Video object extraction and its tracking using background subtraction in complex environments

Satrughan Kumar*, Jigyendra Sen Yadav

Department of Electronics and Communication, MANIT, Bhopal 462003, India

Received 5 February 2016; accepted 12 April 2016
Available online 26 April 2016

KEYWORDS
Background subtraction; Entropy; Initial motion field; Object tracking

Summary Background subtraction is an efficient way to localize and obtain the centroid of the connected pixels moving on the foreground despite the prior information of the scene. It is suitable under fixed camera arrangement, which incorporates many vision applications such as object tracking, human monitoring, etc. However, the moving object extraction task becomes sophisticated and challenging due to some annoying factors such as local motion in background (waving tree, rippling water, etc.), camouflage region, sleeping object, which in turn degrades the tracking performance. In order to alleviate these problems, an efficient background subtraction algorithm is proposed to support the object-tracking task under static and dynamic background conditions. The work is focus to realize the relevant moving blobs on foreground by aiding the proper initialization and updating of the background module in order to improve the tracking accuracy. It generates an initial motion field using spatial-temporal filtering on the consecutive video frames. The block-wise entropy is evaluated above a certain range of the pixels of the difference image in order to extract the relevant moving pixels from the initial motion field. A suitable threshold value is estimated to assign an appropriate label to the moving blobs on the foreground mask. Finally, an adapting Kalman filter is integrated to the object extraction module in order to track the object on the foreground. Extensive quantitative experiments prove that the proposed method competently handles the object extraction, which in turn improves the tracking task under static and dynamic background conditions.

© 2016 Published by Elsevier GmbH. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Introduction

Many computer vision applications requires real time segmentation of moving object, which is a vital step in target localization, traffic monitoring and intruder activity analysis (Bouwman, 2014). The algorithm based on background subtraction is one of the most admired method to realize moving objects on the foreground (Vosters et al., 2012; Nikolov and Kostov, 2014; Xue et al., 2012). The background subtraction task in object tracking offers several advantages such as it provides higher accuracy at pixel, frame as well as in regional level operation. In addition, it is computationally inexpensive and does not require prior information about the scene to localize the moving object on the foreground mask. Moreover, the background initialization and reconstruction are proving to be effective under static camera arrangement in order to get the initial estimate of motion. The intention behind extracting the blob of moving object is to provide the complementary statistical information about the regions. As seen, the Kalman filter requires an initial estimate of position of an object (blob) in order to get the improved trajectory during tracking (Ha, 2012; Weng et al., 2006). In this case, the background subtraction provides the sufficient size of tracking samples or blob region, which is not obtained through corners and edge detectors.

In this work, a background model and object extraction module are proposed to enhance the object segmentation process that also provide a reliable input to the tracking mechanism. Simultaneously, it provides the potential to obtain the foreground blobs, which are free from any artificial trails, ghost, aperture and holes distortion (Cucchiara et al., 2003). The remaining sections of this paper are organized as follow: second section presents some background subtraction schemes used in tracking. In third section, we present our background model and object extraction procedures. Experimental results and performance analysis are shown in fourth section. The concluding remarks are given in final section.

Related works

Background subtraction plays a vital role in order to detect the interesting moving objects and tracking of such objects in the consecutives frame of video sequence. In Bouwman (2014), some background models are reviewed to observe their characteristics. The review provides the clue to take the benefits of the combined approach of basic and statistical models. Since, the integral approach can be robust to noise and operate at pixel level to solve many problems. An adaptive thresholding scheme is combined with temporal averaging method is used to solve the problem of traditional frame difference method (FD) (Nikolov and Kostov, 2014). However, it could not detect the complete blob on the foreground. The same problem is found in (Ha, 2012; Mazinan and Amir-Latif, 2012). In Rahman et al. (2013), the authors used the second order filter in the gradient direction to get relevant edges of moving object by applying the traditional background-updating scheme. The method could not support during real time applications and in noise suppression. In Zhang and Ding (2012), an adaptive background model using mean filter is proposed to eliminate the impulsive noise. As seen, the mean filter is less robust to hollow space and not adapting to the variance of pixels.

In Xue et al. (2012), author proposed a phased based background model. It extracts the foreground pixel using distance transform. Although the method is robust to find the foreground pixels in changing illumination and bootstrapping condition, yet it suffers from time complexity.

