Domain-Aware Video Summarization with Motion Information and Foreground Object Fusion

S. Alangarom[1], A. Vijay[2]
[1,2]*Department of IT, JEC, Chennai, Tamil Nadu, India*
[1]*alangarom123@gmail.com*

Abstract - *Surveillance cameras produce tremendous amounts of continuous video data every day, which is hard and time consuming to determine important events by hand. In this paper, the authors propose a "video summarization method based on foreground" object detection and motion information from spatial and frequency domains to solve this issue. The approach models the background and identifies motion clues to extract foreground objects successfully. "Motion in the spatial domain" is obtained from frame changes, whereas "frequency domain motion" is obtained through the Phase Correlation (PC) method. Through the combination of motion information and foreground object data in both domains, the system detects and retrieves key frames optimally representing the content of a video. Experimental results validate that the suggested method performs better than other approaches in effectively and accurately summarizing surveillance videos.*

Keywords - *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. A strong video summarization (VS) system has to meet three key requirements: first, it has to extract and integrate salient events of the original video; second, it has to produce a summarized version of a long video; and third, it has to omit redundant or irrelevant information. The primary aim of video summarization is to provide the core content of an extended and special video in an abbreviated form so that the viewers can understand the entire context in a short duration of time.

In daily life, a vast amount of surveillance video is recorded round the clock, 24/7, everywhere in the world for security monitoring, crime deterrence, and traffic management. Surveillance cameras tend to be strategically placed in several critical locations inside buildings, enterprises, or populous public places. Cameras usually pump their information to centralized monitoring systems for storage and processing. Yet, the massive amount of video data produced necessitates huge storage, which is both challenging to manage and difficult to review in a timely manner. 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 paper, introduce an object detection and motion analysis driven video summarization method motivated by the growing necessity for effective representation of video content. Foreground object data are used through the incorporation of parametric background modeling (BGM) based on Gaussian Mixture [3], which efficiently dissects moving objects from static environments. To efficiently capture complete object motion, movement "information is taken in both spatial and frequency domains" by our approach. "Motion in the spatial domain" is calculated via sequential frame differencing, whereas motion in the "frequency domain is calculated with the phase correlation" approach [4]. In contrast with other video summarization methods that hardly ever use phase correlation, the proposed approach incorporates it as a fundamental building block. The most significant contribution of this work lies in the new use of phase correlation for video summarization. This method provides low computational complexity with rich motion information capture, thus being very appropriate for real-time or large-scale surveillance video analysis.

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. Various methods have been put forward for video summarization, aiming at different domains and

data modalities. The approach in [6] summarizes narrative-driven egocentric videos by extracting the most significant objects within the video content. “Gaze tracking information” is utilized in [7] to control the process of summarization based on visual attention. For situations where user-created video summaries are present, an “adaptive submodular maximization function” is used in [8]. “Collaborative sparse coding” is proposed in [9] to summarize videos of the same kind efficiently. For better summarization of user-generated content, web images are used in [10]. Multimodal methods that integrate audio, visual, and linguistic features are investigated in [11] to provide more thorough video summaries. A role community network is employed in [12] to organize video summarization, whereas eye-tracking data is once more utilized in [13] to produce a film comic. Besides, certain methods have been suggested for summarization of wireless capsule endoscopy videos, such as in [14], [15], [16], and [17], where effective data abstraction is critical owing to the long and uninterrupted nature of medical videos. [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. 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.

3. Methodology

The suggested method is based on examining the foreground object motion in both frequency and spatial domains. It involves the following key steps: Moving Foreground Object Extraction: Foreground objects are detected and isolated by applying background modeling methods to extract dynamic scene elements. Calculation of Motion Data in the Spatial Domain: Sequential frame differencing is used to estimate motion based on inter-frame changes. Motion Approximation in the Frequency Domain: Frequency variation across frames is analyzed to extract motion information using phase correlation. Integration of “Foreground Object” Range with “Spatial and Frequency Domain Motion Features”: Features of motion in both domains are fused with the identified foreground areas to create an inclusive motion profile. Video Summary Generation: Utilizing the combined information about motion and objects, the most important content is represented through keyframes, which are used to create the final video summary.

3.1. Object Extraction

In the work presented here, Gaussian Mixture-based Background Modeling (BGM) [3] is used for foreground object separation. This technique represents a pixel with a mixture of K Gaussian distributions (K=3). Every Gaussian component accounts for either static background or dynamic foreground feature at any time. (K=1) is initialized with an initial mean, standard deviation, and weight. For every subsequent observation at the same pixel location, the system tries to match it with a previously stored Gaussian model. If it is found within a specified threshold, the parameters are updated; otherwise, a new Gaussian is inserted (up to I(t)−B(t) | ≥Thr1). If the difference is more than a fixed threshold Thr1, the pixel is labeled as a foreground pixel (given a value of 1); otherwise, it is labeled as a background pixel (given a value of 0). This produces a binary foreground mask that indicates the moving object regions in the frame.

