EXTRACTION OF HIGH-LEVEL VIDEO CONTENT FOR ADVANCED VIDEO APPLICATIONS

Aishy Amer
Concordia University, Electrical and Computer Engineering
Montréal, Québec, Canada *amer@ece.concordia.ca*

Abstract

Automated and effective techniques to extract video content such as moving objects and related semantic features have many applications in new video technologies. Video-based technologies include video database, video surveillance and video delivery over mobile terminals. Commentators predict that video surveillance (e.g., automated detection of unauthorised access) will grow dramatically in the coming years. The introduction of next-generation mobile services will make video contents accessible via mobile terminals. Automated content extraction would signi£cantly facilitate the use and reduce the costs of these applications.

This paper proposes a system for stable real-time extraction of high-level video content. The system consists of four interacting levels: enhancement to estimate and reduce noise, stabilisation to compensate for global changes such as global motion, analysis to extract moving objects, and interpretation to extract semantic features.

Our system represents a video in terms of moving objects and related semantic features such as events. To achieve higher applicability, content is extracted independently of the context of the input video. Our system, implemented on over 6000 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.

Keywords: *Object segmentation, event detection, context independence, video interpretation, video surveillance.*

1. INTRODUCTION AND RELATED WORK

Because of the ever-increasing needs for video storage, maintenance, and processing, developing automatic and effective techniques for *content-based* video representation have become crucial. A video shot consists, in general, of moving objects and their low and high-level features within a given environment and context. Developing advanced and ef£cient content-based video representation requires the resolution of two key issues: de£ning what are the most *important* and most *common* video contents and what level of features are suitable to represent these contents. 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 [7, 4]. A second observation is that the human visual system (HVS) is able to search a video by quickly scanning (”¤ipping through”) it for activities and events. In addition, to design widely applicable content-based video representations, the extraction of video content independently of the context of the video data is required.

Much work on video representation deals with the development of a generally applicable solution; however, there are few tests in the presence of noise and other artifacts. Most representation systems use mainly low-level features. Studies have shown that low-level features are not suf£cient for effective video representation and that objects must be assigned high-level features as well [4, 8]. Current systems use one or two processing levels to represent video content: analysis to extract low-level content [5, 2] and/or interpretation to describe content in semantic-related terms [6, 9]. Most high-level video representation systems are developed for narrow applications [4] and little work on context-*independent* representation exist.

2. PROPOSED SYSTEM

Without real-time consideration, a representation can lose its applicability. On the other hand, system stability is important for successful use. The objective of this paper is to develop an *automatic low-complexity modular* system for *stable* representation of video shots of real environments such as those with occlusions and coding artifacts. The system objective is achieved by 1) adaptation to noise, 2) correction or compensation of estimation errors at the various levels, and 3) division of the system into simple but effective tasks avoiding complex operations. The proposed system builds on the work in [3] and consists of two levels:

CCECE 2003 - CCGEI 2003, Montréal, May/mai 2003 0-7803-7781-8/03/$17.00

analysis (Sec. 2.1) and interpretation (Sec. 2.2). Fig. 1 displays a block diagram of the proposed system where $R(n)$ represents a background image of the shot and σ_n is the estimated noise standard deviation. An important feature of this system is that it is layered. For example, low-level object segments are used for tracking and tracking can correct and merge these segments if needed. Tracking together with merging are then used to detect events.

