Spatio-Temporal Metadata Querying for CCTV Video Retrieval: Application in Forensic

Franck Jeveme Panta, Mahmoud Qodseya, Geoffrey Roman Jimenez, André Péninou, Florence Sèdes

Spatio-Temporal Metadata Querying for CCTV Video Retrieval : Application in Forensic

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

Mahmoud Qodseya
IRIT
Université Paul Sabatier
Toulouse, France
mahmoud.qodseya@irit.fr

Geoffrey Roman-Jimenez
IRIT
Université Paul Sabatier
Toulouse, France
geoffrey.roman-jimenez@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

The use of mobile devices and the development of geo-positioning technologies make applications that use location-based services very attractive and useful. These applications are composed of sensors that generate various and heterogeneous spatio-temporal data. Exploiting this spatio-temporal data to support video surveillance systems remains a relevant purpose for video content filtering. Since the data processed in such a context are heterogeneous (indoor and outdoor environment, various position types and reference systems, various data format), interoperability and management of these data remains a problem to be solved.

In this paper, we propose a new generic trajectory based query in order to handle trajectory segment's heterogeneity. for both environments (indoor and outdoor).The proposed data model integrate multi-source metadata and enable to handle interoperability issue of data. Our querying mechanism enable to automatically retrieve video segments that could contain relevant information for the CCTV operator (suspects, trajectories, etc.). To ensure industrial transferability, we have implemented the metadata dictionary of the Standard ISO 22311/IEC 79 (interoperability of CCTV systems).

We provide an experimental evaluation demonstrating the utility of our approach in a real-world case. Results show that the proposed approach enhances the efficiency of investigators by reducing the search space, as the operator will analyze only the relevant data, therefore he needs less time for video processing (video reviewing).

KEYWORDS

Spatio-temporal (meta)data, spatial query, trajectory, CCTV system, interoperability, indoor/outdoor location

1 INTRODUCTION

Nowadays, Location Based services (LBS) are widely used for many applications like surveillance, detection, navigation, etc. Such applications are based on object/device positioning and provide users with geo-located data via spatio-temporal queries. Applications use in most cases location models based on GPS sensors that are widely embedded (cars, smartphones, etc.) and other location sensors deployed (wifi, RFID, ultra wide bande, etc.). The metadata generated by these sensors are heterogeneous and vary according to the environment (indoor, outdoor): (i) **indoor environment**. Positions are generated by different types of sensors, expressed according to different references systems and can be either geometric (coordinates relative to a reference system as map of a building) or symbolic (more semantic description related to points of interest, parts of a building, etc.). For instance, location based on wireless gives geometric positions relative to 2 dimensional (2D) coordinate system while cellular network and RFID technologies give symbolic positions (e.g. an area) ; (ii) **outdoor environment**. The objects movement can be constrained for example by a road network (e.g., the cars movement follows the streets of a city) or a transport network (e.g., the buses movement follows predefined lines)[4]. So, positions are geometric or symbolic and can be expressed according to the following reference systems: geodesic system, road network, transport network.

Such heterogeneity of spatio-temporal (meta)data related to objects, devices or sensors cause a problem of interoperability

https://doi.org/10.1145/3282461.3282465

of location-based applications. It becomes difficult to tr ack objects/devices in different environments, or objects/devices located by different sensors, since metadata generated are heterogeneous.

In this paper, we propose the search and filtering of videos based on the modeling of spatio-temporal metadata from sensors and mobile devices/objects and metadata related to video content. For example: based on a trajectory constructed using locations of a person and a time interval, we can retrieve videos that may have filmed a scene of interest, then use video content features obtained via automatic image processing algorithms to filter or rank videos according to their relevance. In this context, interoperability problem can be tackled at different levels: (i) interoperability issues of spatio-temporal metadata from sensors and mobile objects/devices (described above); (ii) interoperability issues for CCTV systems : CCTV cameras record continuously and therefore generate a huge amount of heterogeneous data. Such a heterogeneity of data is due to the different contextual installation (indoor, outdoor) and camera specificities (manufacturers, data formats etc.); (iii) interoperability of all this multi-source metadata is one of the goals of this paper.

We propose a generic data model allowing efficient management and interoperability of spatio-temporal metadata from sensors and mobile devices/objects and CCTV metadata. Metadata modelling takes into account the metadata dictionary described in ISO 22311/IEC 79 (standard that aims to facilitate interoperability of CCTV systems). This data model is supported by a robust querying mechanism based on metadata for automatic retrieval of relevant video segments from a CCTV system during research of evidence in accident/crime event.

To sum up, the main contributions of this paper are summarized as follows:

- We define a new generic trajectory based query in order to handle trajectory segment's heterogeneity for both environments (indoor and outdoor);
- We define new elements to enrich the metadata dictionary of ISO 22311/IEC 79 standard.
- We provide a generic and scalable data model for multi-source metadata management and interoperability.
- We propose a method to automatically retrieve video segments that could contain relevant information for investigators
- We conduct experimental evaluations demonstrating the utility of our approach in a real-world case.

