Dialogue-driven Search in Surveillance Videos

Vincenzo Deufemia, Massimiliano Giordano, Giuseppe Polese, Genoveffa Tortora
Facoltà di Scienze MM.FF.NN., Università di Salerno
Via Ponte don Melillo, 84084 Fisciano(SA), Italy
{deufemia, mgiordano, gpolese, tortora}@unisa.it

Abstract
Video surveillance is a key tool for enabling security personnel to monitor complex and dangerous environments, and the access to the semantic content of surveillance video has become a challenging research area. In this paper we present a query-processing system for investigating human activities in surveillance videos. The interaction between the user and the system happens through a gesture-based dialogue, which implements a novel question-answering model to support users during the information-seeking process. We also present a visual interface devised to facilitate the query-specification process.

1. Introduction
With the increasing need of security in today’s society surveillance systems have become of fundamental importance. Video cameras and monitors pervade buildings, factories, streets, and offices. Thus, video surveillance is a key tool for enabling security personnel to safely monitor complex and dangerous environments. However, even in simple environments, a video surveillance operator may face an enormous information overload. It is nearly impossible to monitor individual objects scattered across multiple views of the environment. It thus becomes vital to develop interfaces making the investigation process on the overwhelming quantity of videos more intuitive and effective.

In the recent years intelligent user interfaces (IUIs) have been investigated for multimedia applications, aiming to improve efficiency, effectiveness, and naturalness of human-machine interaction by representing, reasoning, and acting on models of the user, domain, task, discourse, and media [12]. IUIs have to make the dialogue between the user and the system possible. Real interaction occurs when there is a need to ask for information during a computation. This need actually arises during the computation and cannot be shifted to the starting point of the computation process. This kind of interaction affects the computation and only the interfaces able to realize it can be considered intelligent and able to manage the interaction between system and user.

A problem arising in the application of this view is the need of a powerful language, like it happens among people using natural language for dialoguing. Natural language interfaces are difficult to realize as they yield difficult problems related to natural language processing. The Question-Answering (Q/A) paradigm [17] is a suitable mean to interact with video surveillance systems. Indeed, Q/A implements the investigative dialogue and supports the guided investigation by foreseeing the user actions. On the other hand, it is essential for interfaces to have human-like perception and interaction capabilities that can be utilized for effective human–computer interaction (HCI).

In this paper we present a query-processing system for investigating human activities in surveillance videos. The interface exploits the information computed by the recognition system to support users (security operators) in the investigation process. The interaction dialogue provides a query language based on gesture that enables users to easily specify various kinds of questions about both actions and states, and the nature of responses one wishes. In particular, the user makes use of hand-drawn symbols for describing the situations s/he needs to investigate, whereas the system uses visual symbols to represent the questions to the user. The contribution of this research is twofold: (i) proposing an intuitive interaction mechanism for surveillance video investigation, and (ii) proposing a question-answering model to support users during the information-seeking process.

The paper is organized as follows. Section 2 surveys related work. Section 3 presents the components of the video understanding system. Section 4 introduces the main notions about the question-answering system, while the proposed user interface for investigating human activities is described in Section 5. Finally, Section 6 provides conclusions and outlines future research directions.

2. Related Work
In the domain of video surveillance much attention has been devoted to the problem of using visualization techniques for clustering and anomaly detection [3][15]. Little work has been devoted to the development of
interfaces and interaction paradigms to support users in the investigation process.

The set of interface components presented in [7] aims to improve the ability of security personnel to locate and follow important activities within security videos. The components include the recognition and visualization of localized activities in a video feed and provide activity-highlighting video summaries in the form of enhanced keyframes, timelines, and storyboards to give users quick access to interesting events in recorded videos.

In the recent years many video retrieval frameworks for visual surveillance have been proposed [9]. They support various query mechanisms, because queries by keywords have a limited expressive power. In particular, query-by-sketch mechanisms have been adopted to express queries such as "a vehicle moved in this way". A drawing interface allows users to draw motion trajectories which approximately represent the activities that users expect to query.