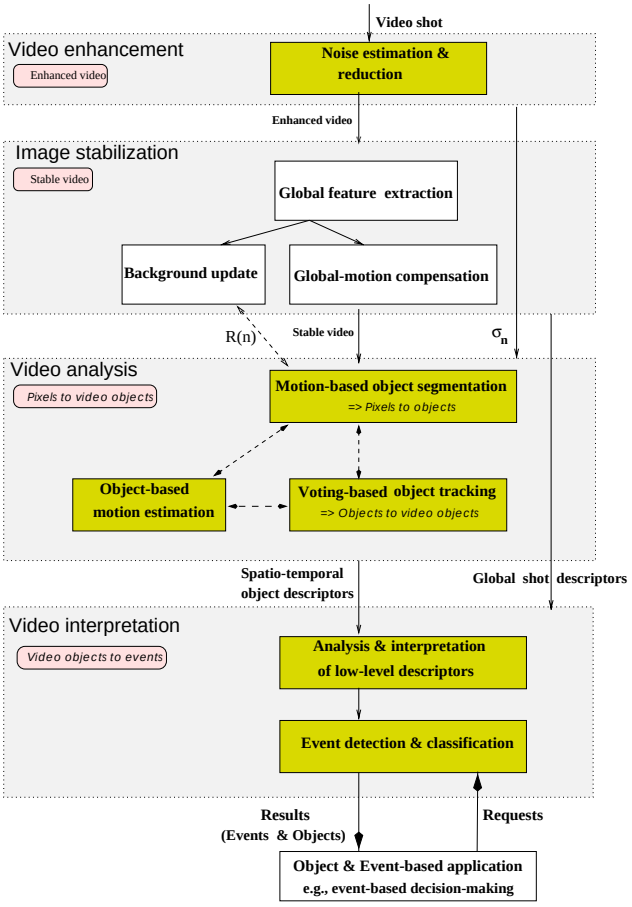

Fig. 1. Proposed framework for video representation.

This system outputs at each instant n a list of objects with their *identity* throughout the shot, *low-level features* (location, shape, size, motion), *trajectory*, *life span or age*, *event descriptions*, and *spatio-temporal relationship*. This can be used in applications where an interpretation ("what is this shot about?") is needed . For example, in video surveillance, event-oriented alarms can be activated. In video retrieval, high-level queries can be facilitated.

2.1. Object-oriented video analysis

The proposed video analysis 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 [3] 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. This paper proposes to compensate errors of low-level steps at higher levels when more reliable information is available. In each module of the proposed video analysis, complex operations are avoided. The current implementation of the proposed video analysis requires on average between 0.11 and 0.35 seconds to analyze the content of two images on a SUN-SPARC-5 360 MHz. For comparison, the current version (v.4x) of the reference method, COST-AM [2], takes on average 175 seconds to segment objects of an image.

The critical task of the proposed video analysis is motion detection which must remain reliable throughout the shot. Parameters of the proposed motion detection are automatically adapted to the estimated noise [3] and temporal adaptation which makes the procedure reliable throughout the shot is introduced. Furthermore, the elimination of small objects is adapted spatially to a homogeneity criterion and temporally to corresponding objects.

2.2. Context-independent high-level video interpretation

High-level features are generally related to object movement [4] and can be divided into context independent and context dependent features. High-level features such as events are generally applicable when they convey a £xed meaning that is independent of context. An event expresses a particular behavior of a £nite set of objects in a sequence of a small number of consecutive images of a shot. An event consists of components associated with time and location whose meaning may vary with context. For example, a *deposit* event has a £xed meaning (an object is added to the scene) but can have a variable meaning in different contexts.

With the information provided from the video analysis step, events are extracted in a straightforward intuitive manner by combining trajectory information with spatial features. Objects and their features are represented in temporally linked lists. Data is analyzed *as it arrives* and events are detected as they occur. The following are events automatically detected by our system (due to space constraints only two events are de£ned here):

• An object **enters**, **appears**, **exits**, **disappear**, and **stops**.