The rest of the paper is organized as follows. Section 2 reviews the related works; Section 3 formulates the problem related to heterogeneous metadata and CCTV system; Section 4, Section 5, Section 6 present our approach; Section 7 shows an real-case experimentation that we performed to evaluate the capability of our method to retrieve relevant video content. Finally, in section 8, we discuss and conclude with suggestions for future research.

2 RELATED WORKS

Some approaches have proposed the use of spatio-temporal information related to video content or objects in information systems. These approaches differ in:

- the main objective of the application (e.g.: annotation of videos and images with text tags, development of decision support systems based on querying geospatial information, development of traffic management systems);
- the metadata on which these systems are based: position of objects, geometry of the observed scene, time, technical characteristics of the camera;
- the type of processed positions (geometric, symbolic);
- how this information are represented in each data model: the data are continuous/discrete, the camera's field of view is represented as a moving region (a geometry) computed for each frame/minute or second, etc.;
- types of spatial query the system can respond to (e.g., position queries [1, 2], range queries [10, 16, 17], visibility queries [15], nearest neighbor queries [13], nearest surrounder queries [9], predictive queries [7]).

In [18], the authors present a spatio-temporal extension named STOC (a PL/SQL package) of Oracle Spatial. Moving regions are represented as geometries (SDO GEOMETRY) that move over time. The use case presented is a traffic information management system that answers questions such as: "which vehicles have crossed a given region?".

In [11], the authors propose a system (SEVA) that annotates each frame of a video by the location, timestamp and objects present in the frame. The system consists of (1) a video camera, (2) a digital compass, (3) a location system, (4) a wifi radio associated with the camera. They also construct the geometry of the field of view for each second of video.

In [14], a similar approach to SEVA is presented, with the following differences: (1) objects should not transmit their position and (2) their geometry is taken into account, not just the location point. For each second of the video, the authors computes the field of view associated with the camera and query two external databases (OpenStreetMaps and GeoDec) in order to extract the objects that are in the captured scene. The list of objects is refined by eliminating objects that are not visible (by computing a horizontal and vertical visibility).

In [6], an approach to annotating images based on camera location and orientation is presented. The originality lies in the fact that between the location and the optical characteristics of the camera (viewing angle), since the proposed system (TagPix) computes a distance between the user and different objects located in the visibility area of the camera in order to choose the most relevant tag. The main similarities with our approach lie in the computation of the field of view and distance seen by the camera without having access to the content. TagPix aims at annotating photos so does not consider the mobility of objects and cameras nor trajectory queries.

3 PROBLEM STATEMENT

An analysis of the existing approaches leads us to conclude that:

- due to the increasing volume of video content acquired by the large number of video sensors implemented in CCTV systems that deployed in the streets, in transportation and inside buildings, train or subway stations, etc., and in general in all aspects of everyday life (mobile devices, cars, etc.), there is a growing need to rely on elements describing the context (e.g., geolocation, orientation, installation, technical context) of video sensors to develop methods and tools for

filtering video content. Most existing approaches focus on developing video content analysis tools that can automatically detect activities, people, events with performance that depend on use case and variability of content quality, without considering pre-filtering of content, based on contextual elements

- most approaches that deal with objects movement in road or transport networks assume the existence of an upstream digitization step that generates the road graph (the nodes of the graph that are the intersections of roads and possibly intermediate points where the curvature of the road changes), which is a fairly strong hypothesis; digitizing, storing and managing an urban road network is a heavy task;
- the need for using standardized data models (for the description of road and transport networks) is not considered.
- approaches based on spatio-temporal data for different applications such as traffic management or filtering and searching video content do not offer data models that integrate information about all the elements we are interested in: road network, transport network, mobile objects and cameras.
- in most existing works the geometries of the camera fields of view are constructed (at the moment of entry into the system) for each frame and are stored as they are; in the case of a CCTV system this can quickly lead to a significant overload.