An approach similar to the one presented in this paper has been developed by Katz et al. [10]. They integrate video and speech analysis to support question-answering about moving objects appearing within surveillance videos. Their prototype system, called Spot, analyzes objects and trajectories from surveillance footage and is able to interpret natural language queries such as "Show me all cars leaving the garage". Spot replies to such a query with a video clip showing only cars exiting the garage.

In the recent years several video retrieval systems have been developed to assist the user in searching and finding video scenes. In particular, interactive video retrieval systems are becoming popular. They try to reduce the effect of the semantic gap, i.e., the difference between the low-level data representation of videos and the higher-level concepts a user associates with videos. An important strategy to improve retrieval results is the query reformulation, whereas strategies to identify relevant results are based on relevance feedback and interaction with the system. The system proposed in [1] combines relevance feedbacks and storyboard interfaces for shot-based video retrieval.

The interaction dialogue proposed in our approach is a generalization of relevance feedback. Indeed, the relevance questions are asked to catch the user's information need, whereas the question-answering process we propose implement a real dialogue between the user and the system making the investigation process more effective.

3. A Video Understanding System Based on Conceptual Dependency

Video understanding aims to automatically recognize activities occurring in a complex environment observed through video cameras [6]. The goals of the proposed activity recognition system are to detect predefined violations observed in the input video streams and to answer specific queries about events that have already occurred in the archived video [4].

We exploit Artificial Intelligence techniques to enable the system “understand” events captured by cameras. In particular, our approach is based on the Schank’s theory [14], a “non-logical” approach that has been widely used in natural language processing. Two main reasons led us to use this theory in video surveillance systems. First, the presence of well-studied primitives to represent details about the actions; second, the possibility to use highly structured representations like scripts, which are a natural way to manage prototypical knowledge. Thus, we are able to associate different levels of meanings to a situation: conceptualization, scene, and script level, which allow us to deeply understand the current situations and to detect anomalies at different levels. Moreover, the structured information makes the design and reasoning process easier.

In order to detect anomalies and to raise alert messages, the system tries to interpret a scene based on its knowledge about “normal” situations, using the conceptual dependencies to describe single events and scripts for complex situations. Therefore, the proposed video-surveillance system is an intelligent system associating semantical representations to images.

Fig. 1 gives an overview of our video understanding system which is composed of three main modules: detection and tracking of multiple objects, scene understanding, and reasoning. The module for tracking multiple objects is implemented by use the codebook based adaptive background subtraction algorithm proposed in
We are concerned with tracking three kinds of objects - human, vehicles and packages.

The reasoning module has twofold functions: understanding the situations that happen and managing the dialogue with the interface. The first task is accomplished by the scene understanding module, whose aim is of associating a semantic representation to the content of the scenes. This module recognizes events and actions using the knowledge about the standard events and situations stored in its knowledge base. In particular, the information on the tracked objects, i.e., trajectories and features (such as color, size, etc.), are synthesized by constructing conceptualizations, which are given in output to the next module. As an example, the following conceptualization expresses the fact that a given car moves from the garage entry to a car place.

\[
\text{car} \leftrightarrow \text{TRANS} \rightarrow \text{car} \leftrightarrow \text{Place} \rightarrow \text{Entry}
\]

The scene understanding module also activates pertinent scripts and appropriate scenes from the script produced by the tracking module in order to identify possible anomalies. In particular, when a script is activated, the conceptualizations belonging to the scenes that might occur are sent to the tracking module to work in a predictive mode. To correctly understand the scene structure, we label various areas of the background, such as doors, elevators, ATM, and so on. The conceptualizations are generated based on object properties and their interactions with these labeled background regions. The output of the understanding module is the scripts describing the occurred situations.

The reasoning tasks of the scene understanding module are:
1. to understand events: The task of representing current events using the stored knowledge is accomplished both reducing events to simple ones and instantiating the objects in the conceptualization with actual data;
2. to reason about events: Once an event has been interpreted using the existing knowledge it is possible to make inferences and to supply the lack of information in occurred events.

The object detection & tracking module and the scene understanding module exchange information themselves since the first one passes information to be conceptualized to the second one, and, in turn, the latter passes expectations to the first one. Expectations are events or actions which typically follow the last recognized event, making easier the low level recognizing of events.

