Video summarization is the process to extract the most significant contents of a video and to represent it in a concise form so that a user can understand about all the important events of a long video. Existing methods for video summarization fail to achieve a satisfactory result for a video with camera movement and significant illumination changes. To solve this problem, a new saliency prediction method is proposed based on human visual field using human eye’s fixation data obtained by Tobii X120 Eye Tracker. Three different circular regions are considered around a fixation point similar to foveal, parafoveal and peripheral regions of human visual field. The inner (foveal), middle (parafoveal), and outer (peripheral) regions are assigned highest, mid and lowest salient values respectively. The motivation is that human pay more attention in foveal region and less attention in parifoveal region. Based on this concept, a visual saliency score is calculated from eye tracker fixation data for each frame and a set of key-frames are selected based on used preferences. The proposed method is implemented on Office video dataset that contains video with camera movement and illumination change. Experimental results show superior performance compared to the existing GMM based method.

Index Terms— Human visual field, eye tracker, fixation point, saliency map, video summary

1. INTRODUCTION

With the increased demand for providing security, controlling crowd, and monitoring traffic, video surveillance cameras are usually setup in public places, like shopping malls, service stations, banks, universities, business organizations and road junctions. These surveillance cameras capture huge amount of videos on daily basis. To find out suspicious/informative events, captured huge video contents are analyzed. This process is very tedious and boring if it is done manually. Therefore, an automatic system is required to detect important events in a video. This system is called video summarization in computer vision field. The main purpose of video summarization method is to represent the long video in a shorter form containing all informative contents and neglecting unnecessary frames so that a user can get whole concept about the important events within the video.

To retrieve all the important events within a video, contents of the video are extracted. However, in case of changing in lighting condition and camera movement, it is not always possible to extract video contents correctly [1][2][3]. When changes in illumination occur, each pixel of an image experiences variation in intensity/color values. As a result, the existing techniques, which use intensity/color based features, fail to detect this incident successfully as they detect abrupt changes in motion or foreground object areas. In case of camera movement, a global motion happens in a video [1] and affects dynamic saliency significantly [4]. Consequently, some unnecessary video frames can be selected from video with camera motion. Therefore, to summarize video containing these phenomena is still a challenging problem.

To solve these problems, an innovative framework for generating visual saliency map based on human visual field using eye tracker data for summarization has been proposed in this paper. According to the findings of psychophysical experiments, eye can easily perceive objects in case of camera movement and/or illumination changes [5]. This motivates to implement human fixation data obtained by Tobii eye Tracker extract informative contents of a video containing illumination change and/or camera movement for generating summarization. According the structure of a human retina, visual field of human eye has three different regions around the fixation point, namely foveal, parafoveal, and peripheral [6]. Human eyes pay high, mid and low level concentration in foveal, parafoveal and periphery regions respectively. Based on this model, a saliency map is generated around the fixation point obtained by Tobii X120 Eye Tracker machine. Three circles with different radii are drawn centering the fixation point. The inner, mid and outer circles are assigned high, mid and low values similar to foveal, parafoveal and periphery regions of human visual field. The saliency maps obtained by several observers are
multiplied and normalized for each frame. A saliency score is calculated for each frame by summing all the saliency values within a frame. Later, saliency scores for all frames of a video are sorted and a summary of video is generated from the sorted saliency score based on user preferences. Therefore, the main contributions of the paper are to apply human visual field for saliency map generation and use it for summarizing video containing changes in illumination and/or camera movement.

The remaining structure of the paper is organized as follows. Section 2 describes the related work proposed in the literature on video summarization methods in recent years; the detail of the proposed method is discussed in Section 3. Extensive experimental results as well as analytical discussion are provided in Section 4 and a concluding remark is presented in Section 5.