4 HYBRID TRAJECTORY BASED QUERY

The main queries in this field are expressed in the form of trajectories whose segments are described by geometric positions (points) or symbolic positions (descriptions that can be reduced to a point or a geometry) in relation to different outdoor or indoor reference systems such as the geodesic system, the road or transport network, the ICCARD reader network, etc. The aim of any " query " sent to a system that manages a video surveillance network is to find the video sequences (with sufficient image quality) that contain the objects (e.g., person, vehicle, abandoned luggage) or target events. Relevant information for an investigation may include: the location (or sequence of locations), date and time interval of the incident, and any information that allows identification of the suspect. The spatio-temporal information required for a query during an investigation are defined in [5] as outdoor hybrid trajectory based query (Fig. 1). For indoor context, [12] used an indoor hybrid trajectory based query (Fig. 2). In this paper we define a new generic trajectory based query in order to handle trajectory segments' heterogeneity for both environments (indoor and outdoor); an example is shown in Fig. 3. This trajectory based query is constructed by investigators based on facts, testimonies, etc. It is composed of two main parts: a spatial part and a temporal one. The spatial part can contain indoor and/or outdoor sub-parts, each of them consists of a sequence of segments, each segment consisting of a reference system identifier and a sequence of positions (geometric or symbolic) expressed relatively to the corresponding reference system. The temporal part is an interval of time $[t_1, t_2]$. This hybrid trajectory will constitute the entry point of our querying framework.


```
{
    "query": {
        "spatial": [
            {
                "WGS84": [
                    [43.56077,1.46298],
                    [43.56088,1.46292],
                    [43.56366,1.46156]
                ]
            },
            {
                "Street Montesquieu": [14,19]
            },
            {
                "Bus2": ["Université Paul Sabatier"]
            }
        ],
        "temporal": {
            "start": "07-02-2014 14:00:00",
            "end": "07-02-2014 14:30:00"
        }
    }
}
```


Figure 1: Example of outdoor trajectory based query.


```
{
    "query": {
        "spatial": [
            {
                "Floor8": [
                    [50.21627,433.53088],
                    [65.84485,409.75342],
                    [60.56707,441.10219],
                    [93.67555,371.18614]
                ]
            },
            {
                "ICCARD": [0000110015010005]
            }
        ],
        "temporal": {
            "start": "10-02-14 12:55:34",
            "end": "10-02-2014 13:20:00"
        }
    }
}
```


Figure 2: Example of indoor trajectory based query.

```
{
    "query": {
        "spatial": [
            "indoor":[
                {
                "Floor8": [ [85.12254,425.3325],
                            [92.4178,468.58037] ]
                },
                {
                "ICCARD": ["0000110015010005"]
                }
            ]
            "outdoor":[
                {
                "WGS84": [ [43.12254,1.0124],
                           [43.4178,1.02045] ]
                },
                {
                "RRTLSE": [ "Route de Narbonne" : 116,
                            "Route de Narbonne" : 118
                          ]
                }
            ]
        ],
        "temporal":
            {
            "start": "14-02-2018 05:00:00",
            "end": "14-02-2018 05:50:00"
            }
    }
}
```

Spatial part

Temporal part

Figure 3: Example of generic trajectory based query.

5 METADATA MODELING (MANAGEMENT AND INTEROPERABILITY OF DATA)

Metadata is defined as structured information that describes, locates and facilitates the retrieval, use and management of a resource. One contribution in this paper is the design of a generic data model that enables the management of heterogeneous (meta)data from sensors, mobile objects/devices and CCTV systems. We focused on metadata related on camera (motion, field of view, geolocation), spatio-temporal metadata from geolocation sensors, and metadata from content analysis algorithms.

5.1 Metadata related on camera

Camera field of view and geolocation are two key elements of the proposed approach. A camera located at a given position, with a specific orientation and installation can capture a given area. The field of view represents the area of the scene shot by the camera. Fig. 4 illustrates a camera field of view.

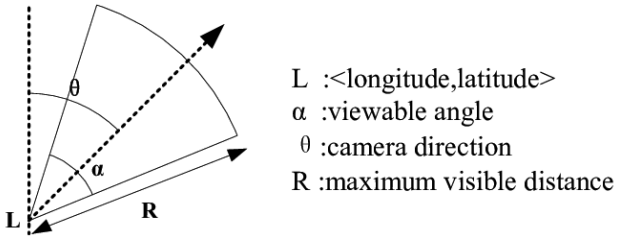

Figure 4: Camera field of view.

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. The captured scenes can change with possible camera rotations (Pan Tilt Zoom camera). Our data model handles all these requirements. Thus you can observe on Fig. 5 the relationships between the camera, the field of view and the location, specifying that a camera can have several positions with variable fields of view at different times. Other data such as the observed scene (e.g. building entrance) and the image quality of the camera are represented in the model. This data model can be consider as the implementation of sensors metadata described in the standard ISO 22311 (which describes the operational requirements for CCTV systems).

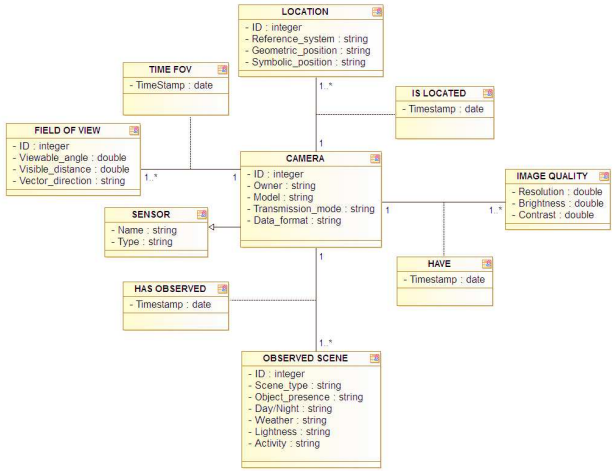

Figure 5: Metadata related to the camera.

