VIDEO PRE-ANALYZING AND CODING IN THE CONTEXT OF VIDEO SURVEILLANCE APPLICATIONS

Amal Ben Hamida¹,², Mohamed Koubaa¹, Henri Nicolas², Chokri Ben Amar¹

¹ REGIM: REsearch Groups on Intelligent Machines, University of Sfax, ENIS: National Engineering School of Sfax, Route de Soukra, B.P. 1173, 3038 Sfax, TUNISIA
² LaBRI: Laboratoire Bordelais de Recherche en Informatique, University of Bordeaux 1, 351 Cours de la Libération, 33405 Talence, FRANCE
{benhamida.amal, mohamed.koubaa, chokri.benamar}@ieee.org, henri.nicolas@labri.fr

ABSTRACT
In many video surveillance applications, only a very small proportion of the information provided by the cameras is really useful. To reduce the video data stream transmitted, one solution is to pre-analyze the video at the sensor, compress and send out only the valuable information that will be analyzed more precisely at the control station. In this paper, we propose a video pre-analyzing and coding system in the context of video surveillance applications aiming to simplify as possible as we can the captured data without altering the interesting information for traffic analysis. The originality of the method described here lies in the gain in term of video H.264/AVC compression by reducing the prediction error values during the motion estimation step.

Index Terms— video surveillance, background subtraction, spatio-temporal filtering, analysis, compression.

1. INTRODUCTION
In the context of video surveillance applications, the purpose is usually to be able to detect events of interest such as: accidents, thefts, or the presence of unauthorized persons. The typical use of such data is either to have a human operator monitoring the screens directly connected to the sensor via a broadband network, or to use the records subsequently. To enable a more efficient use of the videos, it is necessary to devise methods that can automatically detect events of interest and provide an efficient compression, such as H.264/AVC, when the available bandwidth cannot transmit the raw video in real time. Therefore, the typically used scheme is: compress, transmit, decompress and analyze the video. Through this paper, we focus on the pre-analysis stage by presenting a Regions Of Interest (ROI) extraction and a spatio-temporal video filtering methods to detect and simplify areas that may contain information of interest. We differ from many others works, such as [1,2,3,4], for which ROI are extracted to be coded with high quality. Our main goal is to improve the compression efficiency by degrading the ROI quality but with keeping sufficient information for efficient traffic analysis purposes. Experiments will be carried out especially in the context of road traffic and public places monitoring.

2. VIDEO PRE-ANALYZING
Appending the pre-analysis stage to the habitual used scheme requires some additional computations at the sensor level. In return, it accelerates the data transmission and analysis. To be effective, it should reduce the data amount but not destroy the areas bearing useful information for the later events detection and interpretation. This phase involves two main stages: The ROI are first detected and then treated by a spatio-temporal filtering method to obtain a simplified video sequence.

2.1. Regions of interest extraction
Since we focus on captured videos for roads and public places monitoring, areas containing moving objects such as pedestrians, vehicles represent the required regions. This areas are full of information about the events of interest. As presented in Fig. 1, this stage combines two major steps.
Generally, in video surveillance applications, the separation between the background and the moving objects is a crucial step. The popular method used for background modeling is the probabilistic approach: Mixture of Gaussians Model (GMM) presented in [5]. It consists in a weighted sum of Gaussian densities, which allows the color distribution for each pixel to be multi-modal. The values of a particular pixel is modeled by $k=3$ or $k=5$ distributions of Gaussians. The mixture of Gaussians method is robust to slow lighting changes, periodic motions from a cluttered background, slow moving objects, long term scene changes, and camera noises. In our system, the background is detected using an improved version of [5]: improved adaptive mixture of Gaussian by Zivkovic et al. [6]. Recursive equations are used to constantly select the appropriate number of components for each pixel. By choosing this number in an on-line procedure, the algorithm can automatically fully adapt to the scene. The segmentation results and the processing time are improved. Since our system is adapted for roads and public places monitoring, moving objects shadows are noising information that may disrupt the regions of interest extraction step. That's why, we proceed to shadows removal using the method proposed by Prati et al. [7]. Those calculated foregrounds represent the ROI describing the moving objects and their motions. The background model is regularly updated. The insignificant pixels (singular pixels not belonging to any object) are then removed by using morphological operations, essentially erosion and dilation. The binary mask is estimated by using the connected components and flood filling techniques. Hence, via these masks, the ROI are extracted from the video sequence, as shown in Fig. 2.