• Objects **occlude/occluded** and **remove/removed**.
• **Abnormal movements** is de£ned when an object *stays long* or *moves (too) fast/slow*.
• A **dominant object** 1) is related to a signi£cant event, 2) has the largest size, or 3) has the largest speed or age.
• An object, $O_i \in I(n)$, **moves** in image $I(n)$ if
  ◦ $M_i : O_p \rightarrow O_i$ (a function assigning O_p at time $n-1$ to an object O_i at time n) where $O_p \in I(n-1)$, and
  ◦ the median of the motion magnitudes of O_i in the last k images is larger than a threshold.
• O_i **deposits** O_j (or O_j is **deposited** by O_i) if
  ◦ $O_p \in I(n-1)$, $O_i, O_j \in I(n)$, and M_i,
  ◦ $O_j \notin I(n-1)$, i.e., no match $M_0 :\dashv O_j$,
  ◦ $\frac{A_j}{A_i} < t_a$, $t_a < 1$ being a threshold,
  ◦ $A_i + A_j \simeq A_p \quad \wedge[(H_i + H_j \simeq H_p) \vee (W_i + W_j \simeq W_p)]$, where A_i, H_i, and W_i are area, height, and width of O_i,
  ◦ O_j is close to a side, s, of the minimum bounding box (MBB) of O_i. $s \in \{r_{\min_i}, r_{\max_i}, c_{\min_i}, c_{\max_i}\}$. Let d_{is} be the distance between the MBB-side s and O_j. O_j is close to s if $t_{c_{\min}} < d_{is} < t_{c_{\max}}$ with thresholds $t_{c_{\min}}$ and $t_{c_{\max}}$, and
  ◦ O_i changes in height or width between $I(n-1)$ and $I(n)$ at the MBB-side s.

Only if the distance between the deposited object and depositor is large the event *deposit* is considered. Otherwise O_j is assumed have split from O_i and is merged to O_i. To reduce false alarms, *deposit* is declared if the deposited object remains in the scene for some time. This is important to differentiate between deposit and segmentation errors. Stopping objects (e.g., seated person or stopped car) and deposited objects are also considered.

These proposed events are suf£ciently broad for a wide range of applications to assist on-line supervision of, for example, the removal/deposit of objects in a surveillance site, the behavior of traf£c objects, and the behavior of customers in stores or subways. Other events and composite events, such as stand, sit, walk, object lost, and found, can be easily extracted based on our event detection strategy. For example, *approach a restricted site* can be easily extracted when the location of the restricted site is known.

3. RESULTS

Extensive experiments using widely referenced shots have shown the effectiveness of the proposed framework. Tests has been conducted on more than 10 shots containing a total of 6071 images with multi-object occlusion, noise, and artifacts of indoor and outdoor environments. The proposed system works in real time for surveillance applications with a rate of up to 10 frames/second.

Samples of our results for the proposed video analysis are shown in Fig. 2–5. Fig. 2 shows that the proposed motion detection method is more reliable than a statistical motion detection method [11, 1] especially in images with local illumination change and noise. Fig. 3 shows objectively that the proposed object segmentation is better than the reference method, the COST-AM method, with respect to spatial accuracy $sQM(dB)$ criterion[10]. Fig. 4 displays subjectively that the proposed method remains robust to variable object size and is spatially more accurate. Fig. 5 displays reliably estimated trajectories using the proposed tracking method.

The performance of the proposed interpretation is illustrated here by samples of our results for surveillance applications. Fig. 6 shows images of key events extracted automatically where only objects performing events are annotated (ID and MBB). The good performance of the system is a result of special considerations to handle inaccuracies and false alarms such as differentiating between deposited objects and split objects.

4. CONCLUSION

This paper presented a computational framework to automatically and ef£ciently extract 1) meaningful video objects and 2) useful context-independent events. The proposed events are suf£ciently broad to assist applications such as monitoring 1) of removal/deposit of objects, e.g., computing devices, 2) of traf£c objects, and 3) behaviors of customers, e.g., in stores. The reliability of the proposed system has been demonstrated by extensive experimentations on more than 10 indoor and outdoor shots containing a total of 6371 images with object occlusion, noise, and coding artifacts. In this framework, special consideration is given to processing inaccuracies and false alarms. For example, the system is able to differentiate between deposited objects, split objects, and objects at an obstacle. Errors are corrected or compensated at higher level level where more information is available. The proposed system provides a response in real-time for surveillance applications with a rate of up to 10 frames per second on a SUN-SPARC-5 360 MHz. Further research is planned in classi£cation of motion as 'with purpose' (vehicle or people) and 'without purpose' (trees). In addition, the detection of background objects that move during the shot needs to be explicitly processed.

Acknowledgment: This work was supported, in part, by the the Natural Sciences and Engineering Research Council (NSERC) of Canada.