5.2 Metadata from geolocation sensors

In the specific case of indoor environments, many sensors are installed in order to locate devices attached to objects (people, robots, etc.). The (meta)data generated by these sensors can be used to reconstruct object trajectories and can be combined with metadata related to cameras installed for identification purposes. Based on the geolocation sensor used (Wifi, ICCARD, Ultra-wide band, Radio Frequency Identification, etc.) [3, 8] the locations generated are geometric (e.g. 2D or 3D coordinates, latitude/longitude) or symbolic (more semantic description related to points of interest, addresses, etc.) with regards to different reference systems. We have implemented the data model shown in Fig. 6, which handles all these heterogeneous data. A Reader connected to a location sensor can record the different locations of the many devices over time.

5.3 Metadata from content analysis algorithms

Metadata from content analysis algorithms are considered in the proposed approach. We are focused on features such as the appearance of objects (people, trains, cars, etc.) in videos. The features are extracted by frame, using YOLO[1]: a real-time object detection algorithm; but any other algorithm may be used instead. A data model for these features is presented in Fig. 7. This model handles video elements such as: video segments, frames and events. The

[1] https://pjreddie.com/darknet/yolo/

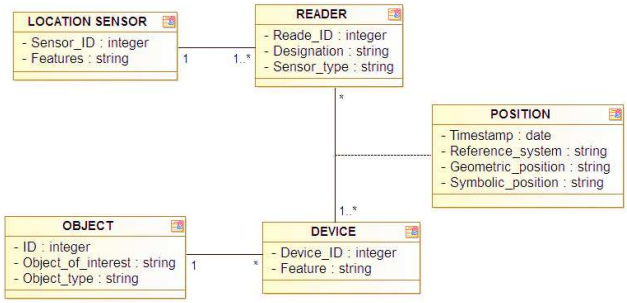

Figure 6: Metadata from geolocation sensors.

relationships between the classes "VIDEO","SEGMENT","FRAME" and "EVENT" are defined as follows: a video contains at least one segment, and the segment contains at least one frame; An event is an action involving content items at a particular place and over a particular time interval (e.g. suspicious hooded persons leaving a building and entering a car parked in an inappropriate location). A video can contain more than one event. The event time interval can be larger than the video segment, so the event is directly related to the video in the model. Object detection is done frame by frame. A frame can contain several objects. The model is able to handle the presence of objects in different frames, also it is able to indicate if there is a movement in the frame comparing with the previous one. Low-level characteristics such as entropy, brightness, contrast, etc. are taken into account in the model and will enable us to filter video content in future work ("negative" filtering).

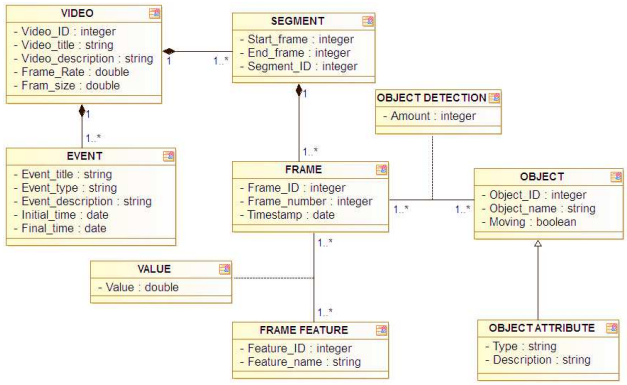

Figure 7: Metadata from content analysis algorithms.

All metadata described in the previous sections are aggregated in a generic data model shown in Fig. 8. It enables to integrate all the previously modeled metadata and then to handle the interoperability of heterogeneous data related to CCTV systems (one of the goals of this paper).This data model is scalable and enables the integration of other metadata.

6 QUERYING MECHANISM

This section develops our querying mechanisms for automatic retrieval of relevant video segments from a CCTV system during

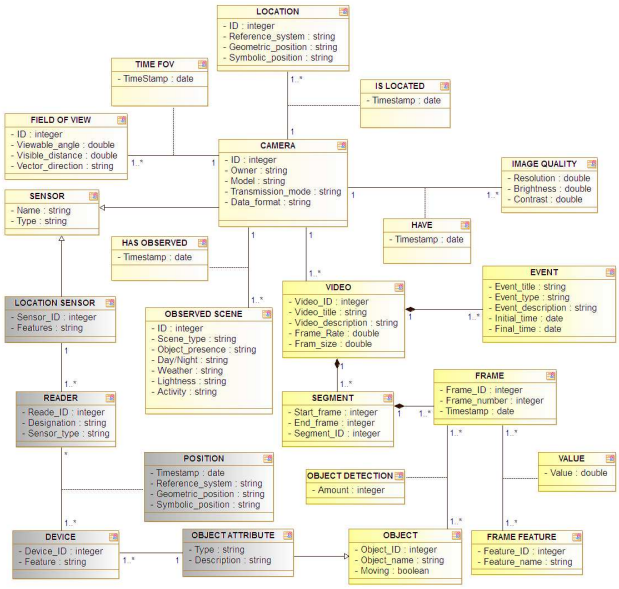

Figure 8: Generic data model.

