Adaptive Summarisation of Surveillance Video Sequences

Jian Li\textsuperscript{1}, Stavri G Nikolov\textsuperscript{2}, Christopher P Benton\textsuperscript{1}, Nicholas E Scott-Samuel\textsuperscript{1}
\textsuperscript{1}Department of Experimental Psychology, 12A, Priory Road, BS8 1TU
\textsuperscript{2}Department of Electrical and Electronic Engineering, MVB, BS8 1UB
University of Bristol, UK

Abstract

We describe our studies on summarising surveillance videos using optical flow information. The proposed method incorporates motion analysis into a video skimming scheme in which the playback speed is determined by the detectability of interesting motion behaviours according to prior information. A psycho-visual experiment was conducted to compare human performance and viewing strategy for summarised videos using standard video skimming techniques and a proposed motion-based adaptive summarisation technique.

1 Introduction

Visual surveillance systems can generally be used for two main tasks: performing real-time online event detection, crime prevention and scene analysis; and performing offline analysis and the retrieval of interesting events from archived surveillance footage. For the second task, a control room operator has to go through an entire video. Because interesting events, such as suspicious targets and criminal behaviour, occur infrequently, an operator may spend most of the time viewing frames without any useful information. Psychologically, going through a long uneventful surveillance video may impair operator performance.

In this paper, we present our current work on adaptive summarisation of surveillance videos, aiming to use shorter time to browse an archived surveillance video without reducing accuracy of event detection.

1.1 Video Summarisation Techniques

Video summarisation techniques usually aim to generate a summary of a video using key frames or video skims. Key frame extraction techniques select one or several still images from a video which could largely represent the content. For example, Gong and Liu [3] adopted Singular Value Decomposition (SVD) to remove redundant information so that an optimal or nonredundant summarisation could be obtained. Alternatively, Wolf [10] utilised motion energy in terms of optical flow information to identify stillness. In this work, it was assumed that either the camera will stop moving or the gesture will be held when important information is to be illustrated. In these situations, the motion energy reaches a local minimum and the corresponding frames can be used as the key frames of the shot. Porter [9] incorporated tracking techniques into the summarisation scheme to distinguish motions caused by gradual translation and motions caused by camera or object movement so that the key frames selected were the ones which were largely due to changes of visual content.

Another technique to summarise videos is called video skimming. Rather than using key frames to represent the whole video, video skimming aims to generate a short image sequence in which the temporal information can also be illustrated. In this technique, the simplest method is to drop video frames at a constant rate. However, without considering any visual content, the dropping procedure could erase informative content from the original video. To alleviate this problem, adaptive skimming schemes have been proposed. Ma and Zhang [6] assumed that the human visual system will concentrate on those local regions with high motion intensities and frames containing such information were retained in the summarised video. Peker and Divakaran [8] computed visual complexity using motion information and used this to determine which frames to retain.

Psychology researchers have studied the influence of video summarisation on human perception. Ding and Marchionini [2] examined the effects of different display speeds of key frames on the tasks of target identification and video comprehension. They concluded that identification accuracy decreases as the display speed increases, and once the speed is higher than a certain level the identification task becomes impossible. Compared to identification, the video comprehension task was less influenced by higher speeds of display and the top speed limit for comprehension was higher than that for identification.

1.2 Methods

Summarising surveillance videos aims to generate a concise representation of unsafe or threatening behaviours so
that control room operators or security officers are able to review and analyse these events precisely and efficiently. Our method combines motion detection with video skimming using constant frame dropping. We detect interesting events by analysing optical flow based motion vectors and matching these with prior information. If no interesting events are detected, we use constant frame dropping. Otherwise, frames with interesting events are retained in the summarised video.

Applying a constant frame rate to the summarised video results in adaptive playback speeds based on the detectability of interesting events. Besides a basic way of evaluating performance by measuring missing frames and false alarms, we conducted a psycho-visual experiment to examine human perception and performance on summarised videos using the adaptive and constant playback speeds. During our experiment, we recorded gaze fixations to try to analyse and understand strategies used for viewing the summarised videos.

2 Adaptive Video Skimming

The standard video skimming technique generates a summarised video by constantly dropping a number of frames, for example \(N - 1\) frames in every \(N\) frames, from a normal speed video. With the same frame rate, the summarised video is then displayed at \(N\) times the normal speed, denoted as \(N\times\). Without considering visual content, the use of a large value of \(N\) results in significantly reducing the time to browse a video, but at the expense of limiting time to detect and analyse interesting events (which may be very short in duration). This problem motivates the generation of an adaptive summarisation scheme.

