Advancement in Video Summarization using Domain Motion Information with Fusion of Foreground Object

Aseena Shaik Babu
Assoc. Prof.
Department of CSE
Siddhartha Institute of Technology and Sciences, Narapally, Hyderabad, Telangana, India

Dr.Akella Satyanarayana
Professor
Department of CSE
Siddhartha Institute of Technology and Sciences, Narapally, Hyderabad, Telangana, India

Abstract **- Every day, a surveillance video camera gathers a large amount of consistent video stream. It is a difficult and tiring task to study or research any significant events from massive video data. This research presents a video summarizing technique that combines foreground objects and movement data in the spatial and frequency domain to overcome this problem. We use foundation demonstrating and movement data in both the spatial and frequency domains to eliminate foreground objects. The acquisition of movement data in the spatial domain is linked to frame transition. The phase correlation (PC) approach is used to generate frequency domain movement data. Then, in the spatial and frequency domains, foreground objects and movements are integrated, and important frames are isolated. According to the findings of the experiments, the proposed strategy outperforms the technique.**

Key Terms: Surveillance video, video summarizing, spatial and frequency domains, foreground objects

1. Introduction

Video summarizing (VS) is a technique for selecting the most interesting frames from a video in order to include all of the important events while excluding extraneous stuff, resulting in a summarized video that is as compact as possible. In this way, a good video summarized approach is one that possesses a few key characteristics. To begin with, it must be capable of incorporating happenings from the first video. Second, it should be able to make a smaller version of the rendered long video. Third, it should not include tiresome information. The primary objective behind VS is to depict a long unique video in a consolidated version so that a viewer can receive the complete notion of the video in a constrained amount of time. In our daily lives, a massive amount of surveillance footage is captured 24 hours a day, all over the world, for the purposes of providing security, monitoring, preventing crime, and controlling traffic, among other things. Various surveillance video cameras are typically installed in various distinct locations within a building, business, or congested zone. For storage and investigation, these cameras are linked to a monitoring cell. To store this massive amount of video data, you'll need a lot of memory. Administrators, on the other hand, must access the saved films in order to find any important events for reviewing or conducting investigations. This method is incredibly sluggish, time-consuming, and expensive. To address these challenges, an approach for generating a reduced version of the original movie that includes essential events is ideal for memory management and data recovery.

Foreground objects in a video typically have greater detail data [1]. Humans, once again, are prone to focusing on the progression of items [2]. As a result, objects, as well as their movement, are essential components of a movie. In this study, a video summarizing technique based on objects and their movement in a film is offered, fueled by these results. Gaussian mixture-based parametric background modelling (BGM) [3] was used to incorporate foreground object data. Object movement is removed not only in the spatial domain but also in the frequency domain to adopt the whole data of item movement in a movie. The technique of sequential frame contrast is

used to obtain movement data in the spatial domain. The phase correlation process [4] is required to achieve object movement in frequency. Phase correlation is not required for video summarizing algorithms, as far as we know. As a result, the main goal of this paper is to use phase correlation in the video summarizing technique. The phase correlation system [4] has a short calculation time and collects a large amount of movement data.

2. RELATED WORK

Various approaches for summarizing various types of films have been proposed in the literature. In [5], location saliency is predicted using a regression model for egocentric video summarization, and a storyboard is constructed based on the region significance score. The technique suggested in [6] summarizes story-driven egocentric video by locating the most influential things within the video. For summarization, gaze tracking data is used in [7]. In the event that a client-produced video summary occurs, the adaptive sub modular maximization function is used in [8]. [9] uses a collaborative sparse coding methodology to create a summary of the same type of video. Web images are used in [10] to improve the process of summarizing the user-generated video. To summarize the film, the audio, visual, and linguistic elements are all combined in [11]. In [12], a role community network is used. In [13], eye tracking data is used to produce a film comic. Despite this, devices for wireless capsule endoscopic video summarizing were proposed in [14] [15] [16] [17].

