6-2015

# A Modular Approach for Key-Frame Selection in Wide Area Surveillance Video Analysis

Almabrok Essa
*University of Dayton*

Paheding Sidike
*University of Dayton*

Vijayan K. Asari
*University of Dayton*, vasari1@udayton.edu

# A Modular Approach for Key-Frame Selection in Wide Area Surveillance Video Analysis

Almabrok Essa, Paheding Sidike, and Vijayan Asari
Department of Electrical and Computer Engineering
University of Dayton, 300 College Park, Dayton, OH, USA 45469

***Abstract* - This paper presents an efficient preprocessing algorithm for big data analysis. Our proposed key-frame selection method utilizes the statistical differences among subsequent frames to automatically select only the frames that contain the desired contextual information and discard the rest of the insignificant frames. We anticipate that such key frame selection technique will have significant impact on wide area surveillance applications such as automatic object detection and recognition in aerial imagery. Three real-world datasets are used for evaluation and testing and the observed results are encouraging.**

***Keywords - content based approach; contextual information, statistical difference; key-frame selection***

## I. INTRODUCTION

An important step in content based video processing is key frame selection which is an essential part in video summarization in terms of speed and accuracy. Key frame is the frame which can represent salient contextual information of the video. The key frame selection techniques can be classified into three categories: energy minimization based methods, cluster based techniques and sequential processing based methods [1]. The energy minimization based methods [2] extract the key frames by solving an energy minimization problem. The clustering based approaches [3] take all the frames of a shot together and identify cluster centers as key frames. The disadvantages of these approaches are that they ignore the temporal information of a video sequence and they use iterative techniques to perform minimization which in general computationally expensive. The sequential processing based methods [4] consider a frame as a key frame when the content difference from the previous frame exceeds a predefined threshold that is determined by the user. Our proposed technique is a sequential processing based methods that is able to automatically select only the frames which contain desired contextual information and discard the rest which are the insignificant ones.

Analyzing all frames in an aerial surveillance video is not a meaningful process when some of the frames do not contain significant information. For example, video frames captured by aircrafts flying over hundreds of miles could be formidable and time costly for computer vision and image understanding algorithms to analyze this enormous amount of data. Therefore, the aim of our proposed method is to automatically select only the frames that contain important information from the big data so that the entire computation time could be reduced significantly. To achieve this, we introduce a modular key frame (MKF) selection strategy which consists of two functional stages: batch processing and sub-region processing as illustrated in Fig. 1. The first stage is to divide all video frames into $M$ batches where each batch contains $N$ frames. In the second stage, we first partition each frame into $m \times n$ sub-regions and calculate the statistical difference between each corresponding sub-regions in two consecutive frames.

The rest of the paper is organized as follows: Section II reviews the related work in key frame extraction. The proposed MKF selection technique is presented in section III. In section IV, the experimental results and analysis are provided. Conclusion is drawn in section V.

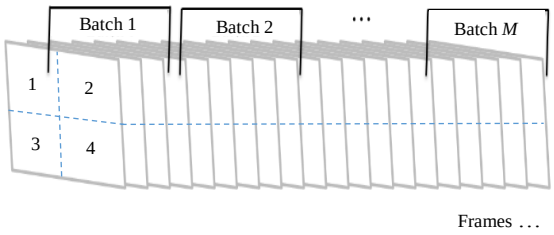


Fig. 1. Illustration of two functional stages of the proposed method, where 1, 2, 3, and 4 are the sub-regions of each frame.

## II. RELATED WORK

### A. Motion Analysis Based Approach

Wolf identified key frames based on motion analysis [5]. He computed a simple motion metric (local minima of motion) based on the optical flow of each frame and then selected the key frames at the local minima of motion in a shot-stillness that emphasizes the image for the viewer. The justification of this technique is to identify both gestures and camera motion. Gestures are considered where the characters emphasize their importance by holding gestures. Camera motion is considered when the camera stops on a new positon where the frame is important.

### B. Shot Activity Based Method

Gresle and Huang extracted the key frames based on a shot activity [6]. They computed the activity indictor by computing the intra- and reference histograms. Then based on the activity curve as well as Wolf's approach, they selected the key frames at the local minima. The disadvantage of the previous approach and this approach is that they are computationally expensive.

