Mining Surveillance Video for Independent Motion Detection

Zhongfei (Mark) Zhang
Computer Science Department
Watson School of Engineering and Applied Science
State University of New York (SUNY) at Binghamton
Binghamton, NY 13902, USA
zhongfei@cs.binghamton.edu

Abstract

This paper addresses the special applications of data mining techniques in homeland defense. The problem targeted, which is frequently encountered in military/intelligence surveillance, is to mine a massive surveillance video database automatically collected to retrieve the shots containing independently moving targets. A novel solution to this problem is presented in this paper, which offers a completely qualitative *approach to solving for the automatic independent motion detection problem* directly *from the compressed surveillance video in a faster than real-time mining performance. This approach is based on the linear system consistency analysis, and consequently is called* **QLS**. *Since the* **QLS** *approach only focuses on what exactly is necessary to compute a solution, it saves the computation to a minimum and achieves the efficacy to the maximum. Evaluations from real data show that* **QLS** *delivers effective mining performance at the achieved efficiency.*

1 Introduction

A target in motion from a surveillance video may be interpreted as a potential suspicious activity. If the camera is still, the problem of automatic detection of motion from the video is trivial. However, in many applications, it is *not possible* to have a still camera. An example is in the automatic data collection in military surveillance using unmanned aerial vehicles (UAVs), such as US Predators. In this case, the surveillance goal is to detect any military maneuvers, which are typically manifested as the target motion in the video. Note that since in this case the cameras in UAVs are also in motion, the problem of detecting any target motion is consequently translated to the problem of detecting *independent motion* (IM) — the motion other than the camera motion. Fig. 1 shows exemplary frames from the real US Predator surveillance videos that represent two different scenarios: a scene with IM and a scene without IM. Due to the fact of large scale, automatic data collection (multiple Predators in nonstop data collection) in a typical military surveillance, the data volume is *massive*. It is noted that due to the fast development of unmanned surveillance and data collection technologies, the existence of such massive surveillance databases is ubiquitous. On the other hand, manual mining for the detection of IM is painfully proven to be extremely tedious and prohibitively expensive. Consequently, solutions to automatically mining video data to retrieve shots containing IM are in high demand. This paper addresses this data mining problem, and motivated by this demand, a highly efficient and effective data mining algorithm is developed in this research for automatically retrieving any shots that contain IM from a surveillance video.

There are two scenarios related to IM detection. Given a video sequence, *quantitative IM detection* refers to *temporal* segmentation into those shots that contain the scene in which one or more independently moving targets are present, and *spatial* segmentation and delineation of each of the independently moving targets in each of the frames of these shots. *Qualitative IM detection*, on the other hand, refers to only the *temporal* segmentation of the video sequence to return those shots that contain IM; it does not perform spatial segmentation to identify the independently moving targets in each frame. The focus of this paper is primarily on qualitative IM detection. Taking the motivated applications of the US military surveillance data shown in Fig. 1, once the shots containing IM are *automatically* detected and retrieved, the major painstaking and tedious mining effort (i.e., manual searching the massive video data to detect those shots containing IM) is saved because the majority of the video does not have IM. Therefore, in terms of the data mining concern for detecting IM, the objective is qualitative IM detection from the temporal sequence of the video, as opposed to quantitative IM detection in all the frames.

Motion analysis has been a focused topic in computer vision and image understanding research for many years [8, 3, 2]. IM analysis deals with multiple motion components simultaneously, and therefore, presumably is more challenging.