research of evidence in accident or crime event. The main goal is to reduce the amount of data that need to be reviewed by the investigators and consequently decreasing the time of video(s) reviewing. The proposed method can be summarized in the following steps: (1) query trajectory intersection with camera fields of view in order to retrieve video segments that may be relevant for investigations; (2) content filtering of retrieved video segments according to the presence/absence of objects and movement.

6.1 Query trajectory intersection with camera fields of view

The main goal is to deliver to investigators the video sequences that may contain interesting images for their investigation (suspects, trajectories, etc.). To do this, it is necessary to search for cameras whose field of view (which can be variable) intersected the trajectories of the query in the given interval $[t_1, t_2]$. We used the *hasSeen* operator defined in [15] [16] as follows: given a spatial trajectory composed of segments $t_r = (u_1, \ldots, u_n)$ and the time interval $[t_1, t_2]$, $hasSeen(t_r, t_1, t_2)$ returns the set of cameras $c_i (1 \leq i \leq m)$ that captured at least one segment $u_k (1 \leq k \leq n)$ and a video sequence between two moments t^i_{start} and t^i_{end} within the interval $[t_1, t_2]$ $(t_1 \leq t^i_{start} \leq t^i_{end} \leq t_2)$.

$$hasSeen : u_1, \ldots, u_n, [t_1, t_2] = \begin{cases} c_1 : t^1_{start} \rightarrow t^1_{end}, \ u_k (1 \leq k \leq n) \\ c_2 : t^2_{start} \rightarrow t^2_{end}, \ u_k (1 \leq k \leq n) \\ \ldots \\ c_m : t^k_{start} \rightarrow t^m_{end}, u_k (1 \leq k \leq n) \end{cases}$$

Two algorithms have been developed for the camera selection process: one for fixed camera and the other for mobile camera. Both algorithms proceeds in two steps: the candidate selection step (purely spatial filtering) and the results refining step (temporal filtering):

- The filtering step applies an algorithm corresponding to a “Region Query" type similar to the one presented in [14]. It allows to select for each segment of trajectory, the cameras located at a distance less than or equal to the maximum visibility distance of all existing cameras in the database. This avoids the evaluation of the spatial intersection (expensive operation) for fields of view of the cameras which are located at a distance that makes it impossible to shot the query segments.
- For the previously selected cameras, the refining step calculates the geometries of the field of view during the time interval of the query, and selects those whose geometries “intersect" the trajectories segments and calculates the time interval $[t_a\ ,\ t_b]$ for each of them.

The result is a set of triplets: $R = \{r = (c_i, u_k, [t_a, t_b])\}$, $c_i \in SetOfCamera$, $u_k \in tr$, $t_1 \leq t_a$, $t_b \leq t_2$.

6.2 Content filtering of retrieved video segments

Once the relevant videos segments are retrieved, some of them may contain no object (people, vehicle, etc.) nor movement. Removing these useless sequences will reduce processing time for investigators.Thus, based on the video content features extracted and modeled in section **??**, we have implemented a content metadata-based query to retrieve video sequences that contain moving objects.

We have defined a $videoOfInterest(VoI)$ operator as follows: given a set of cameras $c_i (1 \leq i \leq m)$ and each related interval of time $[t_a, t_b]$ returned by the $hasSeen$ operator ($R = \{r = (c_i, u_k, [t_a, t_b])\}$), $VoI(\{c_i, [t_a^i, t_b^i]\})$ returns each c_i with a set of interval of time $[t_{start}^{i,k}, t_{end}^{i,k}]$, $t_a \leq t_{start}^{i,k} \leq t_{end}^{i,k} \leq t_b$. Each interval of time $[t_{start}^{i,k}, t_{end}^{i,k}]$ represents a video sequence in which there are objects and movement.

$$VoI : \{c_i, [t_a^i, t_b^i]\} = \begin{cases} c_1 : t_{start}^{1,1} \rightarrow t_{end}^{1,1}, \ldots, t_{start}^{1,n} \rightarrow t_{end}^{1,n} \\ c_2 : t_{start}^{2,1} \rightarrow t_{end}^{2,1}, \ldots, t_{start}^{2,n} \rightarrow t_{end}^{2,n} \\ \ldots \\ c_m : t_{start}^{m,1} \rightarrow t_{end}^{m,1}, \ldots, t_{start}^{m,n} \rightarrow t_{end}^{m,n} \end{cases}$$

The result is the following set: $V = \{v = (c_i, \{[t_s^{i,k}, t_e^{i,k}]\})\}$, $c_i \in R$, $t_a \leq t_s^{i,k}$, $t_e^{i,k} \leq t_b$.

