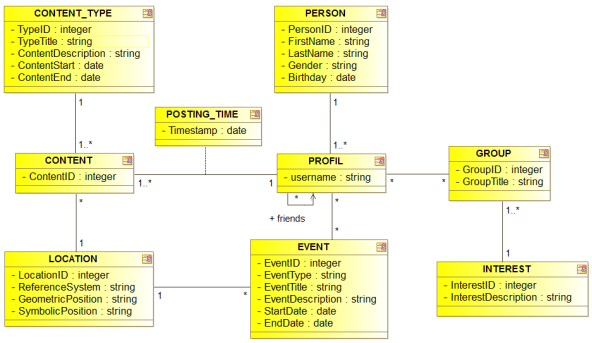

Figure 3: Social media information.

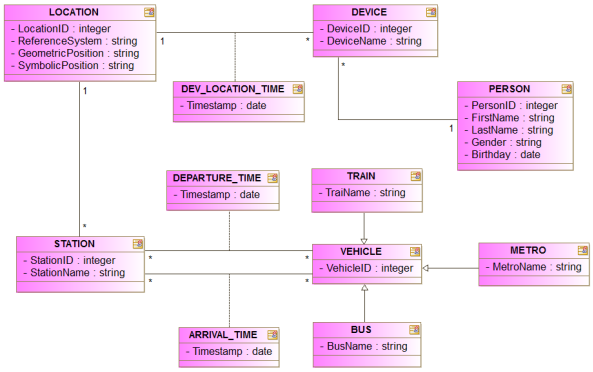

Figure 4: Mobility information.

3.2.5 *Crowdsourcing information.*

Shared information on crowdsourcing platforms can be summarized in terms of content (videos, audio, texts, etc.) and associated tasks. Figure 5 shows that a specific task can have several publications with different content and published by different profiles.

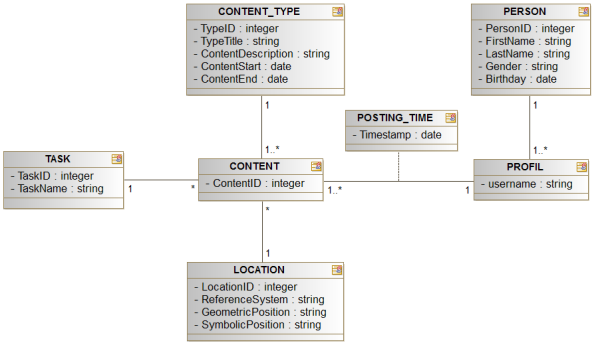

Figure 5: Crowdsourcing information.

All these metadata are aggregated in the generic data model shown on Figure 6. This generic data model allows integration of different contextual information sources and therefore interoperability of heterogeneous contextual information (one of the goals of this paper). Information about crowdsourcing

are almost similar to information about social media (a crowdsourcing platform is considered as a type of social media). So these two information sources are merged into the generic data model and will be distinguishable at the instantiation of the model. We can observe on this data model that interconnection of different information sources is done by spatial information ("LOCATION" class), and spatial information is always connected to temporal information, which further demonstrates that spatio-temporal reasoning must be taken into account in our approach.

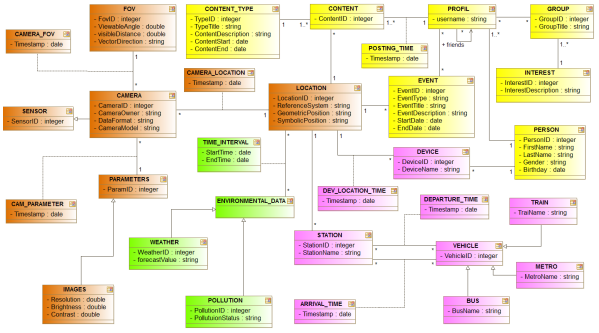

Figure 6: Generic data model.

4 FRAMEWORK ARCHITECTURE

Figure 7 illustrates the framework architecture that we have developed and that implements our filtering and querying mechanism. The main modules of our prototype are: (i) Query interpreter: interprets (transforms) the query submitted by the user into a spatio-temporal query; (ii) Querying mechanism: implements our querying algorithms based on contextual information; (iii) Filtering : implements our filtering method based based on contextual information; (iv) Metadata repository: contains metadata collections based on the generic proposed model. User interface and Metadata sources are two external modules.