Our adaptive summarisation technique combines a motion detection procedure with a standard video skimming scheme, whose general structure is shown in Figure 1 (a). With two successive video frames \(G(t)\) and \(G(t+1)\), optical flow based motion information is estimated. Motion processing and analysis are then applied in a pixelwise manner to extract features which can be compared with prior information of interesting events defined in advance. The form of interesting events varies for different tasks. Here by ‘interesting events’, we mean motions with certain velocity and orientation which happen in pre-defined regions of interest (ROIs). Once an interesting event is detected, the corresponding frame is then regarded as an important frame and is retained in the summarised video. Constant frame dropping is applied to those frames that do not contain interesting events. When using the same frame rate as the normal speed video, those successive frames are displayed at a normal speed while others are displayed at a speed of \(N\times\). An explicit representation of adaptively switching playback speeds has been shown in Figure 1 (b).

It can be seen that the adaptive scheme has two advantages: firstly, slow display of frames with interesting events gives operators more time to analyse the events; secondly, the remaining frames without interesting events connect slow displayed segments so that the temporal coherence and continuity are retained.

2.1 Optical Flow Estimation

Optical flow estimation plays an important role in our adaptive scheme. Optical flow measures displacements of pixels within the same pattern between successive video frames \([5, 1]\). Here we use the framework proposed by Li et al \([4]\) in which a least squares (LS) or a modified total least squares (TLS) regression scheme is associated with an extended optical flow constraint (EOFC) \([7]\). Here the noise level of the input video frames determines which regression method should be adopted. Initially, moving patterns are segmented from backgrounds. In our work, we use a temporal differentiation technique in which a pixel whose absolute temporal derivative is higher than a given threshold is regarded as moving. The LS regression scheme is then applied to those moving pixels only.

2.2 Motion Processing and Analysis

The pre-processing aims to generate meaningful features from pixelwise estimates of the optical flow vectors. The form of motion features to be extracted is determined by the
form of interesting events. For the task of detecting targets moving in a certain direction, we segment a video frame into blocks with equal size. Each block is associated with a dominant motion vector which is computed as the mean of the estimated optical flow vectors within this block. Subsequently, the orientation of a dominant motion vector can be computed as an angle in the interval $(-180^\circ, 180^\circ]$.

### 2.3 Video Summarisation Results

We examined the performance of the adaptive summarisation scheme by generating an example summarised video which slows down the playback speed when targets moving rightwards in Region B of the traffic scene shown in Figure 2 have been detected.

![Figure 2: The Region of Interest.](image)

A clip which consists of 9600 frames was extracted from a longer video recorded by a static camera at a frame rate of 25 frames per second (fps). Frame size was 720x576 pixels. In this task, interesting events were defined as the angles of the dominant vectors falling in the interval $[-20^\circ, 20^\circ]$. If the angles of more than one local dominant motion vectors fall in the angular interval, interesting events were considered as being detected and the corresponding frame was retained in the summarised video. For those frames without interesting events, a constant frame dropping ($N = 16$) was performed. Figure 3 shows an explicit illustration of event detection from the input video with normal playback speed. Here frames with and without interesting events are coloured as red (dark) and green (bright), respectively. The summarised video contains all frames coloured as red and some frames extracted from those coloured as green using constant frame dropping.

![Figure 3: Detection of Interesting Events.](image)

The initial evaluation of the performance of the summarised video is based on four criteria: ratio of length (R-L), false positive (F-P, type I error), false negative from frame missing (F-N FM, type II error). The ratio of length represents the ratio of lengths between the summarised video and the raw video. The false positive (F-P) corresponds to the number of frames mistakenly retained in the summarised video which do not contain interesting events. This is equivalent to the number of false alarms. The false negative from target missing (F-N TM) refers to the number of missed targets with interesting events. The false negative from frame missing (F-N FM) refers to the number of missed frames containing interesting events.

Table 1 shows the performance of the adaptive summarisation technique. It can be seen that the summarised video has a high compression rate with low false positive (or false alarms). The zero F-N TM indicates that all targets moving from left to right in the ROI are retained in the summarised video, i.e. for each interesting event there is at least one frame with the target present in it that has been retained in the summarised video. 113 frames containing the targets with interesting events were missed using the adaptive summarisation scheme. This is largely due to the problem that some motions of the targets cannot be reliably detected.