2. RELATED RESEARCH

In the literature, a number of different approaches have been proposed to summarize video. In [7], audio, visual, and linguistic/textual cues are applied. Audio saliency is obtained by multi-frequency waveform modulations; visual cue is extracted using intensity, color, and orientation and textual cue is gained by subtitles information. These multimodal cues are combined hierarchically to generate saliency curve and video summarization. In [8], salient regions in each video are obtained by video semantics, an active learning method is applied to select candidate key frames and a probabilistic model is introduced for final key frame selection. For summarizing videos, shot boundary is detected in [9] and a set of candidate key frames are selected. Then, region based contrast method is applied for saliency map generation and a set of thresholds is applied for removing non-informative frames. Finally, an online clustering method is used to discard frames with similar contents. In [10], a salient motion detection is proposed for ranking of visual data. This salient motion is obtained by integral-image based temporal gradients. A high-level saliency map is generated based on interaction, frequency, motion and likelihood of human faces in [11]. Then, a regressor is learnt to predict the key frames based on this saliency score. In [12], Deep Event Network (DevNet) is introduced for high level events detection and spatial-temporal important evidence localization in a user generated video. Category specific (e.g. birthday party) user video is summarized in [13] by automatic temporal segmenting of video, scoring each segment applying support vector machine (SVM) and selecting higher segments. For detecting important events from user generated video, low level and semantic level features of visual and audio are applied in [14]. Gaze tracking information is applied in [15] for egocentric video summarization using sub-modular function maximization. Film comic is generated using eye tracking data in [16].

However, the existing approaches did not consider human visual field for saliency map generation. Furthermore, these techniques did not explain their robustness in case of illumination change and camera movement. Therefore, an innovative framework is proposed in this paper to generate summary from video with illumination changes and movement of camera using the saliency map obtained by the eye tracker data based on human visual field.

3. PROPOSED METHOD

The proposed method is based on a new saliency prediction method similar to human visual field. The main steps of the proposed method are (a) candidate key-frame selection based on image features, (b) saliency map generation, and (c) video summary generation based on user preferences. The block diagram of the proposed method is shown in Fig. 1. The detail of each section is described in the following sub-sections.

![Fig. 1. The structure of the proposed method](image)

3.1. Candidate Key-Frame Selection

First step of the proposed method is to select a set of candidate key frames from all the frames within a video using image base features. For this purpose, two important human visual sensitive features- area of foreground and motion information have been applied. Foreground objects are the most informative parts in a video stream as they contain more detail information and play a major role for important events [17]. Furthermore, human being usually pays more attention to the moving objects in a video [18] to understand an event. To obtain the area of foreground in a frame, Gaussian mixture-based Dynamic Background Modeling (DBM) [1] [19] is applied. In the proposed method, each coloured video frame is converted into gray scale image I(t) and DBM [1][19] is applied to obtain its corresponding gray scale background frame B(t). To obtain the foreground objects in a frame, the difference between I(t) and B(t) is calculated and a background-foreground separating threshold (δ1) is applied. In this way, a foreground pixel Gi,j(t) is obtained as follows

\[
G_{i,j}(t) = \begin{cases} 
1, & \text{if } |I_{i,j}(t)-B_{i,j}(t)| \geq \delta_1 \\
0, & \text{Otherwise}
\end{cases}
\]  

(1)

where (i, j) is the pixel position. If the pixel difference is more than 20 between background and foreground in the 0–255 scale, the foreground might be visible for human visual system [1]. Therefore, the value of δ1 is set to 20. After that, the total number of non-zero pixels in Gi,j(t) is used as the...
area of foreground object feature $\Gamma(t)$ which is obtained by the following equation

$$\Gamma(t) = \sum_{i=1}^{r} \sum_{j=1}^{c} G_{ij}(t)$$  \hspace{1cm} (2)

where $r$ and $c$ represent row and column of $G$ respectively.