This project is funded by the Pipeline Research Council International under the contract number: PR-433-133700.

### C. Shot Boundary Based Technique

In this technique, Nagasaka and Tanaka [7] segmented the video into shots and then used the first frame of each shot as a key frame. Even though this approach is comparatively fast, its disadvantages are the number of key frames for each shot is limited to one and does not capture the major visual content of the shot and normally it is not stable.

## III. A Modular Approach for Key-Frame Selection

The aim of key frame selection is to best appear the prominent content and information of the video with the minimal number of frames. In our case, we propose a modular key-frame selection technique to reduce the computation time of analyzing huge amount of data in aerial surveillance imagery. Our proposed technique is able to automatically select only the frames that contain the desired contextual information and discard the rest of the frames that are insignificant. For the detection of key-frames we calculate a statistical difference between subsequent frames. The frames whose statistical differences exceed an adaptively computed threshold value are considered as key-frames.

The proposed modular key-frame selection framework consists of two functional stages: batch processing and sub-region processing. The first stage is to divide the video into $M$ batches of individual frames where each batch contains $N$ frames. In the second stage, we first partition each frame into $m \times n$ sub-regions and the statistical differences (i.e. differences in mean and standard deviation) between the corresponding sub-regions in two consecutive frames are calculated. After that we compute the local means of the batches, which are the means of the statistical differences of all the sub-regions of two successive frames. Finally the global mean and standard deviation are calculated by utilizing all the local means for each batch. During this process, an adaptive threshold is obtained using global mean with its corresponding standard deviation. A frame is considered as a key-frame if the statistical difference exceeds the adaptive threshold of the corresponding frames as described in Eq. (1). Fig. 3 shows the flowchart of the proposed key frame selection technique.

Given an input video, segment it into $M$ batches where each batch contains $N$ number of frames. Divide each frame to $m \times n$ number of regions as illustrated in Fig. 4. Then we calculate the statistical difference between each corresponding sub-regions in two consecutive frames using Eqs. (2) and (3). Finally the local means and standard deviations are calculated for each batch to select an adaptive threshold as shown in Eq. (4).

Let $I(x, y)$ is an original image where $x = 1, \dots, R;\ y = 1, \dots, C$, and R and C are the dimensions of the image. Then the selection of key-frame is defined as follows

$$\begin{cases} D_r > \delta: & Key\ frames \\ D_r \leq \delta: & discard \end{cases} \tag{1}$$

and

$$D_r = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{u} \sum_{v} d_{ij}(u, v), \quad r = 1, \dots, N-1;$$

$$u = \frac{(i-1)R}{m} + 1, \dots, \frac{iR}{m}; \quad v = \frac{(j-1)C}{n} + 1, \dots, \frac{jC}{n} \tag{2}$$

where $D_r$ is the statistical difference between two consecutive frames, $N$ is the total number of frames in a single batch, $m$ and $n$ are the indices of the sub-regions, and there are $m \times n$ number of sub-regions in a single frame as Fig. 2 shows. $d_{ij}$ is the $(ij)^{th}$ statistical differences between each corresponding sub-regions in two consecutive frames computed by

$$d_{ij}(u, v) = I_{r+1}(u, v) - I_r(u, v), \quad r = 1, \dots, N-1;$$

$$i = 1, \dots, m; \quad j = 1, \dots, n;$$

$$u = \frac{(i-1)R}{m} + 1, \dots, \frac{iR}{m}; \quad v = \frac{(j-1)C}{n} + 1, \dots, \frac{jC}{n} \tag{3}$$

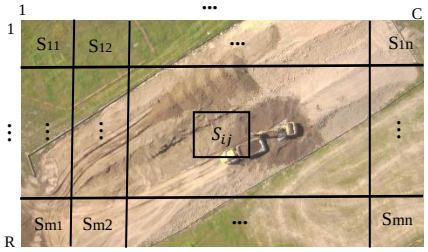


Fig. 2. Illustration of sub-region strategy in a single frame, where $S_{ij}$ is the $(ij)^{th}$ subregions.