2.2. Regions of interest simplification

As described earlier, only small portions of the recorded video represent the relevant information which is the extracted regions in section 2.1. Since we are working for traffic analysis purposes, the goal is to analyze the moving objects behaviors. The ROI generally contain too detailed information for the traffic analysis applications which generates unnecessary coding and computational costs during the H.264/AVC encoding phase. For example, in road monitoring applications, the needed information is the spatio-temporal descriptors, such as the object’s color, form, trajectories, speed and/or the object’s classification (as car, truck, pedestrian...). Otherwise, the rest is useless. Based on this assumption, we present in this part our contribution to filter the extracted data in order to keep only the used ones and at the same time to more simplify the compression computation costs taking into account the application’s needs (traffic analysis).

2.2.1. H.264/AVC Encoding process

During the encoding process, the inter frame prediction is one of the key point to get high compression quality. If the motion compensation is efficient, the prediction error is low and therefore, the compression ratio is high. For this reason, we aim at obtaining a more accurate macro-blocks prediction step in order to minimize as possible the prediction error and then narrow down the computation costs for a better compression stage. This is done using a spatial and temporal filtering as will be shown in the next section. For more detailed explanations about H.264/AVC compression, please refer to [8].

2.2.2. Spatio-temporal filtering

Guaranteeing a precise macro-block prediction means finding a block with color components values almost similar to the current ones. The encoding cost is strongly determined by the motion compensation error. Therefore, it is essential to minimize it. For this reason, the target of this proposed method is to simplify the video data, the way that during the compression process, every two corresponding blocks have almost the same value. And so, the prediction error is too close to zero and the encoding cost becomes minor. Starting from the video containing just the extracted ROI (see section 2.1); for each GOP (Group Of Pictures), the basic idea is to apply a spatial filtering to $I$ frames and then a temporal filtering to every $P$ frame. As expressed in Fig. 3, each filtered $I$ frame represents the reference frame for the encoding phase and the rest filtered frames form the $P$ (predicted) frames of the GOP. We presume here that the GOPs of the encoded video contain only $I$ and $P$ frames.

Spatial filtering: The spatial filtering of the reference frame consists in sliding a window through the image. Every time, the robust mean of the window's RGB pixels values is calculated and is attributed to the value of the window's...
central pixel. It represents the mean of the \( T \) major colors values in the RGB color histogram \((T\) is experimentally fixed). The goal of using a robust mean is to obtain a color value representing the dominant colors in the window. So that, by simplifying the image the detailed color information is reduced to the mean value of the major colors.

**Temporal filtering:** The temporal filtering is used to simplify the \( P \) frames of the GOP based on the frames already filtered. The purpose is to modify the frames blocks depending on their corresponding blocks in their reference image. For each ROI in the \( N\)-th frame of the GOP, we search for its similar ROI in the previously filtered frame using the block matching algorithm. The matching score \( R \) is computed based on the normalized square difference matching method. For each block and if the matching succeeds \( (R < T_H, T_H \text{ experimentally fixed threshold}) \), all the ROI's pixels of the block will be changed by their corresponding values provided by the matching algorithm. Else, the ROI is spatially filtered.

**Coding:** The motion vectors used for the temporal filtering are re-used within the compression scheme. The prediction error is therefore null for the blocks temporally filtered, and reduced for the spatially filtered ones. In this way, the prediction error is generally very low. Then, the compression ratio is higher and more rapidly computed (lower computation costs).

As earlier explained, the application context does not require too detailed information as presented in the original captured video. So that, by the spatio-temporal filtering, we maintain enough information not only for the object's tracking, but also for the object's classification (as car, truck, pedestrian...) and the color's characterization (see Fig.4). It means that sufficient information is kept for most traffic analysis goals (except for the detection of a specific car or pedestrian which requires the knowledge of unique descriptors such as license plate or human face detection).

### 3. PERFORMANCES EVALUATION

Experiments have been done in order to evaluate the performances of the proposed approach. The tested videos belong to ViSOR, i-LIDS and PETS benchmark datasets. They are captured sequences for road traffic and public places with different levels of complexity (pedestrians and / or vehicles, with and without occlusions, large or small objects...). The used GOP size is 12 frames. The used sliding window has a sliding step equal to 1 pixel and a 7x7 pixels size. It was fixed after experimental tests which demonstrate that beyond this size the visual quality of the extracted regions, essentially regions representing small objects (such as pedestrians), is degraded: we can no more distinguish their details (color, texture, body, hair, clothing...). Threshold \( T \) is fixed to 60%. It should be noted that the simplified frames does not only contain the ROI; they are formed by the background model spatially filtered superimposed by the spatio-temporally filtered ROI.