In order to obtain object motion information, consecutive frame difference is computed by considering two consecutive frames $I(t-1)$ and $I(t)$ at time $t-1$ and $t$ in video respectively. To find out spatial motion information, the color difference in red, green and blue channel between these frames is calculated. A pixel is considered as motion pixel and set to value ‘1’, if the differences in three different channels at that pixel are greater than or equal to a threshold ($\delta_2$). Otherwise, it is considered that this pixel does not contain any motion information and set to value ‘0’. Therefore, the motion information $S_{ij}(t)$ in pixel $(i,j)$ at time $t$ can be obtained by the following equation

$$S_{ij}(t) = \begin{cases} 
1, & \text{if } |I_{ij}^r(t) - I_{ij}^r(t-1)| \geq \delta_2 \\
\text{and } |I_{ij}^g(t) - I_{ij}^g(t-1)| \geq \delta_2 \\
\text{and } |I_{ij}^b(t) - I_{ij}^b(t-1)| \geq \delta_2 \\
0, & \text{otherwise}
\end{cases}$$  \hspace{1cm} (3)

where $I_{ij}^r$, $I_{ij}^g$, and $I_{ij}^b$ represent red, green and blue color at $(i,j)$ pixel respectively. The value of $\delta_2$ is set to 20. Because, this threshold value may make human visual system to understand the difference between motion and non-motion pixel [1]. Therefore, the motion information $Y(t)$ is obtained at time $t$ by summing all values in $S_{ij}(t)$ as follows

$$Y(t) = \sum_{i=1}^{r} \sum_{j=1}^{c} S_{ij}(t)$$  \hspace{1cm} (4)

where $r$ and $c$ represent row and column of $S$ respectively.

After obtaining area of foreground $\Gamma(t)$ and motion information $Y(t)$ of a frame, a frame is considered as a candidate key-frame ($P(t)$) if either $\Gamma(t)$ or $Y(t)$ are greater than thresholds $\lambda_1$ and $\lambda_2$ respectively. These conditions is represented as follows

$$P(t) = \begin{cases} 
1, & \text{if } \Gamma(t) > \lambda_1 \text{ or } Y(t) > \lambda_2 \\
0, & \text{otherwise}
\end{cases}$$  \hspace{1cm} (5)

In Fig. 2, candidate keyframes obtained by the conditions of Equation (5) is shown. The light blue and black lines represent area of foreground and motion information of office-1 video respectively. The red and green lines indicate threshold values for area of foreground and motion information respectively. The frames above these thresholds are selected as candidate keyframes.

Fig. 2. An example of candidate key-frames selected from office-1 video.

### 3.2. Saliency Map Generation

In this step, visual saliency map is generated in each candidate key-frame obtained by the previous step. To predict saliency, human visual cognitive system is applied. The motivation is that human eyes can detect objects in case of illumination change and/or movement to/from the objects [5]. To locate the interesting/suspicious objects or events, human fixation point ($f_p$) are extracted using an eye tracker machine. Human fixation is the point in the visual field that is fixated by the two eyes in normal vision and for each eye is the point that directly stimulates the fovea of the retina. For generating a saliency map around $f_p$, the concept of human visual field is applied (please see Fig. 3). Based on the structure of human retina, visual field of human eye has three different regions around $f_p$, namely foveal, parafoveal, and peripheral [6]. Human eyes pay high, mid and low level concentration in foveal, parafoveal and periphery regions respectively. Based on this model, a saliency map is generated around $f_p$. Three circles ($C_1, C_2$ and $C_3$) with three different radii $R_1, R_2$, and $R_3$ are drawn centering $f_p$. The inner ($C_1$), mid ($C_2$), and outer ($C_3$) circles are assigned high ($V_1$), mid ($V_2$), and low ($V_3$) values similar to foveal, parafoveal, and periphery regions of human visual field. The saliency maps obtained by several observers are multiplied to highlight the overlap area in a video frame. For example, in Fig. 4 observers are interested to watch moving object in a video frame (Fig. 4(b)). On the other hand, their fixations are scattered when there is no object in a video frame (Fig. 4(d)).

Fig. 3. An example of human visual field. The inner, mid, and outer circles represent foveal, parafoveal, and peripheral regions respectively. Human pays highest, mid, and lowest attention in foveal, parafoveal, and peripheral regions respectively.