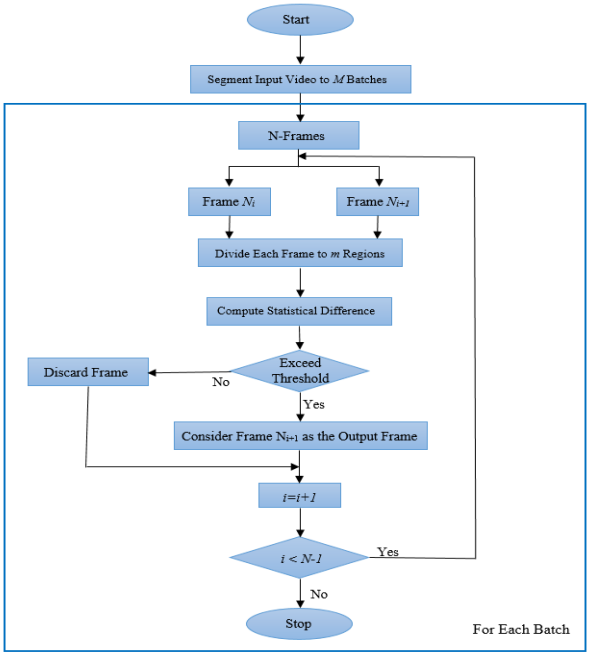


Fig. 3. Flowchart of the proposed key frame selection technique.

$$\delta = \frac{1}{N-1}\sum_{r=1}^{N-1} D_r + \beta \sqrt{\frac{1}{N-1}\sum_{r=1}^{N-1}\left[D_r - \left(\frac{1}{N-1}\sum_{r=1}^{N-1} D_r\right)\right]^2} \quad (4)$$

$\delta$ is the adaptive threshold and $\beta$ is a constant which controls the number of key-frames.

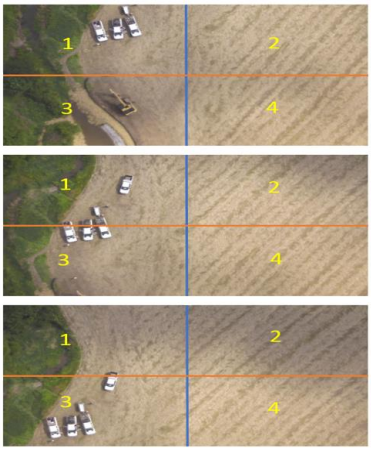


Fig. 4. Illustration of modular (sub-region) in multiple frames.

Khurana and Chandak technique [8] used the edge difference to calculate the difference between two sequential frames which is computationally expensive. On the other hand, we use the intensity space directly to calculate the statistical difference between two successive frames which is comparatively faster. Khurana and Chandak computed the edge differences between the entire two connected frames which may cause false negatives in case of some important contents that have relatively small edges compared to the background. Our modular technique can capture the small changes that appear in the scene by partitioning each frame into fine sub-regions and then calculating the statistical difference between each corresponding sub-regions.

## IV. Experimental Results and Analysis

To evaluate the proposed technique, we utilize three real-world aerial imagery datasets which were captured by a small aircraft at altitudes of 1000~2000 feet during several data acquisition sessions at different seasons and environmental conditions. The image size is 1920×1080. Some sample frames can be seen in Fig. 5. In this experiment, our goal is to keep the frames which include prominent contextual information and discard the rest which are redundant or undesired frames. For the implementation stage, we use Xeon(R) CPU, 2 GHz, 12 GB (RAM) PC in Python 2.7 software environment.

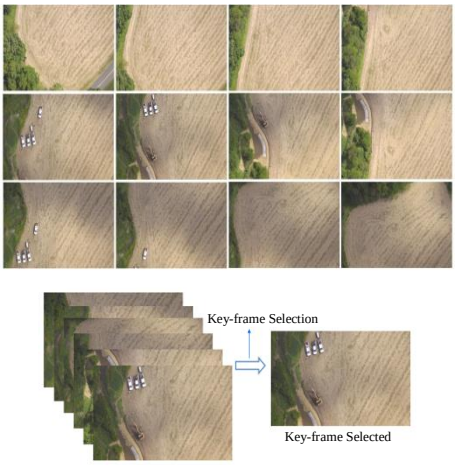


Fig. 5. Samples of one dataset.