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

3.2. Motion and Fusion of Foreground Information

To improve frame selection accuracy, the method in question combines foreground object information with motion data through a weighted linear fusion process. The fusion of spatial and frequency-based motion features with foreground regions detected enables a more accurate identification of key frames, as seen in the test video. Prior to using the fusion process, every feature is normalized to z-score level to make all features contribute proportionally irrespective of their original size.

$$Z(t) = (X(t) - \mu) / \sigma \quad (2)$$

4. Discussion

A comparative study between the current GMM algorithm and the new Bayesian approach in terms of Color Difference Observation and Probability Assignment. For both methods, the corresponding values of probability are calculated and graphed at each level of observed color difference as shown in Figure 1. The findings show that the suggested Bayesian algorithm

always gives higher probabilities at different levels of color difference, proving to be a more robust response in separating foreground objects. In comparison with the conventional GMM method, the Bayesian algorithm shows better performance in terms of probability assignment correctness, eventually leading to better foreground detection and video summarization quality.

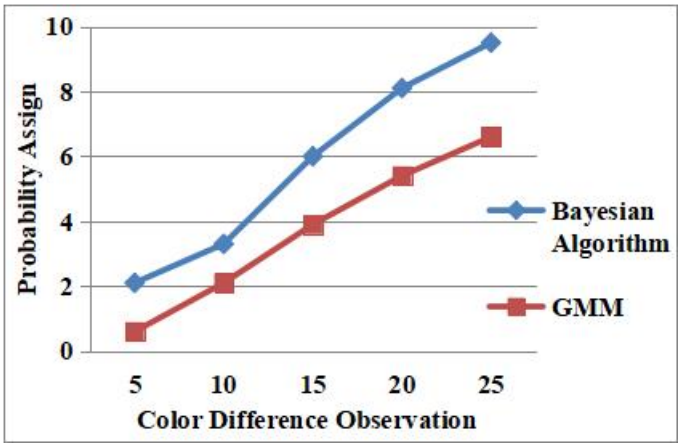

Fig. 1 Color Difference Observation vs. Probability Assign

Figure 2 shows a comparison between the new Bayesian algorithm and the current GMM algorithm, specifically on Position Difference Observation and the resultant Probability Assignment. The position difference is divided into several levels, and for each level, the probability values assigned by both methods are computed and graphed. The findings show that the suggested Bayesian method always yields higher probability assignments under different position differences, demonstrating better sensitivity to positional variation. This improved performance indicates that the Bayesian approach yields a more precise spatial variation, thus supporting more accurate foreground detection and efficient video summarization.

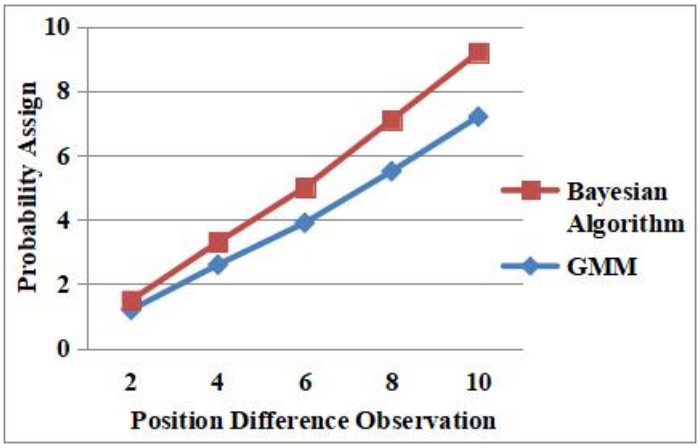

Fig. 2 Position Difference Observation vs. Probability Assign

5. Conclusion

In summary, we introduce a new method for summarizing surveillance videos through the integration of foreground object details with motion information derived from the spatial and frequency domains. As illustrated in [1], foreground objects hold extremely descriptive and relevant information regarding video content. Additionally, existing research shows that observers pay greater attention to object motion while making sense of video scenes [2]. Hence, this approach combines two essential elements of visual information—object presence and motion dynamics.

To extract frequency domain motion, the new solution uses the phase correlation technique, which to our knowledge for

the first time is applied to video summarization. Results of experiments validate that our method performs better than current state-of-the-art solutions, providing a more efficient and precise solution for summarizing long security videos.

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).

[23] Han, J., Kamber, M., Pei, J.: Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan Kaufmann (2006).