Most of the existing techniques for IM detection in the literature aim at quantitative detection [7, 4, 1]. Due to this fact, very few of them can afford efficient performance, as their solutions to temporal IM detection depend on spatial IM segmentations. While quantitative detection is useful in general, due to the specific applications that have motivated this project, a qualitative approach is sufficient. This is based on the following two reasons. (i) In the military and intelligence applications, the time issue, i.e., the detection efficiency, is always an important concern. Obviously the qualitative approach saves time as the spatial segmentation in the image domain in each frame is avoided. (ii) It is not necessary to take a quantitative approach in these applications. Even if the independently moving targets are all segmented and identified in each frame in the quantitative approaches, given the current status of computer vision and artificial intelligence in general, it is *not possible* to have a fully automated capability to interpret whether the segmented and identified IM in the frames indicates any military or intelligence significance without human expertise' interaction. Therefore, these detected shots must be left to the Image Analysts for further analysis anyway, *regardless* of whether or not the independently moving targets are segmented and identified in each frames of these shots.

The other observation is that in the literature, most of the existing techniques for IM detection are based on image sequences, as opposed to compressed video streams. This restriction (or assumption) significantly hinders these techniques from practical applications, as in today's world, information volume grows explosively, and all the video sequences are archived in compressed forms. This is particularly true in the applications this paper concerns, in which the data volume is *massive* and they must be archived in a compressed form, such as MPEG.

Based on these considerations, we have developed a completely *qualitative* approach to solving for the automatic IM detection problem *directly* from the compressed surveillance video in an *efficient* performance. By an efficient performance, it is meant that the data mining speed is faster than the real-time performance. This capability allows two possible application scenarios for this technology. The first is to equip this algorithm with the sensors to allow real-time data mining while the sensors are in surveillance. The second is to mine an archived surveillance video database in which all the video data are stored in a compressed format; the fast scanning performance allows efficiently automatic mining the data to retrieve shots containing IM. This qualitative approach is based on the linear system consistency analysis, and consequently is called **QLS**.

2 QLS

Assuming that the camera model is a 3D to 2D affine [5], it can be shown [10] that given n macroblocks in a frame of an MPEG compressed video stream, we can build a linear system:

$$\boldsymbol{D_m} = \boldsymbol{\xi_m} \boldsymbol{b_m} \qquad (1)$$

with the following theorem:

Theorem 2.1 *Given n macroblocks in a video frame represented in the linear system in Eq. 1, if there is no IM with any of these macroblocks, then the linear system is consistent.*

The consistency of Eq. 1 is defined by determining the value of the statistic R:

$$R = \frac{\sigma_{min}(\boldsymbol{D_m})}{\sigma_{min}(\boldsymbol{D_m b_m})} \qquad (2)$$

where $\sigma_{min}(\boldsymbol{D_m})$ and $\sigma_{min}(\boldsymbol{D_m b_m})$ are the smallest singular values of the coefficient matrix $\boldsymbol{D_m}$ and the augmented matrix $\boldsymbol{D_m b_m}$, respectively, assuming Eq. 1 has unique solution if it is consistent; multiple solution cases may be handled similarly. Consequently, Eq. 1 is consistent *iff* R is above a threshold.

In MPEG compression standard, for each macroblock in a frame, if this macroblock is inter-coded, there is a motion vector available. Since the macroblock information (including the motion vector and the center coordinates) can be easily obtained directly from a compressed MPEG video stream, we have a linear system Eq. 1 that can directly work on the MPEG compressed data without having to depend on a specific algorithm to compute the correspondence or optical flow between the two frames, and without having to decompress the video stream [6]. If the macroblock is intra-coded, we just exclude this macroblock from the linear system of Eq. 1. If the frame is an I frame in which all the macroblocks are intra-coded, $R = 1$. This could be a false positive, which can be easily removed by filtering the R statistics, resulting in rejection of this false positive in the final detection.

We use the normal flow [9, 7] to detect IM. The rationale is that if the normal flow is low, the motion vector is probably not accurately estimated; consequently this macroblock should be rejected from Eq. 1.