5. REFERENCES

[1] T. Aach, A. Kaup, and R. Mester. Statistical model-based change detection in moving video. *Signal Process.*, 31(2):165–180, 1993.

[2] A. Alatan, L. Onural, M. Wollborn, R. Mech, E. Tunceland, and T. Sikora. Image sequence(1) analysis for emerging interactive multimedia services - the European COST

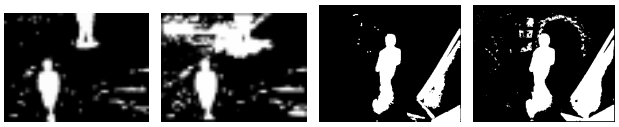

Fig. 2. Better detection: motion detection comparison

211 Framework. *IEEE Trans. Circuits Syst. Video Technol.*, 8(7):802–813, Nov. 1998.

[3] A. Amer. *Object and Event Extraction for Video Processing and Representation in On-Line Video Applications*. PhD thesis, INRS-Télécommunications, Univ. du Québec, Dec. 2001. www.ece.concordia.ca/ amer/phd.

[4] A. Bobick. Movement, activity, and action: the role of knowledge in the perception of motion. Technical Report 413, M.I.T. Media Laboratory, 1997.

[5] S. Chang, W. Chen, H. Meng, H. Sundaram, and D. Zhong. A fully automatic content-based video search engine supporting multi-object spatio-temporal queries. *IEEE Trans. Circuits Syst. Video Techn.*, 8(5):602–615, 1998. Special Issue.

[6] J. Courtney. Automatic video indexing via object motion analysis. *Pattern Recognit.*, 30(4):607–625, 1997.

[7] A. Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(1):107–119, Jan. 2000.

[8] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval: the end of the early years. *IEEE Trans. Pattern Anal. Machine Intell.*, 22(12):1349–1380, Dec. 2000.

[9] E. Stringa and C. Regazzoni. Content-based retrieval and real time detection from video sequences acquired by surveillance systems. In *Proc. IEEE Int. Conf. Image Processing*, pages 138–142, Chicago, IL, Oct. 1998.

[10] P. Villegas, X. Marichal, and A. Salcedo. Objective evaluation of segmentation masks in video sequences. In *Proc. Workshop on Image Analysis for Multimedia Interactive Services*, pages 85–88, Berlin, Germany, May 1999.

[11] F. Ziliani and A. Cavallaro. Image analysis for video surveillance based on spatial regularization of a statistical model-based change detection. In *Proc. Int. Conf. on Image Analysis and Processing*, pages 1108–1111, Venice, Italy, Sept. 1999.

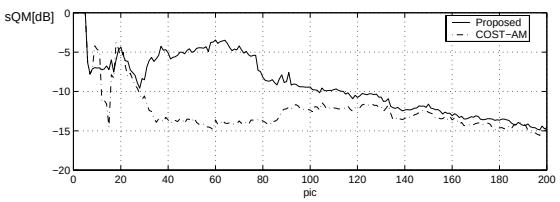

Fig. 3. More accurate results: objective comparison of object masks for the 'Hall' shot.

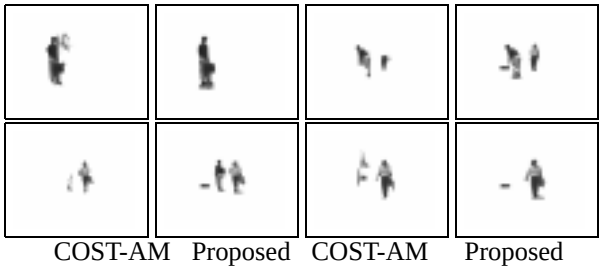

Fig. 4. Better segmentation: subjective comparison of object segmentation for the 'Hall' video.

Fig. 5. Accurate estimated trajectories of the objects in the shot 'Highway' using the proposed video analysis system. 'StartP' is start position.

Fig. 6. Indoor surveillance: key events of the 300 images of the 'Hall' shot, e.g., O_6 is *deposited* by object O_1.