7 EXPERIMENT STUDY

7.1 Prototype architecture

Fig. 9 illustrates the prototype architecture that we have developed and that implements the data model and the operators described in the previous sections.

The main modules of our prototype are:

- Query interpreter: interprets (transforms) the generic trajectory query submitted by the user into a spatio-temporal query;
- Search Engine: implements the research operators ($hasSeen$, $videoOfInterest$) defined in the previous sections;
- Database-MongoDB: contains metadata collections based on the generic proposed model.

The following external modules are used:

- User interface: can be used to build the generic trajectory query. It also enables data and results visualization;
- Video Metadata Extraction: is used to extract data related to the video (object detection, movement, etc.);
- Data Collecting: is used to collect spatio-temporal data and sensor-related data (e.g. data related to the camera field of view).

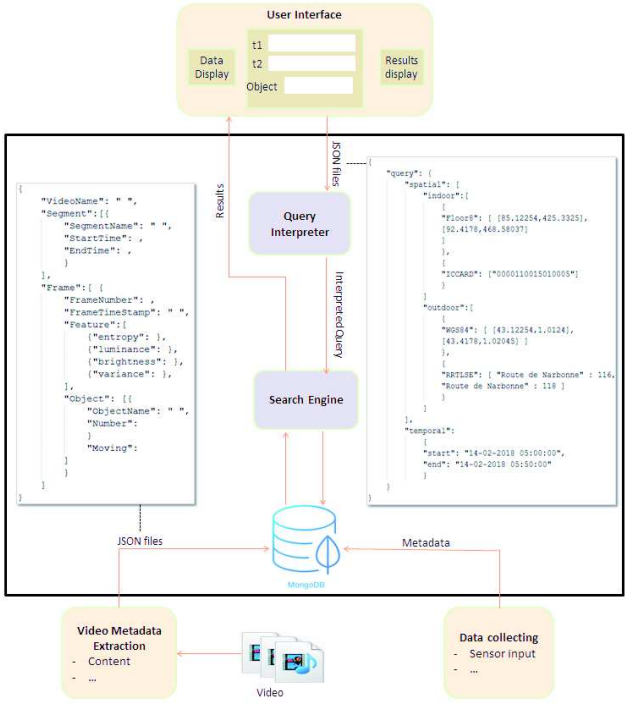

Figure 9: Architecture of the implementation of our method for relevant video segments retrieval.

7.2 Dataset and experiments

The datasets come from geolocation sensors installed on the Kyushu University campus in Japan. These datasets consist of 646109 objects detected on the eighth floor from 08/02/14 at 17:29:04 to 19/02/14 at 16:48:28. The scenario consists of 17 CCTV cameras and its description is summarized as an important document theft that occurred on the eighth floor in room 804 on 16/02/14 between 14:05:47 and 17:29:02. The goal is to find elements to identify the suspects.

With our approach, we defined a query whose incident region is the area of the 804 office and the query time interval is [16/02/14 at 14:05:47, 16/02/14 at 17:29:02]. In this scenario, 13 suspected devices have been detected. 6 cameras have been returned with time intervals for each camera, which provided enough information for the investigation. The total number of hours to view for all selected cameras was 00:57:47 (approximately 58 minutes). Manual analysis (without prototype) should consist in watching approximately 3 hours of videos per camera (for 17 cameras), giving a total of 17*3 = 51 hours of video to analyze. Thus, in this scenario, results shows

Table 1: Video retrieval comparisons.

Selected devices (suspects)	Selected cameras		Time (minutes)	
	Ground truth	Proposed approach	Ground truth	Proposed approach
1	2	2	9	10
2	5	5	15	15
3	4	4	18	18
4	3	4	15	17
5	7	7	28	28
6	4	4	19	20

that the developed prototype drastically reduced research space and time for CCTV operator.

However, since there is no easy way to get the "ground truth" for the query result set, it is difficult to evaluate the accuracy of the matching video segments for the given query. One possible way is the using of object recognition algorithms to extract all video frames in which a given object is visible. However, such object recognition algorithms have their own limitations that would also need to be integrated in the evaluation process. Thus, to evaluate our algorithm, the safest way is to watch all videos and manually set the time intervals in which the desired objects have appeared in each video.

Thereby, we performed experiments in which we defined several queries considering different number of suspects (from 1 to 6). We intersected the trajectories of these devices by the fields of view of the cameras installed in the building. For each query, we performed manual checks of cameras selection and time interval selection, and compared them with our automated approach. Table 1 summarized the results obtained. **"Selected cameras"** represents all cameras that filmed the suspects during their trajectories; **"Time"** is the total of time intervals for which the selected cameras intersected the suspects trajectories.