To assess the ROI extraction stage, precision and recall values are estimated for each sequence's image by calculating the percentage of pixels correctly belonging to moving objects. In Table 1, we evaluate the performance of our used method [6,7] versus the GMM method proposed in [5]. We compare the results with the ground truth, which consists here in a manual extraction of the objects in each frame. Recall and precision values obtained with our method are better than the GMM (an increase from 2% to 20% is observed). They vary between 0.78 and 0.98 depending on the complexity of the scene. As earlier detailed, the ROI simplification leads to a better compression. To evaluate this stage, the obtained simplified sequences are encoded to H.264/AVC format using the H.264/14496-10 AVC reference software (http://iphone.hhi.de/suehring/tml). The performance is depicted in Fig. 5 where we compare the prediction errors values computed for every macro-block of the original sequence's frames and its simplified version. We can distinguish the huge reduction of its values after the spatio-temporal simplification. These values are very small that is to say that the matched blocks are very close. In the studied figure, many values are null which proves the exact similarity between the macro-blocks. In Fig. 5(b), the first obtained values are very close to the original ones. In fact, they are computed for blocks belonging to a new moving object entering the scene (Red encircled object in the joined

---

Figure 3. The filtered I frame represents the I frame for the encoding phase and the rest filtered frames forms the P frames of the GOP (GOP size=12).

Figure 4. (Up) Frames 808,923 of traffic monitoring sequence in VISOR and S1 sequence in PETS. (Low) Extracted simplified ROI.
levels during the final analysis stage. In order to exploit different simplified video scalability objects. Secondly, we will use the spatio-temporal filtering depending on the nature and the size of the moving contribution to an intelligent adaptive spatio-temporal have several perspective ideas. Firstly, we will improve our contribution versions with an increase between 1% and 11% relatively to the original version. This proves that almost all the moving objects are tracked correctly. Even if they have been significantly simplified, they still contain enough information for traffic analysis which mainly consists in objects behavior analysis and classification.

To more validate the approach, we applied an objects tracking algorithm to the decoded simplified sequence. We aim at successfully tracking the moving objects in the scene even after they have been simplified. We used the blob tracking system of the OpenCV library. Precision and recall values in Table 2 (Bottom) reach over 0.78 for the simplified version with an increase between 1% and 11% relatively to the original version. This proves that almost all the moving objects are tracked correctly. Even if they have been significantly simplified, they still contain enough information for traffic analysis which mainly consists in objects behavior analysis and classification.

4. CONCLUSION

In this paper, we presented a pre-analysis and coding method for video surveillance applications: roads and public places monitoring. It was validated experimentally that we succeeded in extracting and simplifying the ROI without altering the requested information by the fact that even higher recall and precision values for object tracking have been obtained using the simplified decoded video. At the same time, we fulfilled a significant reduction of the H.264/AVC coding cost. Through experiments, we demonstrate that the achieved reduction rate is over than 70%. Bitrates and time consumption are reduced too. This proves that almost all the moving objects are tracked correctly. Even if they have been significantly simplified, they still contain enough information for traffic analysis which mainly consists in objects behavior analysis and classification.

5. REFERENCES


<table>
<thead>
<tr>
<th>Table 1. Precision and recall computed for the ROI extraction stage using the GMM method [5] (GM) and the used method [6,7] (UM)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
</tr>
<tr>
<td><strong>Precision</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. (Top) Bitrate reduction, (Middle) Motion estimation time, (Bottom) Precision and recall values computed for moving objects tracking in the (O) original and the (S) simplified versions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td><strong>Bitrate (bits/s)</strong></td>
</tr>
<tr>
<td>O</td>
</tr>
<tr>
<td>S</td>
</tr>
<tr>
<td><strong>Reduction rate (%)</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ME time (ms)</th>
<th>Visor</th>
<th>Camer a</th>
<th>Video1</th>
<th>PVHard</th>
<th>China</th>
</tr>
</thead>
<tbody>
<tr>
<td>O</td>
<td>216009</td>
<td>171010</td>
<td>979652</td>
<td>998054</td>
<td>471242</td>
</tr>
<tr>
<td>S</td>
<td>134876</td>
<td>113808</td>
<td>664791</td>
<td>788462</td>
<td>350903</td>
</tr>
<tr>
<td><strong>Reduction rate (%)</strong></td>
<td>37.63</td>
<td>33.45</td>
<td>35.15</td>
<td>21.01</td>
<td>25.54</td>
</tr>
</tbody>
</table>

Figure 5. Prediction error values computed for macro-blocks in VISOR sequence (frame 63) and camera1 sequence (frame 760)