High-Level Video Representation for Advanced Video Applications

Aishy Amer[a] Eric Dubois[b]

[a]INRS-Télécommunications; Montréal, Qc; H5A 1K6 Canada
[b]University of Ottawa; Ottawa, On; K1N 6N5 Canada

ABSTRACT

Video is increasingly used in various advanced applications. Many of these applications require common video representations that should be oriented towards how people describe video content. In this paper we first discuss the background of high-level video representations. We then introduce a computational framework for high-level video representation that evolves towards how people describe video content. Our framework represents a video shot in terms of its moving objects and their related semantic features such as events and other high-level motion features. To achieve higher applicability, content should be extracted independently of the type and the context of the input video. Our representation system, implemented on 6371 images with multi-object occlusion and artifacts, produces stable results in real-time. This is due to the adaptation to noise, the compensation of estimation errors at the various processing levels, and the division of the processing system into simple but effective tasks.

1. BACKGROUND

Video is becoming integrated in various personal and professional applications such as entertainment, education, telemedicine, databases, security applications and even low-bandwidth wireless applications. Providing means for fast, automated, and effective techniques to represent video based on its high-level content, such as objects and meanings, are important topics of research.[1–3] In a surveillance application, for instance, object extraction is necessary to classify object behavior, and with video databases, effective retrieval must be based on high-level features. Automated content representation would significantly facilitate the use and reduce the costs of video retrieval and surveillance by humans.

Typically, a video is a set of stories, scenes, and shots. There is little semantic change in the visual content of a shot, i.e., within a shot there is a short-term temporal consistency. To facilitate extraction and analysis of video contents, a video has to be first segmented into shots. A shot consists, in general, of multiple *objects*, their *semantic interpretation* (i.e., objects' meaning), their *dynamics* (i.e., objects' movement, activities, action, or related events), and their *syntax* (i.e., the way objects are spatially and temporally related, e.g., 'a person is close to a vehicle'). (In this paper the term video refers to a video shot.)

Developing advanced content-based video representation requires the resolution of two key issues: defining what are interesting video contents and what features are suitable to represent these contents. In various video applications, common video representations are needed. Since the focus when developing a video representation is on the needs of a human operator a representation should evolve towards how people describe video content. The main questions are: what level of object features and semantic content is most *important* and most *common* for advanced video applications? What type of high-level features are appropriate for various applications? Are high-level intentional descriptions such as what a person is thinking needed? Is the context of a video data necessary to extract useful content? An important observation is that the subject of the majority of video is related to moving objects, in particular people, that perform activities and interact creating object meaning such as events. A second observation is that the human visual system is able to search a video by quickly scanning it for activities and interesting events. In addition, to design widely applicable content-based video representations, the extraction of content independently of the context of the video data is required.

Moving objects and event-oriented semantic features are important and common for a wide range of video applications and video representation using solely low-level objects does not fully account for the meaning of a video. To fully

Correspondence to amer@inrs-telecom.uquebec.ca.

represent a video, objects need to be assigned high-level features as well. However, the difficulty is how to develop stable representation (e.g., in case of noise or occlusion) for large video shots that can be used in various applications.

Most current video representation systems are based on low-level quantitative features or focus on narrow domains. There are few representation schemes based on high-level features; most of these are context-dependent and focus on the constraints of a narrow application and they lack, therefore, generality and flexibility. Most systems assume simple environments, for example, without multi-object occlusion or noise. Therefore, current visual representation systems limit the user capability of taking decision based on automatically retrieved information from an input video. Since the focus when developing a video representation is the human operator of a video system, in the remainder of this paper we will explore positive answers to how to automatically extract stable features that are closer to how humans abstract video and closer to the needs of users of a video system.

2. A HIGH-LEVEL VIDEO REPRESENTATION FRAMEWORK

A high-level video representation can be based on key-frames, objects, activities, events, and/or semantic relations (e.g., interaction, object to object, object to event). High-level object features are generally related to the movement of objects and are divided into context-independent and context-dependent features. Features that have context-independent components include movement, activity, and related events. Here, *movement* is the trajectory of the object within the video and *activity* is a sequence of movements that are semantically related (e.g., *pitching a ball*).[2] Context-dependent high-level features include object *action* which is the semantic feature of a movement related to a context (e.g., *following a player*).[2]

We have developed a computational framework[4] that is designed to provide stable high-level video representations in terms of moving objects and their related semantic features. Moving objects are represented using quantitative and qualitative low-level features and generic semantic features using motion-related features such as events. To achieve higher applicability, content is extracted independently of the context of the input video. The reliability of our system is due to noise adaptation and due to correction or compensation of estimation errors at one step by processing at subsequent steps where higher-level information is available.

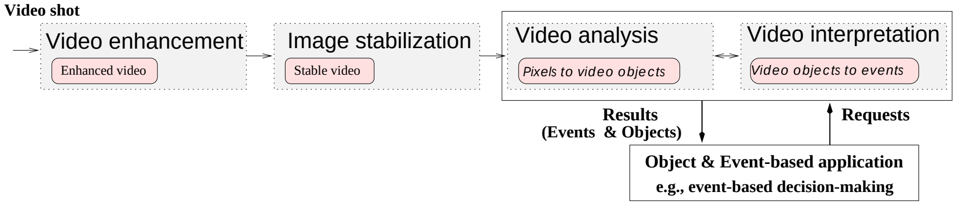

Figure 1: A framework for high-level stable video representation.[4]

Our system aims at: 1) flexible content representation, 2) reliable stable processing that foregoes the need for precision, and 3) low computational cost. Our system targets video of real environments such as those with multi-object occlusions and artifacts. To achieve these goals, four processing levels are proposed: video enhancement to estimate and reduce noise, image stabilization to compensate global motion and background changes, video analysis to extract meaningful objects and their spatio-temporal features, and video interpretation to extract context-independent semantic-related and qualitative video features. To extract semantic content, the motion behavior and low-level features of video objects are analyzed to represent important events and actions. Our system is thus modular, and layered where low (e.g., motion-based), middle (e.g., trajectory-based), and high (e.g., event based) layers interact.