The task of managing the dialogue with the interface is realized by the dialogue reasoning module, whose main task is to answer questions about occurred events. In order to do so, the module tries to start a real dialogue, constituted by a follow up of questions and answers, with the interface. The function of this module will be treated in deepen in the next section.

The sketch-based user interface allows users to interact with the videos through a language which is natural and intuitive for the user and complete to support a dialogue between the user and the system.

4. Dialogues for Investigation

The classical investigation process in multimedia retrieval is accomplished through the interactive search process shown in Fig. 2. The user repetitively submits a query to the system based on the topic under investigation, the previous queries, and the results obtained so far. The introduction of relevance feedback has allowed to improvement this process. Relevance feedback is the method of reformulation and improvement of the original search request, based on information from the user about the relevance of the retrieved data [13]. However, this method suffers from several limitations such as the constraint to browse the results to give feedback to the system.

![Figure 2. Interactive search: human (re)formulates query, based on topic, query, and/or results.](image)

If we think of both the user and the system as interogative reasoners, the interface can be interpreted as an oracle for both user and system. In fact, the system does not know who the user is, but it is sure that s/he tells the truth; analogously the user trusts the system, thinking that it tells the truth. Therefore, the interface represents the system for the user and it represents, in turn, the user for the system.

4.1 Metaquestioning

Question-answering systems (henceforth Q/A) have the goal of finding and presenting answers to questions that the user makes. In these systems, the interface has the role of managing a common language between user and system in order to enable the former to make questions and the latter to provide answers. However, Q/A systems do not realize a real dialogue, because the user can only ask questions and the system can only answer. As observed by Driver in [5], it is possible and often desirable that a question be followed by another one, as in the following examples:

- \(q_1\) What happened yesterday?
- \(q_2\) Would you like a short or a long response?

Question \(q_2\) in the previous examples are called metaquestions. They occur between an inquirer (questioner) asking a first order question and a responder (answerer/metaquestioner) answering through a metaquestion.
The importance of metaquestions in the context of Q/A systems is due to the fact that they can be used to overcome obstacles to answering the first order questions and, hence, they have an active role in the Q/A process itself. In a sense, metaquestions can be seen as a generalization of feedbacks.

4.2 MetaReasoning

Metaquestioning involves many features deriving from the fact that it is related to different research areas, like dialogue theory, problem solving, and metareasoning. From the point of view of dialogue theory [16], metaquestioning can be seen as the general process underlying the information-seeking type of dialogue, whose goal is to exchange information and where the goals of user and system are to acquire information and to give information, respectively. According to this view, it turns out that metaquestioning is a version of the analytic method involving two reasoners: the user and the system. This observation tells us that metaquestioning is a heuristic process and that, as a consequence, it is characterized by certain but incomplete rules, like, for instance, the problem specialization or generalization.

In our model, we suppose that there are the user $U$ (questioner) and the answerer/metaquestioner $A/M$ such that $U$ asks a question $q_1$ to $A/M$ and $A/M$ responds with $q_2$. Questions are represented according to the logic programming style. As an example, the question “Who entered from the main entrance between 10 and 11 p.m.?” is represented by the clause $\text{act}(X, \text{enter}, \text{MainEntry}, [10-11\ p.m.])$, whereas the question “What are the paths of John between 10 and 11 p.m.?” is represented by the clause $\text{act}(\text{John}, \text{path}, X, [10-11\ p.m.])$. Thus, a question has an unknown (indicated with an $X$) and a body (represented by the action $\text{enter}$ in the previous examples).

5. Sketch-based Dialogues for Human Activity Investigation

A problem arising in the development of dialogue interfaces is the need of a powerful language, like the natural languages people use for dialoguing. Natural language interfaces are very difficult to realize, since Natural Language Processing has already proven to raise many problems.

We propose the use of hand-drawn gestures as a dialogue language for investigating human activities in surveillance videos. The language is not as versatile as natural languages, but it allows users to query the system in a natural way.