In Stauffer and Grimson (2000), authors proposed a multi-valued background-modelling scheme in which each pixel of the background model is modelled using Gaussian mixture model (GMM). The method is found to be effective against dynamic background, yet it does not discriminate the background and foreground pixel near camouflage region and in changing illumination condition. The statistical method employed in (Oral and Deniz, 2007), does not update the background model that makes it less useful under varying illumination conditions. In Vosters et al. (2012), the authors proposed the combined approach of Eigen background and statistical illumination model to solve the disruption due to rapid illumination. However, it could not solve the problem of local motion because Eigen space of background model was not incorporating the next position of object.

We can recapitulate here that some effective models have higher computational complexity, while the simpler background models are not effective under complex conditions. It is noted that the effective background reconstruction and sufficient object sample size are necessary to extract and track the object on the foreground (Mandelllos et al., 2011). In this work, we have included the spatial and temporal constraint to get the object identification and removed unnecessary background pixels by integrating the region level operation (Yao and Ling, 2014).

Proposed method

This section illustrates the proposed method into two stages. The first stage is about the background modelling and objects extraction, while the second stage is related to object tracking using Kalman filtering (Weng et al., 2006). It is noted that the work is focussed on grey scale videos, taken under static camera arrangement.

Background modelling and object extraction

The steps of the proposed method are followed according to Algorithm 1 that deals the background modelling and object extraction phase. Initially, average of some initial frames is taken as reference background. As seen, the temporal processing creates holes and avoids spatial correlation amongst the moving pixels. Therefore, an approximate motion field is derived using the background subtraction and temporal difference mechanism. The diversity state of each pixel is examined through spatio-temporal filtering and an approximate initial motion field is estimated by thresholding the background pixel. A background model should adapt to the dynamic changes, whether it is local (swaying tree, rippling water, and fountain) or global (resolution or illumination change). In this work, the proposed background model adapts temporal changes that efficiently avoided the common unwanted variables and extracted the complementary region in the scene.
Algorithm 1. Object extraction module

1. Average 'R' initial frames \(l_1, l_2, l_3, \ldots, l_k\) to generate reference background \(B^0_i(x, y)\), where \(0 < K < 11\)

2. for \(F \leftarrow 1\) to \(N\) do // \(N = \) total number of frames

3. \[X = \text{abs}(l_i(x, y) - B^0_i(x, y))\]
   \[FD = \text{abs}(l_i(x, y) - l_{i-1}(x, y))\]

   // \(l_i(x, y)\) and \(l_{i-1}(x, y)\) are current and previous frame respectively.

4. Initialize thresholds \(\tau_1\) and \(\tau_2\)
   \[\tau_1 = \text{mean}(X) + \gamma'(\text{std}(X)).\]
   \[\tau_2 = \text{mean}(FD) + \gamma''(\text{std}(FD)).\]

5. for \(x \leftarrow 1\) to rows of frame do

6. for \(y \leftarrow 1\) to columns of frame do

7. if \(X(x, y) > \tau_1\) then
   \[B^P_i(x, y) = 1;\]
   else
   \[B^P_i(x, y) = 0;\]

8. if \(\text{abs}(l_i(x, y) - l_{i-1}(x, y)) > \tau_2\) then
   \[B^D_i(x, y) = 1;\]
   else
   \[B^D_i(x, y) = 0;\]

9. if \((B^P_i \cup B^D_i)\) true then
   \[B^B_i(x, y) = X(x, y);\]
   else
   \[B^B_i(x, y) = 0;\]

end end

// Calculate entropy of initial motion field

10. Define a window of size \(r \times r\), where \(r = 8\).

11. for \(i \leftarrow 1\) to number of rows incremented by \('r'\)

12. for \(j \leftarrow 1\) to number of columns incremented by \('r'\)

13. \[M = B^0_i(i \rightarrow i + r, j \rightarrow j + r)\]

14. Discard the pixel with intensity smaller than \('3'\).

15. Calculate 'pdf' using the pixels of \(M\)

16. // pdf \(\rightarrow\) probability density function

17. Calculate entropy

\[E_i = -\sum_{R=\text{Rmin}}^{R=\text{Rmax}} \text{pdf}(R) \log(\text{pdf}(R))\]

18. if \(E_i(i, j) > \tau_1\) then // \(\tau_3 = 3\)
   \[M(i, j) = B^H_i(i, j);\]
   else
   \[M(i, j) = 0;\]

19. To calculate \(\tau_4\)
   - Initialize \(\tau_4\) by taking the mean of 'X'.
   - Update \(\tau_4\) using the given equation:
     \[\tau_4 = X(x, y) + (x \times \mu_M - \beta Y(x, y))\]
     where \(\mu_M\) is mean of 'M' and 'Y' is the current average of updated background frame.