However, the importance of surveillance video for industrial applications is far greater than that of other types of recordings (e.g., egocentric, user created, motion picture, and so forth). [18] uses an object focused technique to compress surveillance video. [19] proposes a Dynamic Video Book for showing surveillance video in a hierarchical way. [20] presents a learned separation metric for summarizing nursery school surveillance video. The salient motion data is linked in [21]. For the production of synopses, [22] use maximum a posteriori probability (MAP). A technique for multi-view surveillance video summarizing is now proposed in [1]. To begin, this process creates a single view summarization for each sensor on its own. As a result, every video frame must have an MPEG-7 color format descriptor, and clustering is done using an Online-Gaussian mixture model (GMM). The key frames are selected based on the cluster's specifications. A video segment is extracted rather than key frames since the decision to select or discard a frame is dependent on the consistent updates of these clustering parameters. Finally, multi-view summarization is achieved using a distributed view selection technique that makes use of the video segments that were deleted for each sensor in the previous stage. To our knowledge, the phase correlation method has never been used for video summarization. In order to incorporate movement data in the frequency domain and fuse it with moving foreground objects and spatial movement data in this suggested strategy, a phase correlation approach is required.

3. PROPOSED WORK

The proposed technique is based on the spatial and frequency domain movement data of moving foreground objects. (1) moving foreground object extraction (2) movement data count in spatial domain (3) movement approximation in frequency domain (4) combination of foreground object range and spatial and frequency movement data (5) video summary generation are the major steps of the suggested technique.

3.1. Foreground Object Extraction

Gaussian mixture-based parametric BGM [3] is used in the proposed approach. The K Gaussian distributions (K=3) display every pixel in this parametric BGM, and each Gaussian model addresses either a static background or a dynamic foreground item on a time frame. Assume that a pixel intensity x_t at time t is represented by k^{th} Gaussian with recent measure γ_k^t, mean μ_k^t, standard deviation σ_k^t and weight ω_k^t such that $\sum \omega_k^t = 1$. The learning parameter is used to rework parameter measures such as mean, standard deviation, and so on. The framework has an unfilled arrangement of Gaussian models at the start. Following the discovery of the principal pixel (t=1), a Gaussian model (K=1) with $\gamma_k^t = \mu_k^t = x_t$, standard deviation $\sigma_k^t = 30$ and weight $\omega_k^t = 0.001$.It then attempts to locate a coordinated model from the present models

such that $|x_t - \mu_k| \leq 2.5\sigma_k$ for each additional notification of pixel intensity of the same area at t.

Background modelling [3] is used after turning each color frame into a grey scale picture to provide a grey scale background frame. A color video frame captured at time t is converted into a grey picture I(t) and subtracted from the corresponding grey background frame B(t) acquired through background modelling. If the pixel intensity difference between I(t) and B(t) is greater than or equal to a given threshold, a pixel is designated as a foreground pixel and the measure is set to one (Thr1). If the pixel intensity does not meet this requirement, it is considered a background pixel and set to zero. As a result, a foreground area pixel is obtained as follows:

$$G_{i,j}(t) = \begin{cases} 1 & if\ |I_{i,j}(t) - B_{i,j}(t)| \geq Thr1 \\ 0 & otherwise \end{cases} \quad (1)$$

The pixel position is represented by (i, j). To avoid unnoticeable transitions between background and foreground, the estimation of Thr1 is adjusted to 20. As stated in [24], it is common practice to set the threshold limit to 20 to identify objects from the background. The total number of non-zero pixels in $G_{i,j}(t)$ is then used as the region of foreground object feature F(t), which is calculated using the equation below, where r and c denote the row and column of F, respectively.

$$G(t) = \sum_{i=1}^{r}\sum_{j=1}^{c} G_{i,j}(t) \quad (2)$$

According to the psychological theories of human consideration, the movement data is far larger than static consideration indications [2]. As a result, regardless of the foreground object, movement data is incorporated into the proposed approach.

3.2. Fusion of Foreground and Motion Information

The foreground object and movement data are linked with the purpose of selecting more precise frame sequences. A weighted linear fusion is used in this method to connect the features for locating each frame as shown in the sample video. Every feature is turned into z-score standardization using the mathematical equation below before using the fusion process.