Now the **QLS** algorithm is summarized as follows, which takes four parameters: the normal flow threshold T_n, the scan window width r, the R statistic threshold T_R, and the defined minimum number of frames T_f of a segment that contains IM.

```
Scan an input video stream in compressed MPEG
For every pair of consecutive frames
    Start to build up the linear system Eq. 1
    For each macroblock $\boldsymbol{M}$ of frame $l$ of the pair
      Estimate the normal flow $\nabla I(\boldsymbol{M})$ of $\boldsymbol{M}$
      If $\nabla I(\boldsymbol{M}) > T_n$
          Incorporate $\boldsymbol{M}$ into Eq. 1
    Compute $R$ of the linear system Eq. 1
Compute the median filtered $\bar{R}$ over a window of $r$
If $\bar{R} - 1 > T_R$
    Label $l$ as no IM (NIM)
Else, label $l$ as a frame with IM (IM)
Any IM segment $> T_f$ is retrieved
```

3 Experimental Evaluations

We have implemented the **QLS** as a stand alone version in a Windows2000 platform with Pentium III 800 MHz CPU and 512 MB memory. Fig. 1(c) and (d) show the *original* R statistics computed at every frames for the two shots from two surveillance videos in Fig. 1(a) and (b), respectively. The statistics are obvious to tell whether and where there is IM in the video. The first shot containing 1119 frames describes an IM of a missile launcher moving to its destination. The mean of the original R is 1.0 and the deviation is 0.00122 over the 1119 frames. The second shot containing 1058 frames surveys an area of ground terrain with no IM. The mean of the original R is 1.389 and the deviation is 0.169 over the 1058 frames. A separate evaluation with over 160,000 frames of real surveillance data indicates an 81.27% precision and a 93.6% recall of **QLS** [10].

Since **QLS** essentially just needs to compute the R value for each frame, and since in each frame there is typically a very limited number of macroblocks, the complexity of **QLS** is very low. The current implemented version of **QLS** scans a compressed MPEG video with a typical frame resolution of 240 by 350 at the speed of 35 frames/second under the current platform, which is already faster than real-time. Note that this implementation is just for proof of the concept and the code has not been optimized. This shows that **QLS** holds great promise and vitality in the future applications in both proposed scenarios: real time data mining equipped with the sensors and fast data mining for an archived database.

4 Conclusions

This paper presents an efficient and effective approach to automatically mining surveillance video data for IM based on a qualitative, linear system approach called **QLS**. As compared with the existing techniques and available technologies, the **QLS** has the following distinctive advantages: (i) No camera calibration is required or necessary, i.e., image coordinates directly from the video frame may be used without having to convert them into calibrated coordinates. (ii) The statistics computed in the algorithm are stable due to the *Low condition numbers* of the matrices, resulting in avoiding the unstable matrix computation problem of high condition numbers typically existing in many computer vision and image understanding techniques. (iii) No specific motion model is assumed, i.e., **QLS** is able to detect IM for any motion models, either planar or parallax motion, or either dense parallax or sparse parallax camera motion. (iv) **QLS** is able to detect IM only based on two frames, as opposed to some techniques in the literature requiring more than two frames. (v) Due to the qualitative nature, the **QLS** complexity is very low, and is able to have efficient detection. (vi) **QLS** directly works on the compressed data; it does not need to decompress a video before applying the detection. (vii) **QLS** only requires one camera video stream to be able to detect IM, as opposed to some techniques in the literature that require stereo video streams.

Acknowledgment

This work is supported in part by NEC Research Institute, Inc. at Princeton, NJ, DARPA, and AFOSR. The author acknowledges Stoyan Kourtev and Xunyin Wang at SUNY Binghamton, Binghamton, NY, for part of the implementation of **QLS**. The author is also grateful to David W. Jacobs at NEC Research Institute for many useful and enlightening discussions.

References

[1] A.A. Argyros and S.C. Orphanoudakis. Independent 3D motion detection based on depth elimination in normal flow fields. In *Proc. International Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society Press, 1997.

[2] O. Faugeras. *Three-Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, 1993.

[3] T.S. Huang and C.H. Lee. Motion and structure from orthographic views. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 11:536–540, 1989.