20. Update current background using
     \[B_i(i, j) = B_{i-1}(i, j) + \text{signum}(l_1(i, j) - l_{i-1}(i, j))\]

end end

In complex environment, the state of pixels in \(B^0_i(x, y)\) is affected due to heavy local or global changes and causing the appearance of irrelevant pixels on the foreground. In this regard, we monitor the regional entropy to get the actual moving pixels in the approximate initial motion field. The pixels belonging to the actual moving object has low entropy as compared to those false negatives pixels that arise due to either dynamics or illumination changes. During the probability density function evaluation, we have discarded lower grey level having the range below '3'. Based on the block-wise entropy, an initial motion field is estimated using proper threshold. Finally, the largest area of moving blob is labelled after a morphological open operation followed by a close operation on \(D_i(i, j)\).

Object tracking using Kalman filtering

An adaptive Kalman filter algorithm proposed in Weng et al. (2006), has been adopted to estimate the centre of the largest moving blob inside the rectangle. Despite the presence of error due to unconstrained measurement, local motion in background, the Kalman filter has an ability to estimate tracking positions using the minimum information about the blob. Initially, the state \(\hat{S}(t)\) and measurement \(m(t)\) model is specified to estimate (predict) the next position. The model matrices are defined as:

\[\hat{S}(t) = A\hat{S}(t - 1) + g(t)\]  \hspace{1cm} (1)

\[m(t) = H(t)\hat{S}(t) + v(t)\]  \hspace{1cm} (2)

where \(A\) and \(H(t)\) refer to state transition matrix and measurement matrix respectively. The white Gaussian noise \(g(t)\) and \(v(t)\) with zero mean may produce in the model due to unconstrained measurement. Therefore, the covariance matrix \(Q\) and \(R\) are estimated using \(g(t)\) and \(v(t)\) in conjunction with korneckor delta function to define noise at certain instant.

The prediction of the next state \(\hat{S}^t(t)\) is accomplished by incorporating the actual measurement with prior estimate of state \(\hat{S}^t(t)\).

\[\hat{S}^t(t) = \hat{S}^t(t) + K(t)(m(t) - H(t)\hat{S}^t(t))\]  \hspace{1cm} (3)
The variable \( K(t) \) refers to the Kalman gain and is written as:
\[
K^+(t) = \hat{P}^-(t)H(t)^T(H(t)\hat{P}^-(t)H(t)^T + R(t))^{-1}
\]

(4)

The Kalman gain includes a prior error covariance matrix \( \hat{P}^-(t) \) that is derived using covariance of \( \hat{S}^- = \hat{S}(t) - \hat{S}(t). \) In a similar manner a posterior error covariance matrix \( \hat{P}^+(t) \) is derived using the covariance of \( \hat{S}^+ = \hat{S}(t) - \hat{S}(t). \) This \( \hat{P}^-(t) \) along with \( K^+(t) \) and current input is utilized to find the next state and process is repeated recursively.

For next measurement, a prior state and covariance error is measured recursively as follows.
\[
\hat{S}^- = A\hat{S}^-(t-1)
\]
\[
\hat{P}^- = A\hat{P}^-(t-1)A^T + Q(t-1)
\]

(5)
(6)

The main goal is to estimate correct state using Eq. (3) by correcting the Kalman gain through Eq. (4). As seen, more would be the Kalman gain lesser would be the measurement error. The final step is to obtain a posterior error covariance matrix using Eq. (7). The previous posterior estimate is used to create a new prior estimate to improve the measurement.
\[
\hat{P}^+(t) = (I - K(t)H(t))\hat{P}^-(t)
\]

(7)

It is assumed that object covers equal distance in every interval, which is represented as:
\[
f(t) = f(t-1) + f(t-1) - f(t-2)
\]

(8)

where \( f(t-1) \) and \( f(t-2) \) is the distance covered by object in frame \( t-1 \) and \( t-2 \), respectively.

The state matrix \( \hat{S}(t) \) and \( \hat{S}(t-1) \) of Eq. (1) can be derived using the given equation.
\[
\hat{S}(t) = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} f(t) \\ f(t-1) \end{bmatrix} + \begin{bmatrix} g(t) \\ 0 \end{bmatrix}
\]

(9)

Experimental analysis

This section presents the qualitative and quantitative analysis on different video sequences using the proposed algorithm and its performance parameters are compared with other existing methods. The ‘IR’, ‘WS’ and ‘Office’ experimental sequences depict great diversity in its successive frames due to the illumination variation, local motion in background, aperture problem. The ‘WS’ sequences has camouflage region below the knee of person and depict the local motion in the background due to rippling water. The moving object changes its dimensionality in the successive frame of ‘IR’, while the ‘Office’ sequences depict the slow and stationary behaviour of the object for a long time.