Figure 10 shows the number of selected cameras for each query by: ground truth (*Ground_Truth*) and the approach we have proposed (*Proposed_Approach*). Ground truth is video time really to be watched because they always contain suspicious persons/objects

Results show that, for the most part (5/6), the proposed approach returns the same number of cameras as ground truth. The cameras selected by our approach are the correct ones. The only case where "*Proposed_Approach*" returns one more camera can be explained by inaccuracy in the generated positions of devices or in camera positions.

Figure 11 shows the total time in minute for selected cameras. Results show that, in addition to the time given by *Ground_Truth*, *Proposed_Approach* adds one or two more minutes. This is due to the fact that different cameras can film the same segment at a given time, so this redundancy of information increases the total time.

We computed Precision and Recall to evaluate the degree of accuracy and comprehensiveness of our resulting set. We denoted P_C and R_C, the precision and recall measures related to the cameras retrieved by our method. Besides, we denoted P_T and R_T, the precision and recall measures related to the total time of video segments retrieved by our method. In our context, $P_C < 1$ means that the resulting set contains non-relevant cameras, and $R_C < 1$ implies

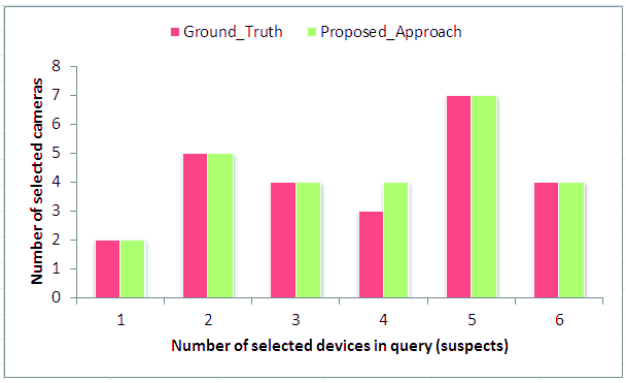

Figure 10: Number of selected cameras per number of suspects.

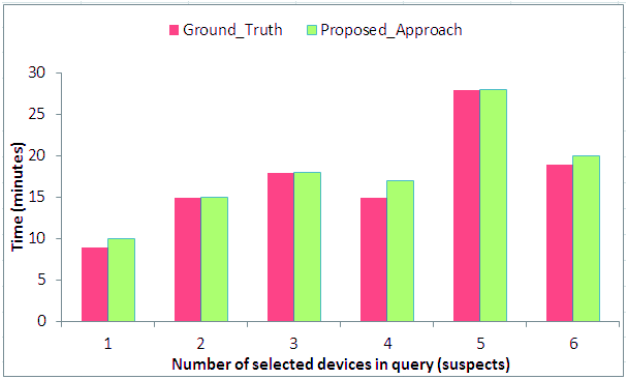

Figure 11: Total time per query.

that some relevant cameras have been ignored. Similarly, $P_T < 1$ means the resulting set contains non-relevant time interval, and $R_T < 1$ implies that some relevant cameras have been ignored.

We computed P_C, R_C, P_T and R_T as follows:

$$P_C = \frac{|PA_C(i) \cap GT_C(i)|}{|PA_C(i)|} = 0.96,$$

$$R_C = \frac{|PA_C(i) \cap GT_C(i)|}{|MC_C(i)|} = 1,$$

$$P_T = \frac{|PA_T(i) \cap GT_T(i)|}{|PA_T(i)|} = 0.96,$$

$$R_T = \frac{|PA_T(i) \cap GT_T(i)|}{|MC_T(i)|} = 1,$$

with $PA_C(i)$ and $GT_C(i)$ being respectively the sets of cameras retrieved by *Proposed_Approach* and by *Ground_Truth*, regarding the number of suspects i. Similarly, where $PA_T(i)$ and $GT_T(i)$ are respectively the total time of video segment retrieved by *Proposed_Approach* and by *Ground_Truth*, regarding the number of suspects i.

We can observed that our method *Proposed_Approach* almost performed perfect match with the *Ground_Truth* except for precision. This shows that our algorithm retrieved the totality of the relevant content without adding too much non-relevant content.

We are aware that the completeness of the results from a single dataset may not imply the same for the general case. Further experiments on a large amount of dataset are needed to evaluate the global performances of the proposed method.

8 CONCLUSION

In this paper, we presented a modelling approach of spatio-temporal metadata related to sensors, objects, devices and CCTV systems. In our context, data heterogeneity (differents types of position, various references systems, indoor en outdoor environment, multi-source data) faces us to a problem of management and interoperability. We defined a generic trajectory based query in order to handle trajectory segment's heterogeneity for both environments (indoor and outdoor); We defined new elements to enrich the metadata dictionary of ISO 22311/IEC 79 standard. We provided a generic and scalable data model for multi-source metadata management and interoperability. Since our work is applied on video analysis for investigative purposes, we proposed a method to automatically retrieve video segments that could contain relevant information for investigators.Conducted experiments show that the proposed approach enhances the efficiency of investigators by reducing the search space, as the operator will analyze only the relevant data, therefore he needs less time for video processing (video reviewing).