[4] M. Irani and P. Anandan. A unified approach to moving object detection in 2D and 3D scenes. In *Proc. of IUW*, 1996.

[5] D.W. Jacobs. *Recognizing 3-D Objects Using 2-D Images*. Ph.D. Dissertation, MIT AI Lab., 1992.

[6] S-W. Lee, Y-M. Kim, and S.W. Choi. Fast scene change detection using direct feature extraction from MPEG compressed videos. *IEEE Trans. Multimedia*, 2(4):240–254, 2000.

[7] R. Sharma and Y. Aloimonos. Early detection of independent motion from active control of normal image flow patterns. *IEEE Trans. SMC*, 26(1):42–53, 1996.

[8] S. Ullman. *The Interpretation of Visual Motion*. MIT Press, 1979.

[9] A. Verri and T. Poggio. Motion field and optical flow: qualitative properties. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 11(5):490–498, 1989.

[10] Z. Zhang. Qualitative independent motion detection. *Computer Science Tech Report*, 2002.

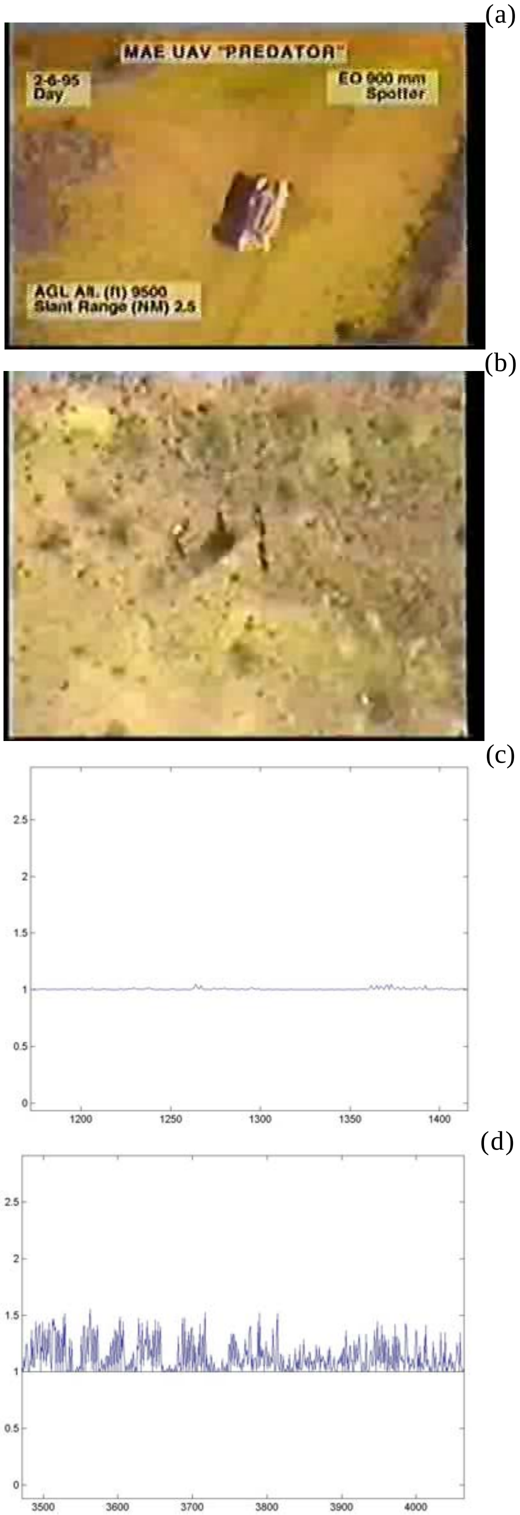

Figure 1: (a) An example of a shot containing an independently moving target (a missile launcher) (b) An example of a shot containing no IM (terrain) (c) The original R statistics computed for the shot in (a) (1119 frames) (d) The original R statistics computed for the shot in (b) (1058 frames).