<table>
<thead>
<tr>
<th>Raw Video</th>
<th>Summ. Video</th>
<th>R-L</th>
<th>F-P</th>
<th>F-N TM</th>
<th>F-N FM</th>
</tr>
</thead>
<tbody>
<tr>
<td>9600</td>
<td>2942</td>
<td>30.65%</td>
<td>46</td>
<td>0</td>
<td>113</td>
</tr>
</tbody>
</table>

### 3 Human Visual Perception and Performance Experiments

In the previous section, we presented the performance of the adaptive summarisation technique in terms of data and information retainment. Below, we examine human performance using the summarised videos.

#### 3.1 Design

The experiment we conducted had three goals: firstly, to gain a better understanding of the relationship between playback speeds and human performance regarding a given task; secondly, to evaluate human performance on the summarised videos using constant frame dropping scheme and the adaptive scheme; and finally, to differentiate psychological strategies used for viewing videos of different playback speeds with the assistance of eye tracking techniques.

In the experiment, the task was to report the total number of people moving rightwards in region B in Figure 2. During the experiment, eye movements were recorded with an eye tracker. When preparing the videos, we recorded three
1-hour long videos of the same scene (see Figure 2 for example). Each video had a frame rate of 25 fps and a frame size of 720x576 pixels. Seven equal-length clips were then manually selected from the videos, each of which consisted of 9600 frames. The visual information in these clips was very similar but not identical. The number of targets for each clip (C1-C7) is shown in Table 2.

Table 2: Correct Number of Targets

<table>
<thead>
<tr>
<th>Number</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>37</td>
<td>30</td>
<td>36</td>
<td>32</td>
<td>41</td>
<td>53</td>
<td></td>
</tr>
</tbody>
</table>

Each clip was used to generate seven summarised videos in which six summarised videos with speed of 1x, 2x, 4x, 8x, 16x, 32x were generated using constant frame dropping and one summarised video was generated using the adaptive scheme. The lengths of videos with constant playback speeds are shown in Table 3. The seven adaptive summarised videos have different lengths because of variation of the visual content. The average length is 2’07” with a standard deviation of 20”. The length is approximately equivalent to the summarised video at 3x speed using constant frame dropping.

Table 3: Lengths of Summarised Videos (Constant Speeds).

<table>
<thead>
<tr>
<th>Speed</th>
<th>1x</th>
<th>2x</th>
<th>4x</th>
<th>8x</th>
<th>16x</th>
<th>32x</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frames</td>
<td>9600</td>
<td>4800</td>
<td>2400</td>
<td>1200</td>
<td>600</td>
<td>300</td>
</tr>
<tr>
<td>Duration</td>
<td>6’24”</td>
<td>3’12”</td>
<td>1’36”</td>
<td>48”</td>
<td>24”</td>
<td>12”</td>
</tr>
</tbody>
</table>

There were therefore 49 summarised videos in the database. During the experiment, each participant was shown seven summarised videos which were randomly selected from the database according to the following rules: 1) visual information is different among videos; 2) each participant will view seven clips that all have different speeds, i.e. there are no two clips that have the same speed.

3.2 Participants

16 participants took part in the experiment (8 males). Ages ranged from 19 to 56 years (mean=28.4, s.d.=10.36). Participants had normal or corrected-to-normal vision.

3.3 Experimental Settings

Video stimuli were shown using the ClearView 2.6.8 software package, on a 3.2 GHz Intel Pentium D processor PC with 2 GB RAM. A 22” flat screen CRT monitor running at 85 Hz was used to display the stimuli. The screen resolution was set to 800x600 pixels. Participants were required to use a chin-rest positioned 68cm from the monitor screen. A Tobii™ x50 remote eye tracker was used to collect eye movement data. This is a table-mounted eye tracker that samples eye positions at 50 Hz with an approximate accuracy of 0.5°.

3.4 Procedure

Seven videos were displayed one by one with a short break between each. After showing one video, participants were asked to report the total number of targets they detected from the video. Participants were also asked after the experiment to provide general comments about the setup and the experiment.

3.5 Data Analysis

The accuracy of each playback speed can be evaluated by calculating the difference between the number of detected targets and the number of existing targets. The latter was determined manually by visually inspecting the videos and it is shown in Table 2. For a certain playback speed, the average and the standard error for all participants were computed. The standard error (SE) is defined as:

\[ SE = \frac{\sigma}{\sqrt{P}}, \]

where \( \sigma \) is the standard deviation of the errors and \( P \) is the total number of participants (\( P=16 \)).

The eye tracker records gaze locations for both eyes at 50 Hz. Each point is associated with a duration. For a certain playback speed, we generated a Fixation-Duration Map (FD Map) using the ClearView software package by grouping all the Type I fixations which last for more than 100 milliseconds within a radius of 30 pixels.