In our work, we relied on spatio-temporal metadata to identify videos that are not related to the spatio-temporal trajectory of interest for the investigation. Other "negative" filtering measures can be developed based on metadata or video characteristics in order to improve the information retrieval capability of our approach. We are currently working in collaboration with industry (Thales Communications & Security SA) and technical scientific police (PTS) to ensure that our approach is applicable and effective in the real-world case. The next step of our work is to extend this approach, in FILTER2 French ANR project and VICTORIA H2020 European project.

REFERENCES

[1] Imad Afyouni, Cyril Ray, and Claramunt Christophe. 2012. Spatial models for context-aware indoor navigation systems: A survey. *Journal of Spatial Information Science* 1, 4 (May 2012), 85–123. https://doi.org/10.5311/JOSIS.2012.4.73

[2] Haidar AL-Khalidi, David Taniar, and Maytham Safar. 2013. Approximate algorithms for static and continuous range queries in mobile navigation. *Computing* 95, 10 (01 Oct 2013), 949–976. https://doi.org/10.1007/s00607-012-0219-7

[3] Abdulrahman Alarifi, AbdulMalik S. Al-Salman, Mansour Alsaleh, Ahmad Alnafessah, Suheer Alhadhrami, Mai A. Al-Ammar, and Hend Suliman Al-Khalifa. 2016. Ultra Wideband Indoor Positioning Technologies: Analysis and Recent Advances †. In *Sensors*.

[4] Khaled Amriki and Pradeep K. Atrey. 2012. Bus surveillance: how many and where cameras should be placed. *Multimedia Tools and Applications* 71 (2012), 1051–1085.

[5] Dana Codreanu, André Péninou, and Florence Sèdes. 2015. Video Spatio-Temporal Filtering Based on Cameras and Target Objects Trajectories – Videosurveillance Forensic Framework. *2015 10th International Conference on Availability, Reliability and Security* (2015), 611–617.

[6] Hillol Debnath and Cristian Borcea. 2013. TagPix: Automatic Real-Time Landscape Photo Tagging for Smartphones. *2013 International Conference on MOBILe Wireless MiddleWARE, Operating Systems, and Applications* (2013), 176–184.

[7] Abdeltawab M. Hendawi and Mohamed F. Mokbel. 2012. Predictive spatio-temporal queries: a comprehensive survey and future directions. In *MobiGIS*.

[8] Sébastien Laborie, Ana-Maria Manzat, and Florence Sèdes. 2009. Managing and Querying Distributed Multimedia Metadata. *IEEE MultiMedia* 16 (2009), 12–21.

[9] K. C. K. Lee, W. C. Lee, and H. V. Leong. 2010. Nearest Surrounder Queries. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (Oct 2010), 1444–1458. https://doi.org/10.1109/TKDE.2009.172

[10] Jongtae Lim, Kyoungsoo Bok, and Jaesoo Yoo. 2016. Processing a Continuous Range Query in Mobile P2P Network Environments. In *Proceedings of the Sixth International Conference on Emerging Databases: Technologies, Applications, and Theory (EDB '16)*. ACM, New York, NY, USA, 102–105. https://doi.org/10.1145/3007818.3007834

[11] Xiaotao Liu, Mark D. Corner, and Prashant J. Shenoy. 2005. SEVA: sensor-enhanced video annotation. *TOMCCAP* 5 (2005), 24:1–24:26.

[12] Dimitrios Lymberopoulos, Jie Liu, Xue Yang, Romit Roy Choudhury, Vlado Handziski, and Souvik Sen. 2015. A realistic evaluation and comparison of indoor location technologies: experiences and lessons learned. In *IPSN*.

[13] Thomas Pajor. 2009. *Multi-modal route planning*. Master's thesis. Karlsruhe Institute of Technology, Germany.

[14] Zhijie Shen, Sakire Arslan Ay, Seon Ho Kim, and Roger Zimmermann. 2011. Automatic tag generation and ranking for sensor-rich outdoor videos. In *ACM Multimedia*.

[15] Rainer Simon and Peter Fröhlich. 2007. A mobile application framework for the geospatial web. In *WWW*.

[16] Guoqing Xiao, Kenli Li, Keqin Li, and Xu Zhou. 2015. Efficient top-(k, l) range query processing for uncertain data based on multicore architectures. *Distributed and Parallel Databases* 33, 3 (2015), 381–413.

[17] Wenjie Yuan and Markus Schneider. 2010. Supporting Continuous Range Queries in Indoor Space. *2010 Eleventh International Conference on Mobile Data Management* (2010), 209–214.

[18] Lei Zhao, Peiquan Jin, Lanlan Zhang, Huaishuai Wang, and Sheng Lin. 2011. Developing an Oracle-Based Spatio-Temporal Information Management System. In *DASFAA Workshops*.