$$Z(t) = (X(t) - \mu)/\sigma \quad (3)$$

Where feature esteem at time is t, mean is μ and standard deviation of the feature measures is σ. Z(t) is a standardized type of X-score (t). Z-score standardization is preferred in this technique because it produces substantial data about each information point and produces better outcomes in the proximity of exceptions than min-max based standardization [23]. Assemble the weighted linear fusion as follows:

$$A(t) = \phi_1 * Z_G(t) + \phi_2 * Z_S(t) + \phi_3 * Z_F(t) \quad (4)$$

where A(t) is the fusion measure, $Z_G(t)$, $Z_S(t)$ and $Z_F(t)$ are the z-score standardization of foreground feature (G(t), spatial movement feature (S(t), and frequency domain movement data (F(t)) at time t, respectively. Experimentally, it has been determined that setting the weight ϕ_1, ϕ_2 and ϕ_3 estimations to 15, 60, and 25 yields superior results for all movies in the BL-7F dataset. The decision to give movement features greater weight than the front zone is based on the fact that, according to psychological theories of human consideration, movement data is more significant than static contemplation inferences [2]. Following that, A(1,..., T) is sorted in descending order, with T denoting the total number of frames in a video.

4. EXPERIMENTAL RESULTS

The existing GMM algorithm's Color Difference Observation and Probability Assign are compared to the suggested Bayesian methods in Figure 1. Both algorithms' corresponding Probability Assign are calculated and plotted at each level of Color Difference Observation. When compared to other algorithms, the suggested Bayesian algorithm performs better in terms of enhanced probability assign. The suggested Bayesian algorithms and the existing GMM algorithm's Position Difference Observation and Probability Assign are contrasted in Figure 2. To get the matching Probability Assign of both techniques, the Position Difference Observation is separated into multiple levels. According to the graph, the proposed Bayesian method performs better with a higher probability assign

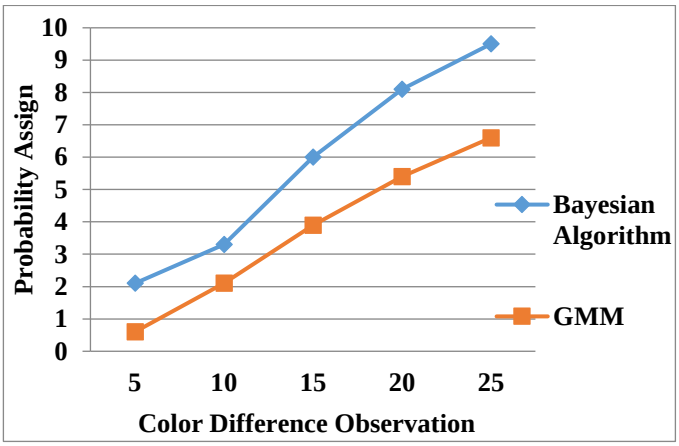

Figure 1: Color Difference Observation vs. Probability Assign

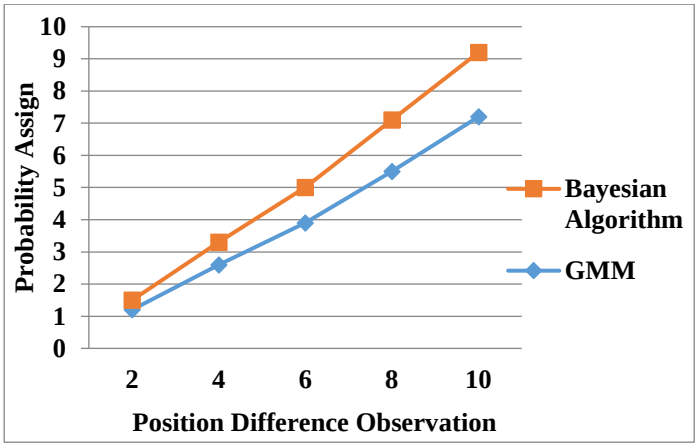

Figure 2: Position Difference Observation vs. Probability Assign

.5. CONCLUSION

As a result, we proposed a unique technique for merging foreground object data with movement data in the spatial and frequency domain to summarize surveillance footage. As shown in [1,] the foreground object contains precise information about the video's contents. In addition, in a video, a person really shows more consideration for object movement [2]. As a result, this methodology incorporates two crucial features of a video. The phase correlation approach [4] is used to incorporate movement data in the frequency domain. To our knowledge, this is the first time the phase correlation method has been used for video summary. The proposed technique outperforms the current state-of-the-art strategy, according to the findings of the experiments.