Fig. 4. An illustration of human fixations during watching a video: (a) and (c) are video frame no 245 and 10495 of office-1 video, (b) and (d)
and (d) are fixation points obtained by Tobii X120 eye tracker from four observers.

The multiplied saliency map for each frame is then normalized by the root of the number of observers. A saliency score \( K(\tau) \) is calculated for each frame by summing all the saliency values within a frame. The saliency scores of a video is smoothed by applying Savitzky-Golay filtering \([20]\) with window size 350. The main advantage of this filtering is that it enhances local maxima \([20]\). In Fig. 5, saliency score and smooth saliency of candidate keyframes office-1 video is shown by light aqua color and red color respectively.

![Saliency score generation using human visual field based saliency map of office-1 video. The light aqua, red, and black lines represent saliency score of candidate frames, smooth saliency score by Savitzky-Golay filtering [20], and ground truth.](image)

**Fig. 5.** Saliency score generation using human visual field based saliency map of office-1 video. The light aqua, red, and black lines represent saliency score of candidate frames, smooth saliency score by Savitzky-Golay filtering [20], and ground truth.

### 3.3. Video Summary Generation

In the final step, the saliency score \( K(\tau) \) for each frame within a video is sorted in descending order. The introduced approach offers the user to select the number of key frames \( \Phi \) in the summarized video. It generates summarized video based on the frame number \( \Phi \) selected by the user. From the sorted saliency scores \( K \); \( \Phi \) number of frames is selected from the top. Finally, summarized video is produced from these selected frames keeping their sequential order in the original video.

### 4. RESULTS AND DISCUSSION

The proposed method is evaluated by the publicly available Office [3] dataset. In Office dataset [3], four videos are collected from stably held with non-fixed cameras. The main difficulties are the vibration of camera and different lighting conditions. The ground truth key frames for Office dataset are also publicly available.

In this experiment, eye fixation points are extracted from four subjects for each videos of Office dataset by Tobii X120 eye tracker machine. These four subjects did not have any idea of the video. The summarization threshold \( \Phi \) is set to the total number of ground truth key frames for each video. The values of radii \( R_1 R_2 \) and \( R_3 \) are set to 20, 30, and 50 pixels respectively. The saliency values of \( V_1, V_2, V_3 \) are set to \( 3^3(= 27), 3^2(= 9), 3^1(= 3) \) respectively.

To evaluate the proposed method, an object comparison has been performed. For this purpose, a set of evaluation metrics including precision, recall and \( F_1 \)-measure are computed. The definition of precision and recall are as follows:

\[
\text{Precision} = \frac{T_p}{(T_p + F_p)}
\]

\[
\text{Recall} = \frac{T_p}{(T_p + F_n)}
\]

where \( T_p, F_p \) and \( F_n \) are the number of frames selected by a method and the ground truth, the number of frames selected by a method but not by the ground truth, the number of frames selected by the ground truth but not a method respectively. However, either precision or recall alone cannot provide a good indication of perfect measurement. For example, a method can offer better precision but poor recall or vice-versa. To be an efficient and robust method, it must achieve both higher precision and recall. To represent this measure, \( F_1 \)-measure is defined combining precision, recall, and represented as follows:

\[
F_1 - \text{measure} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}
\]

A high value for both precision and recall indicate that \( F_1 \)-measure is also very high. Thus, a method with high \( F_1 \)-measure value confirms better summarization technique.

The proposed approach is compared with the single-view video summarization results provided by GMM based method [17]. The GMM method is the most relevant and the state-of-the-art method to summarize video. This method is implemented on Office dataset [3] where camera vibration and illumination change are the main problems. Another reason is that this method does not use any human cognitive (visual) system. Therefore, to evaluate the robustness of the proposed method, it is compared with GMM based method.