5 CONCLUSION AND FUTURE WORK

In the context of over-expansion of information generated by CCTV systems, recommendation of video sequences of interest is a frequent and challenging issue in video analysis. In this paper, we proposed a contextual based approach that takes into account open data information, social media information, mobilty and geolocation information, crowsourcing information and sensors information. We proposed a generic data model that allows integration of different contextual information sources and therefore interoperability of heterogeneous contextual information. We presented architecture that we have developed and that implements our filtering and querying mechanism which will be presented in future work and validated in the French national project FILTER2.

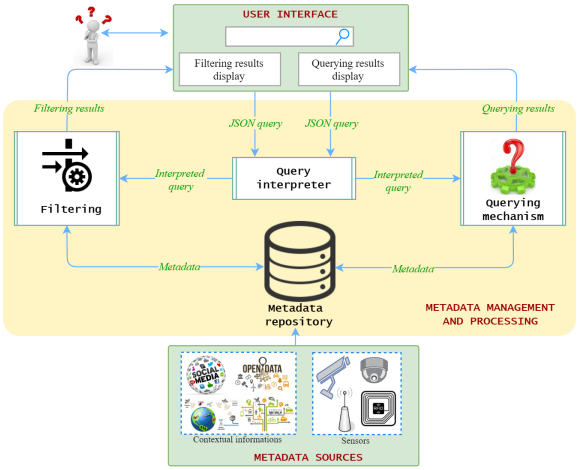

Figure 7: Architecture of CCTV context-aware system

REFERENCES

[1] Siddu P Algur, Prashant Bhat, and Suraj Jain. 2014. Metadata Construction Model for Web Videos: A Domain Specific Approach. *International Journal of Engineering and Computer Science* 3, 12 (2014).

[2] Yi-Ling Chen, Tse-Shih Chen, Tsiao-Wen Huang, Liang-Chun Yin, Shiou-Yaw Wang, and Tzi-cker Chiueh. 2013. Intelligent urban video surveillance system for automatic vehicle detection and tracking in clouds. In *2013 IEEE 27th international conference on advanced information networking and applications (AINA)*. IEEE, 814–821.

[3] Dana Codreanu, Andre Peninou, and Florence Sedes. 2015. Video Spatio-Temporal Filtering Based on Cameras and Target Objects Trajectories–Videosurveillance Forensic Framework. In *2015 10th International Conference on Availability, Reliability and Security*. IEEE, 611–617.

[4] David Gerónimo and Hedvig Kjellström. 2014. Unsupervised surveillance video retrieval based on human action and appearance. In *2014 22nd International Conference on Pattern Recognition*. IEEE, 4630–4635.

[5] Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. 2011. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41, 6 (2011), 797–819.

[6] HC Lee and EM Pagliaro. 2013. Forensic evidence and crime scene investigation. *Journal of Forensic Investigation* 1, 1 (2013), 1–5.

[7] Xue Ling, Li Chao, Li Huan, and Xiong Zhang. 2008. A general method for shot boundary detection. In *2008 International Conference on Multimedia and Ubiquitous Engineering (mue 2008)*. IEEE, 394–397.

[8] Ana-Maria Manzat, Romulus Grigoras, and Florence Sèdes. 2010. Towards a user-aware enrichment of multimedia metadata. In *Workshop on Semantic Multimedia Database Technologies*.

[9] Azra Nasreen and G Shobha. 2013. Key frame extraction from videos-A survey. *International Journal of Computer Science & Communication Networks* 3, 3 (2013), 194.

[10] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. 2008. Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision* 79, 3 (2008), 299–318.

[11] Franck Jeveme Panta, Mahmoud Qodseya, André Péninou, and Florence Sèdes. 2018. Management of Mobile Objects Location for Video Content Filtering. In *Proceedings of the 16th International Conference on Advances in Mobile Computing and Multimedia*. ACM, 44–52.

[12] Franck Jeveme Panta, Geoffrey Roman-Jimenez, and Florence Sèdes. 2018. Modeling metadata of CCTV systems and Indoor Location Sensors for automatic filtering of relevant video content. In *2018 12th International Conference on Research Challenges in Information Science (RCIS)*. IEEE, 1–9.

[13] Albrecht Schmidt, Michael Beigl, and Hans-W Gellersen. 1999. There is more to context than location. *Computers & Graphics* 23, 6 (1999), 893–901.

[14] B Yogameena and K Sindhu Priya. 2015. Synoptic video based human crowd behavior analysis for forensic video surveillance. In *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*. IEEE, 1–6.