Figs. 1 and 2 show the qualitative performance of this proposed method. The first row of Figs. 1 and 2 shows the sampled frames with tracking results, while the last row shows the motion mask obtained through this method. As expected, the proposed method performs better classification of foreground and background under both static and dynamic background conditions. The qualitative comparison on some sampled frames shown in Fig. 3 illustrates the efficacy of this proposed method in complex situations.

The quantitative results are evaluated through Similarity, \( F1 \) and Detection rate metrics (Rahman et al., 2013). These evaluation metrics depend on \( tp \) (true positive), \( tn \) (true negative), \( fp \) (false positive) and \( fn \) (false negative). The ‘tp’ and ‘tn’ are correctly detected foreground and background pixels respectively.

The ‘fp’ and ‘fn’ are the incorrectly detected foreground and background pixels respectively. The parameters Detection Rate, Similarity and \( F1 \) are given as:
\[
\text{Detection Rate} = \frac{tp}{tp + fn}
\]

(10)
\[
\text{Similarity} = \frac{tp}{tp + fp + fn}
\]

(11)
\[
F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

(12)

where Recall and Precision are the relevant and irrelevant true positive pixels respectively. Fig. 4 presents the detection rate on sampled frames of each video sequence through this method along with other existing methods. Here, it is seen that initially some methods such as GMM (Stauffer and Grimson, 2000) and method (Rahman et al., 2013) have good detection rate, but their performance is degraded on subsequent frames due to object either becomes stationary or suffers from ghost on foreground. However, the detection rate through our method is superior to other methods for each video sequence. The average detection rate obtained through this method is up to 40% and 35% greater than GMM and method (Rahman et al., 2013) respectively for
Video object extraction and its tracking

Figure 3 Visual comparison on foreground motion masks.

Figure 4 Detection rate of (a) WS, (b) Office and (c) IR video.

Table 1 Quantitative performance comparison on foreground motion mask.

<table>
<thead>
<tr>
<th>Sequences</th>
<th>Evaluation</th>
<th>Proposed method</th>
<th>Method (Rahman et al., 2013)</th>
<th>GMM</th>
<th>FD</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS</td>
<td>Similarity</td>
<td>0.8745</td>
<td>0.7000</td>
<td>0.3476</td>
<td>0.1542</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>0.9325</td>
<td>0.8444</td>
<td>0.5112</td>
<td>0.2677</td>
</tr>
<tr>
<td>IR</td>
<td>Similarity</td>
<td>0.7791</td>
<td>0.4035</td>
<td>0.5116</td>
<td>0.2064</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>0.8983</td>
<td>0.5750</td>
<td>0.6769</td>
<td>0.3132</td>
</tr>
<tr>
<td>Office</td>
<td>Similarity</td>
<td>0.8189</td>
<td>0.4864</td>
<td>0.2729</td>
<td>0.2451</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>0.9129</td>
<td>0.6545</td>
<td>0.4275</td>
<td>0.3938</td>
</tr>
</tbody>
</table>

all videos. The detection rate obtained through this proposed scheme is also better than traditional FD (Nikolov and Kostov, 2014) as it does not create holes inside moving entity. Table 1 lists the average Similarity and F1 score of the moving mask extracted by method (Rahman et al., 2013). GMM, FD and proposed approach for the each video sequence. Here, it is seen that the average Similarity and F1 score secured through this method are up to 35% and 33% greater than those attained by GMM for ‘WS’ video. Moreover, it significantly outperforms the FD method in providing a promising motion mask on foreground. As seen, only this method attains higher similarity and F1 score exceeding 80% for the Office sequence in which the object becomes stationary for long time.
Conclusion

The proposed algorithm localizes the entire target region with high similarity and accuracy. The system effectively combines the spatio-temporal processing with regional level operation in order to reduce the background clutter. The proposed scheme is found to be effective to eliminate ghost and aperture distortion. It provides the sufficient sample size of an object to Kalman filter in order to predict the target position in the successive frames. Simulation results prove that the method can be implemented in better localization of moving or stationary object than other background subtraction schemes used for tracking.

References