<table>
<thead>
<tr>
<th>Video</th>
<th>The Proposed Method</th>
<th>GMM based Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Office-0</td>
<td>85.3</td>
<td>85.2</td>
</tr>
<tr>
<td>Office-1</td>
<td>78.6</td>
<td>78.6</td>
</tr>
<tr>
<td>Office-2</td>
<td>69.3</td>
<td>69.2</td>
</tr>
<tr>
<td>Office-3</td>
<td>73.5</td>
<td>88.1</td>
</tr>
<tr>
<td>Mean</td>
<td>76.7</td>
<td>80.3</td>
</tr>
<tr>
<td>Std</td>
<td>6.9</td>
<td>8.4</td>
</tr>
</tbody>
</table>

The results of precision, recall and \( F_1 \)-measure of both GMM based method and the proposed method are shown in Table I. It is observed from Table I that the mean of precision, recall, and \( F_1 \)-measure of GMM based method are 38.8, 86.9, and 53.0 while the proposed method obtains 76.7, 80.3, and 78.3. Standard deviations of precision, recall, and \( F_1 \)-measure of GMM based method for Office dataset are 7.3, 7.9, and
6.5. On the other hand, the proposed method attains 6.9, 8.4, and 6.7 respectively. This indicates that the proposed method not only performs in higher accuracy, but also the variance of the performance is more consistent in the different videos of Office dataset compared to GMM based method [17].

The results of F1-measure of the proposed method, the proposed method with only candidate key frames, and GMM based approach [17] are shown in Fig. 6. From this graph, it is observed that the proposed method performs superior all videos of Office dataset to the recently proposed the-state-of-the-art GMM based approach [17] and the proposed method with candidate key frames.

In Fig. 7, a number of frames of office-2 video of Office dataset [3] and the results obtained by GMM based method as well as the proposed method are shown. Both methods can successfully detect keyframe no 517. However, GMM based method [17] selects frame no 428, 5620, and 14117 as key frame wrongly. These frames are selected by the proposed method as candidate key-frames due to illumination changes and camera movement. However, the proposed saliency prediction method ignores these frames. The reason is that these frames do not contain any interesting objects or events (see the last column of Fig. 7).

The main reasons of the superior performance of the proposed method are that it considers two important human visual sensitive features (Area of foreground and motion information) for candidate key frame selection, and applies human visual cognitive system for selecting important contents of video containing illumination change and camera movement. According to the findings of psychophysical experiments, human eye can easily perceive objects in case of camera movement and/or illumination changes [5]. Moreover, to obtain saliency map of a video frame, the concept of human visual field is applied. Therefore, the proposed method performs better in case of illumination change and camera movement.

5. CONCLUSION

<table>
<thead>
<tr>
<th>F. No</th>
<th>Ground Truth (Key Frame)</th>
<th>GMM Key Frame by The Proposed Method</th>
<th>Candidate Key Frame</th>
<th>The proposed Saliency Map</th>
</tr>
</thead>
<tbody>
<tr>
<td>428</td>
<td>Not Key Frame</td>
<td>Not Selected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>517</td>
<td>Not Key Frame</td>
<td>Not Selected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5620</td>
<td>Not Key Frame</td>
<td>Not Selected</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14117</td>
<td>Not Key Frame</td>
<td>Not Selected</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In this paper, an effective and robust framework is proposed for generating visual saliency map based on human visual field using eye tracker data for summarization. According to the findings of psychophysical experiments, human eye can easily perceive objects in case of camera movement and/or illumination changes [5]. This motivates us to implement human fixation data obtained by Tobii eye tracker to extract informative contents of a video containing illumination change and/or camera movement for generating summarization. According the structure of a human retina, human visual field has three different visual regions around the fixation point, namely foveal, parafoveal, and peripheral [6]. Human eyes pay high, mid, and low level concentration in foveal, parafoveal and periphery regions respectively.

Based on this model, a saliency map is generated around the fixation point obtained by Tobii X120 Eye Tracker machine. Three circles with different radii are drawn centering the fixation point. The inner, mid, and outer circles are assigned high, mid and low values similar to foveal, parafoveal, and periphery regions of human visual field. The experimental results confirm that the proposed method provides 25.3 better F1 measure on average compared to the existing GMM-based technique[17].

6. REFERENCES