6.REFERENCES

1. Ou, S., LEE, C., Somayazulu, V., Chen, Y., Chien, S.: On-line Multi-view Video Summarization for Wireless Video Sensor Network. IEEE J. Sel. Top. Signal Process. 9, 165–179 (2015).
2. Gao, D., Mahadevan, V., Vasconcelos, N.: On the plausibility of the discriminant center-surround hypothesis for visual saliency. J. Vis. 8, 1–18 (2008).
3. Paul, M., Lin, W., Lau, C., Lee, B.: Explore and model better I-frames for video coding. IEEE Trans. Circuits Syst. Video Technol. 21, 1242–1254 (2011).
4. Paul, M., Lin, W., Lau, C.T., Lee, B.-S.: Direct intermode selection for H.264 video coding using phase correlation. IEEE Trans. image Process. 20, 461–73 (2011).
5. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering Important People and Objects for Egocentric Video Summarization. IEEE Conf. Comput. Vis. Pattern Recognit. 1346–1353 (2012).
6. Lu, Z., Grauman, K.: Story-Driven Summarization for Egocentric Video. IEEE Conf. Comput. Vis. Pattern Recognit. 2714–2721 (2013).
7. Xu, J., Mukherjee, L., Li, Y., Warner, J., Rehg, J.M., Singh, V.: Gaze-enabled Egocentric Video Summarization via Constrained Submodular Maximization. IEEE Conf. Comput. Vis. Pattern Recognit. 2235–2244 (2015).
8. Gygli, M., Grabner, H., Gool, L. Van: Video Summarization by Learning Submodular Mixtures of Objectives. IEEE Conf. Comput. Vis. Pattern Recognit. 3090–3098 (2015).
9. Liu, Y., Liu, H., Sun, F.: Outlier-attenuating summarization for user-generated-video. IEEE Int. Conf. Multimed. Expo. 1 – 6 (2014).
10. Khosla, A., Hamid, R.: Large-scale video summarization using web-image priors. IEEE Conf. Comput. Vis. Pattern Recognit. 2698 – 2705 (2013).
11. Evangelopoulos, G.: Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. IEEE Trans. Multimed. 15, 1553–1568 (2013).
12. Tsai, C., Kang, L.: Scene-Based Movie Summarization Via Role-Community Networks. IIEEE Trans. Circuits Syst. Video Technol. 23, 1927–1940 (2013).
13. Sawada, T., Toyoura, M., Mao, X.: Film Comic Generation with Eye Tracking. Adv. Multimed. Model. 467–478 (2013).
14. Schoeffmann, K., Del Fabro, M., Szkaliczki, T., Böszörmenyi, L., Keckstein, J.: Keyframe extraction in endoscopic video. Multimed. Tools Appl. (2014).
15. Spyrou, E., Diamantis, D., Iakovidis, D.K.: Panoramic Visual Summaries for Efficient Reading of Capsule

Endoscopy Videos. 2013 8th Int. Work. Semant. Soc. Media Adapt. Pers. 41–46 (2013).
16. Mehmood, I., Sajjad, M., Baik, S.W.: Video summarization based tele-endoscopy: a service to efficiently manage visual data generated during wireless capsule endoscopy procedure. J. Med. Syst. 38, 109 (2014).
17. Ismail, M. Ben: Endoscopy video summarization based on unsupervised learning and feature discrimination. Vis. Commun. Image Process. 1–6 (2013).
18. Fu, W., Wang, J., Zhao, C., Lu, H., Ma, S.: Object-centered narratives for video surveillance. IEEE Int. Conf. Image Process. 29–32 (2012).
19. Sun, L., Ai, H., Lao, S.: The dynamic VideoBook: A hierarchical summarization for surveillance video. IEEE Int. Conf. Image Process. 3963–3966 (2013).
20. Wang, Y., Kato, J.: A distance metric learning based summarization system for nursery school surveillance video. IEEE Int. Conf. Image Process. 37–40 (2012).
21. Mehmood, I., Sajjad, M., Ejaz, W., Wook, S.: Saliency-directed prioritization of visual data in wireless surveillance networks. Inf. Fusion. 24, 16–30 (2015).
22. Huang, C., Chung, P.J.: Maximum a Posteriori Probability Estimation for Online Surveillance Video Synopsis. IEEE Trans. Circuits Syst. Video Technol. 24, 1417–1429 (2014).
28. Han, J., Kamber, M., Pei, J.: Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan Kaufmann (2006).