5.1 The Sketch Language for Activity Investigation

A sketch language is formed by a set of sketch sentences over a set of shapes from a domain-specific alphabet. To support the question-answering process the sketch language should allow users specify:

1. the kind of object to be retrieved (the unknown) and the constraints on it, (e.g., a person in an angle of the room);
2. actions and states involving the scene objects, (e.g., a person opening a door, a person waiting the lift);
3. temporal information on the events to be investigated, (e.g., the time interval of a theft);
4. elements of the metaknowledge, (e.g., some properties of the response).

In the following, we describe the symbols composing the dialogues between users and the system.

Language Symbols. The user can associate a sketched symbol to each kind of object identified by the Object Detection algorithm, and will use it to refer to the objects in the context of questions. As an example, if the detection algorithm is able to categorize the mobile objects in people and packages, then during the specification of the questions the user can refer to them by using the sketched symbols in Fig. 3.

![Figure 3. Sketched symbols of the objects detected in the video scene.](image)

In case the objects involved in the question are part of the scene, the user can select them by hooping them with a hand-drawn circle. As an example, Fig. 4 shows a question where a person is already in the scene and it is included in the question by the red ellipse stroke.

![Figure 4. A question involving a detected object and a path as the unknown.](image)

Actions and States. Useful information during the investigation process is in the actions involving the detected objects, and the states in which they could be. As for relationships, they depend on the actions that the algorithm used to generate the facts that is able to infer. As an example, the action “a person picks up a package” is described by the sketch in Fig. 5(a), while the sketch in Fig. 5(b) describes the action of leaving a package.
Temporal Information. Questions information regarding time intervals is specified by drawing a circle on the timeline at the bottom of window.

Metaquestions. Sketches are also used by the system to represent questions for the user. We have defined the sketch symbols for a set of metaquestions, such as “how long should the response be?”, “which is the path followed by the person?”. The example in the next section shows some metaquestion symbols.

Unknown. The unknown of the question is indicated with a question mark on top of the sketched symbol. As an example, the unknown of the question in Fig. 4 is the walk path of the person enclosed with a circle.

5.2 The Sketch-based Interface

We have built a prototype video surveillance system with a sketch-based interface answering interesting questions about video surveillance footage taken in university offices, corridors, and halls. The scenes contain both persons and packages. A typical segment of the video footage shows persons leaving and entering offices, persons discussing in the hall, and persons putting down and picking up packages in the offices and corridors.

Fig. 6 shows the system interface. The main window contains the (background) image of the selected camera, on which the user can draw the sketch representing a question, and the system can reply with another question. As said above, timeline at the bottom of the image is used to specify the temporal information of the question as we show in the following example. The frame at the bottom of the interface (Fig. 6(b)) contains the images obtained from the user question. A storyboard containing the previous investigations is on the right of the interface (see Fig. 6(c)).

The sketches drawn by the user are recognized by using the approach proposed in [2]. The recognition algorithm employs visual language parsing techniques and exploits contextual information to solve ambiguities in the sketches. On average the algorithm gets a recognition rate of 96% since the symbols of the language are not complex and the number of overlapping symbols is very low.

The interface also includes a Question Visualizer module, which receives the metaquestions from the Dialogue Reasoning module, and adapts their layout to the user input. In particular, it adapts the shape of the metaquestion symbols to the sketch question drawn by the user.

6. Conclusions and Future Work

We have presented a novel interface for investigating human activities in surveillance videos. The interaction with the surveillance system consists of sketch dialogues allowing users to easily specify various kinds of questions about occurred events. The presented proposal contains two innovative solutions: the use of sketching for representing the dialogue between the user and the system, and the use of question-answering to support users in the investigation process.

We have also presented a system prototype that implements several dialogues in a scenario of multi-camera university activity surveillance. In the future we plan to extend the sketch language both for questions and metaquestions. This will allow the system to support dialogues that are more complete and concrete for the user.

We also plan to perform thorough usability experiments. Initial experimental uses of the system have
already shown a potential enhancement in the user support during the investigation process. However, in some cases the system fails to understand the sketch question due to presence of ambiguities in the positioning of the sketched objects. In particular, this happens because in the same position of the image there are different places where the objects can appear. As an example, when a sketch symbol representing a person is drawn on a space that cover more than one street.

Finally, we intend to extend the interface for supporting the specification of questions involving more than one camera.

7. References