With the Type I fixations, we characterised the degree of concentration at a certain playback speed by computing the ratio between the number of fixations within the ROI and the number of all fixations for all participants. To compare the strategies used for viewing videos with different playback speeds, we generated a Fixation-Duration Map (FD Map) using the Type II fixations. Each Type II fixation is associated with a duration. For a given playback speed, we scanned the fixations for all participants and summed the durations at the same fixations. The durations were then normalised by dividing a factor representing the length (duration) of the video for this playback speed. These normalised FD Maps show how participants manage their eye movements.
3.6 Results

The mean errors and the mean absolute errors are shown in Figure 4. It can be seen that when the playback speed is less or equal to 8x, the accuracy is high and the individual differences are small. When the playback speed increases to 16x or 32x, the mean of the absolute errors increases as does the standard error, indicating large individual differences.

From Figure 4, it can be seen that the adoption of the motion-based summarised videos guarantees accuracy. The mean errors and mean absolute errors by using the proposed videos are $-0.63 \pm 0.27$ and $0.75 \pm 0.25$, each of which is the smallest among all playback speeds. When taking the compression rate into consideration, the lengths of the adaptive summarised videos are equivalent to the lengths of the video at the speed 3x using constant frame dropping.

![Figure 4: Human Performance Accuracy.](image)

Figure 5 shows the ratios of the number of the Type I (raw) fixations within the ROI. For all constant playback speeds which are higher than or equal to 4x, and the adaptive speed, the ratios are higher than 95%. In contrast, with the original speed, the ratio falls to 82%. This shows that as the playback speed increases, participants tended to concentrate on detecting targets in the ROI instead of looking around the scene.

The same conclusion can be drawn from the fixation-duration maps shown in Figure 6, in which each fixation is associated with a circle. The center of the circle indicates the position of the fixation and the radius is proportional to the normalised duration of the fixation. It can be seen that as the speed increases, fewer fixations appear outside the ROI. For the speeds 1x, 2x, 4x and 8x, the radii of the circles are small. This means participants made frequent eye movements rather than just fixating the same position. From our observation of the gaze gatherings, the eye movements were used for looking around or performing smooth pursuit of a group of targets to extract the precise number of the targets. At 16x, there are still fixations with short durations. Participants were trying to track and calculate the targets in groups. However, the compromised performance shown in Figure 4 indicates that the tracking was less successful than the lower playback speeds as participants did not have enough time to calculate the precise number of the targets. At 32x the larger radii show that participants preferred to fixate at a limited number of positions rather than tracking targets, as there was no time to track a single group of targets. Finally, the short durations of the fixations for the adaptive playback speed indicate frequent eye movements when viewing the video. The high detecting accuracy confirms the successful target tracking using the adaptive playback speed. To summarise, in order to obtain accurate estimates, both high concentration and effective target tracking are important. The evaluations of the detection accuracy and gaze fixations confirm that human performance was superior when using the proposed adaptive summarised surveillance videos.

3.7 Subjective Comments

Participants also commented on the usefulness of the summarised videos and the overall experimental setup. Most participants thought the 1x speed is slow for successfully accomplishing the task, so for most of the time they could look round the whole scene. In contrast, 16x and 32x were too fast for them so that precise calculation of the number of targets was hard to achieve. Most participants gave positive feedbacks to 2x and 4x for their efficiency and clarity showing the targets. Participants thought 8x of being slightly too fast. Regarding the videos with adaptive playback speeds, most of participants gave positive feedbacks. One subject,
a surveillance officer, reported that adaptively changing the playback speed helped him concentrate on the region where interesting events could happen. When the speed slowed down, he assumed that something would happen and then quickly moved his eyes to the left part of the ROI.

Figure 6: The Fixation-Duration Maps. Top row: 1x, 2x; Second row: 4x, 8x; Third row: 16x, 32x; Bottom row: Adaptive.

4 Conclusions

The current paper presents our work on summarising surveillance videos using adaptive playback speed and examining human performance when identifying interesting events using summarised videos with constant and adaptive playback speeds. The technical results showed that the summarised videos with adaptive playback speeds ensure both accuracy and efficiency for detecting targets. The same conclusion was also obtained from an experiment requiring participants to count visual targets. Gaze-fixation and eye-movement analysis implies that the superior performance was largely due to the fact that the adaptive summarised videos allows more attention to be focused on the ROI and more time to perform successful target tracking.

Acknowledgments

This work has been funded by the UK MOD Data and Information Fusion Defence Technology Centre (DIF DTC).

References