Computer Vision and Image Understanding 144 (2016) 237–257

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Temporal mapping of surveillance video for indexing and summarization

Saeid Bagheri, Jiang Yu Zheng*, Shivank Sinha

Indiana University Purdue University Indianapolis, United States

ARTICLE INFO

Article history:
Received 15 January 2015
Accepted 28 November 2015

Keywords:
Surveillance video
Temporal profile
Image projection
Target counting
Crowd analysis
Video summarization
Video indexing
Video retrieval
Video compression

ABSTRACT

This work converts the surveillance video to a temporal domain image called *temporal profile* that is scrollable and scalable for quick searching of long surveillance video by human operators. Such a profile is sampled with linear pixel lines located at critical locations in the video frames. It has precise time stamp on the target passing events through those locations in the field of view, shows target shapes for identification, and facilitates the target search in long videos. In this paper, we first study the projection and shape properties of dynamic scenes in the temporal profile so as to set sampling lines. Then, we design methods to capture target motion and preserve target shapes for target recognition in the temporal profile. It also provides the uniformed resolution of large crowds passing through so that it is powerful in target counting and flow measuring. We also align multiple sampling lines to visualize the spatial information missed in a single line temporal profile. Finally, we achieve real time adaptive background removal and robust target extraction to ensure long-term surveillance. Compared to the original video or the shortened video, this temporal profile reduced data by one dimension while keeping the majority of information for further video investigation. As an intermediate indexing image, the profile image can be transmitted via network much faster than video for online video searching task by multiple operators. Because the temporal profile can abstract passing targets with efficient computation, an even more compact digest of the surveillance video can be created.

1. Introduction

1.1. Objectives

A video shows location (*where*), time (*when*), people/objects (*who*/*what*), and actions/events (*how*). A surveillance video, however, has a relatively fixed area to monitor and simple actions (passing of object and people) as compared to the entertainment and sports videos. The time is much longer for identifying who and what passed through in the field of view. It is a major task to screen surveillance videos captured day and night to find target objects and persons. Surveillance cameras have been located everywhere for monitoring targets and events, and video data are constantly collected day and night. Searching targets in large video volumes from distributed cameras is not a trivial task. An automatic scanning of candidates in videos and the confirmation by human operators are eventually necessary.

This paper generates a profile image to index the entire surveillance videos emphasizing when targets passed (time) critical locations. The indexing means that an examiner of surveillance video can click at the identified targets or suspects in such a profile image for further investigation in the video frames. The profile image can also be used efficiently for long term survey and direct counting of large group or passing flow. In this paper, we make efforts to preserve the target shape in the temporal profile of video for target identification (*who*/*what*). Detecting and screening targets and their motion in the profiles (*when*/*how*) are easier and faster than watching the video, and so as to visualizing the target moving directions and positions (*where*) for understanding events. Figs. 1 and 2 show the preliminary framework of obtaining temporal image from a video sequence [25].

1.2. Related works

Automatic identification of pedestrians and cars has been one of the hottest topic in video surveillance [12,32]. Background and foreground separation have been extensively explored [10]. Among detected foreground regions, pedestrian detection is carried out by using shape [36] and motion [34,47] information based on supervised learning techniques [33,35]. Motion tracking further extracts trajectories of targets for event understanding [37]. However, these methods have not achieved close to perfect results; many of them have accuracy around 80% depending on the training data. Therefore, current video screening still requires involvement of human operators. In addition, all of the methods require intensive computation that is much longer than playing video in real time.

* Corresponding author.
E-mail addresses: bagheris@cs.iupui.edu (S. Bagheri), jzheng@iupui.edu (J.Y. Zheng).