The core of our system is the video analysis which consists of 1) motion-detection based object segmentation, 2) object-based motion estimation, 3) region merging, and 4) feature-voting based object tracking. The object segmentation module extracts objects based on motion and background data. In the motion estimation module, temporal features are estimated. The tracking module combines spatial and temporal features in an effective voting strategy that accounts for possible inaccuracies. Segmentation may produce objects that are split into sub-regions. Region merging intervenes to improve segmentation using tracking results. Region merging is based on temporal coherence and matching of objects rather than on local features. Video analysis modules may produce inaccuracies and much research has been done to enhance their performance. In our framework, we propose to compensate errors of low-level steps at higher levels when more reliable information is available.

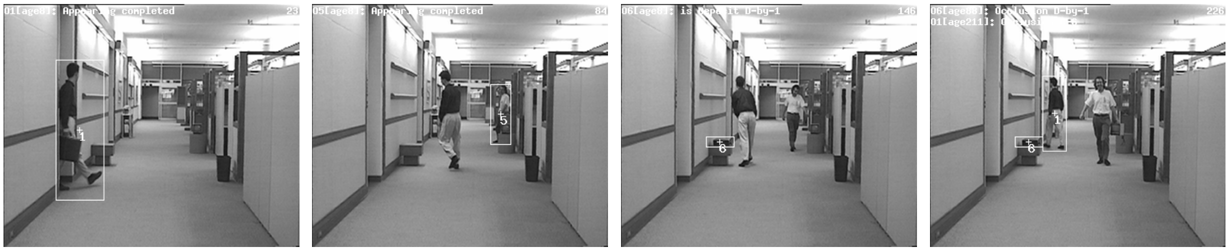

Figure 2: Key images of the 300 images of the 'Hall' shot (e.g., in image 146 O_6 is *deposited* by object O_1.)

3. APPLICATIONS AND RESULTS

In *video surveillance*, our framework can be used for automated monitoring of activity in scenes. Examples include 1) detecting people and related events such as fighting, or over-staying, 2) monitoring traffic and related unusual events, 3) detecting hazards such as fires, and 4) monitoring flying objects such as aircrafts. Tools for extracting and interpreting descriptions are essential for effective use of the upcoming *MPEG-7 standard*. On the other hand, a well-defined MPEG-7 standard will significantly benefit exchange among various video applications. In our system, a video is seen as a collection of video objects, related meaning, local and global features. This supports access to MPEG-7 video content description models. Our framework can be also used in *video database* applications, for example, to retrieve video based on specific events and objects. A high-level video representation can be used in *human motion analysis* applications to detect and recognize humans, to detect and understand human activities, or to analyze and measure human motion (e.g., to find location of joints). Human motion analysis applications include 1) dance performance, 2) athletic activities in sports, and 3) smart environments for human interactions. In *entertainment and telecommunications* applications, a high-level video representation can be used to facilitate video editing and reproduction, dynamic video summarization, browsing of video on Internet, or supporting the use of smart video appliances.

To test the reliability and real-time response of our system we have conducted extensive experimentations on various indoor and outdoor shots containing a total of 6371 images with multi-object occlusion, noise, and coding artifacts. [4] Due to space constraint, we only show here results of two video shots for two applications: key-frame extraction and video summary. Fig. 2 show key images extracted automatically by our system. The following table, generated automatically by our system, show a summary of the 655 images of the shot 'floorp'.

```
 'floorp' Shot Summary based on Objects and Events; StartPic 1/EndPic 655
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
|Pic | Event                 | ObjID | Age | Status    | Position            | Motion    | Size         |
|    |                       |       |     |           | start/present       | present   | start/present|
|--- | ----------------------| ----- | --- | --------- | ------------------- | --------- | -------------|
|13  | Entering completed    | 1     | 8   | Move      | (85 ,234)/(74 ,215) | (2  ,-5 ) | 3671  /3671  |
|--- | ----------------------| ----- | --- | --------- | ------------------- | --------- | -------------|
|460 | is Deposit by ObjID 1 | 2     | 8   | Stop      | (32 ,135)/(32 ,135) | (0  ,0  ) | 266   /266   |
|--- | ----------------------| ----- | --- | --------- | ------------------- | --------- | -------------|
|654 | Disappear             | 1     | 648 | Disappear | (85 ,234)/(51 ,108) | (0  ,1  ) | 3327  /85    |
|--- | ----------------------| ----- | --- | --------- | ------------------- | --------- | -------------|
```

ACKNOWLEDGMENTS This work was supported, in part, by the the Natural Sciences and Engineering Research Council of Canada under Strategic Grant SRT224122.

REFERENCES

1. "Video representation, coding, indexing (panel 2)." *Proc. Int. Workshop on Very Low Bitrate Video Coding*, Urbana, IL, USA, Oct. 1998.
2. A. Bobick, "Movement, activity, and action: the role of knowledge in the perception of motion," Tech. Rep. 413, M.I.T. Media Laboratory, 1997.
3. A. Pentland, "Looking at people: Sensing for ubiquitous and wearable computing," *IEEE Trans. Pattern Anal. Machine Intell.* **22**, pp. 107–119, Jan. 2000.
4. A. Amer, *Object and Event Extraction for Video Processing and Representation in On-Line Video Applications*. PhD thesis, INRS-Télécommunications, Dec. 2001. www.inrs-telecom.uquebec.ca/users/amer/phd/.