http://dx.doi.org/10.1016/j.cviu.2015.11.014
1077-3142/

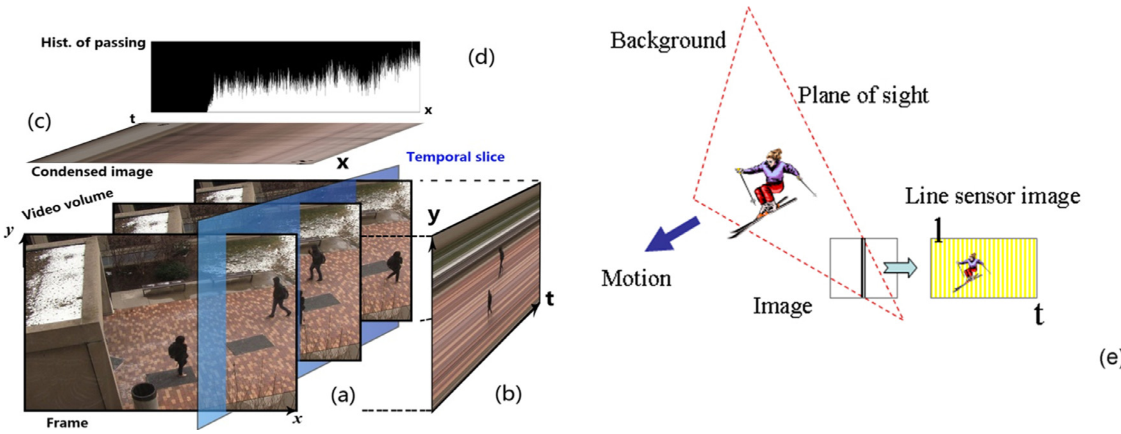

Fig. 1. Principle in generating a temporal slice image from consecutive sampling at a line in the video frame. (a) Video volume. (b) Temporal slice. (c) A condensed image by averaging colors vertically. (d) Passing scope in the condensed image. (e) Target passing a plane of sight that is through a sampling line in the 3D space. Pixels on the sampling line are corrected continuously from consecutive frames to obtain a temporal image.

Fig. 2. A sampling line in the frame and its generated temporal profile. (Left) A frame with the vertical sampling line depicted. (Right) Profile collected from the sampling line recording people both at front and in the hallway behind the window.

To reduce the time of searching video exhaustively, there have been many works categorized as *temporal condensing of frames in spatial domain* in order to shorten the video lengths. One is to remove video segments without foreground object events [39] or even compress different temporal events together [[26]– [28]] into short video clips. Another way along the same line is to condense events in different frames into key frames as index by mosaicing foreground patterns [41]. Although these compressions of frames to the same spatial domain have reduced data sizes, they lose temporal information completely either across clips or within clips. The visual space may also get cluttered quickly by crowded scenes over prolonged time periods such that non-uniformed clip selection and sampling are required to be done offline based on overall target density and event complexity in the video [40].

An alternative way to index video is to focus on *temporal domain* [25,29] at critical positions in the field of view where the targets of interest will pass through, because the surveillance cameras always watch at fixed locations (*where* is known). Similar as the linear CCD camera in the early stage of digital imaging [3,4,6,7,25], we set a fixed pixel line in the video frame for collecting temporal data as used in traffic monitoring [11,25] with the camera mounted at a high position. In many applications such as counting passing people in a park, station, store or exhibition site [21], alarming invasion of a critical facility or cross of border, and monitoring a traffic flow, the target moving directions and speeds are already obvious. We can thus focus on counting targets and recording shapes. If penetrating objects are mainly pursuing a translational movement through some sampling line, the resulting 2D temporal slice image contains information on time, shape and identities [3–6,23,25] but requires less redundant processing than normal video [12,15]. It is also more intuitive to examine the history of passing entities than a video and discrete *synopsis* clips or key frames, because it is temporally continuous and extendible to long periods, which serves the purpose to index temporal data in surveillance videos.

The third method is combining *spatial-temporal* presentations by copying the moving targets to shifted positions along the time axis

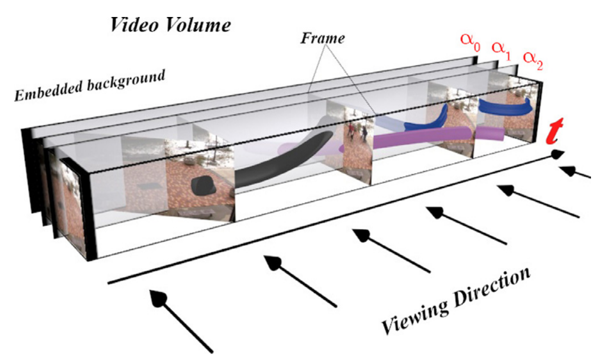

Fig. 3. The scheme of temporal profile from video: watching video volume from right side to have precise temporal information. The time axis is always displayed rightward in video track in video software. This transformation of video to a continuous indexing image can facilitate smooth content visualization over a long period.

in the video volume of a clip [44–46], after the segmentation process to separate foreground from background as *synopsis*. The display of video volume allows viewer's interaction to examine the targets and actions from different angles. This method is more powerful but more costly in visualization, searching and retrieval of videos. A comprehensive comparison of the three methods is given in the discussion section later on.

1.3. Methods proposed and contributions

In this work, we introduce an intermediate image representation named *temporal profile* created from the surveillance video for saving computation cost and data size. The resulting data have one dimension as time; it corresponds to watching the video volume from right side as Fig. 3 depicted. We provide a much faster approach than playing video for finding *when* some targets (*who*/*what*) pass through specified critical locations (*where*), since it is the primary task for