### A. *Comparison Between Modular Key-frame and Non-Modular Key-frame Techniques*

In this section we compare our MKF with the non-modular key frame (NMKF) method. As we can see in Fig. 6, the MKF is able to select the frames which contain the objects of interest (e.g. the excavator and trucks) from the bunch of consecutive frames and discard the undesired ones, while the NMKF sometimes discards the frames that include the objects of interest.

Fig. 7 shows the capability of our MKF technique to capture the frames which have significant information in dataset 2 and neglect the insignificant ones. On the other hand, it also shows how the non-modular one fails to capture the frames which contain the objects of interest. Therefore, from Fig. 6 and Fig. 7 we can conclude that our MKF provides better performance in capturing the small local changes in the scene and produces less miss rate. When it comes to decrease the redundant information, our modular technique not only discard the insignificant frames, it also discards the frames which have repeated information as shown in Fig. 8.

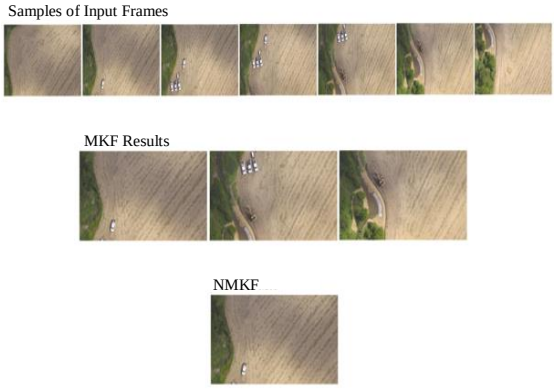


Fig. 6. Sample results on dataset 1.

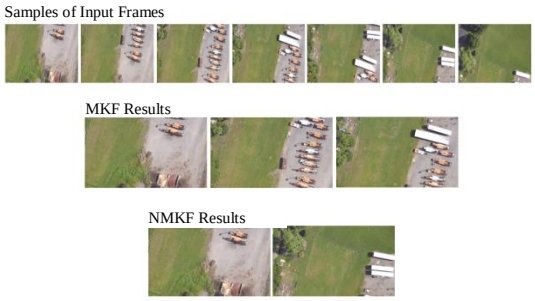


Fig. 7. Samples results on dataset 2.

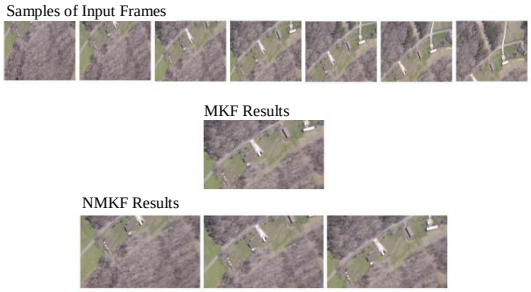


Fig. 8. Samples results on dataset 2.

When it comes to the time consumption, we apply our algorthim as a preprocessing stage for an automatic threat detection system (ATDS) which is devoloped to protect the pipline infrastructure. As seen in Fig. 9, our proposed technique enables the ATDS processes only 1945 frames instead of the whole 4380 frames in dataset 1. In dataset 2, only 2669 frames are used instead of 4510 frames, while in dataset 3, there are only 4840 frames kept as key frames from a set of 10983 frames without losing any important contextual information. This makes the computation time for datastes 1 and 2 enhanced from 54.15 and 52.39 minutes to 29.89 and 34.92 minutes respectively after using MKF. As for dataset 3, the computation time before MKF process was 139.23 minutes and after MKF it is reduced to 71.68 minutes as shown in Fig. 10.

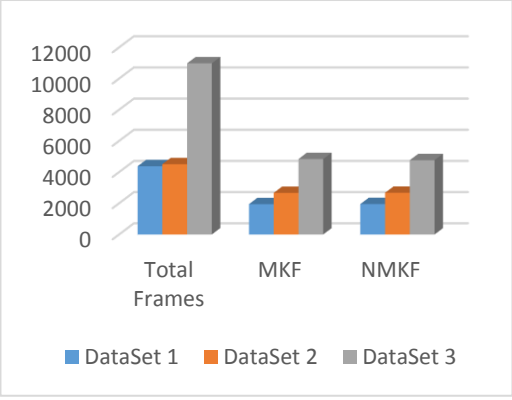


Fig. 9. Data reduction by key-frame selection.

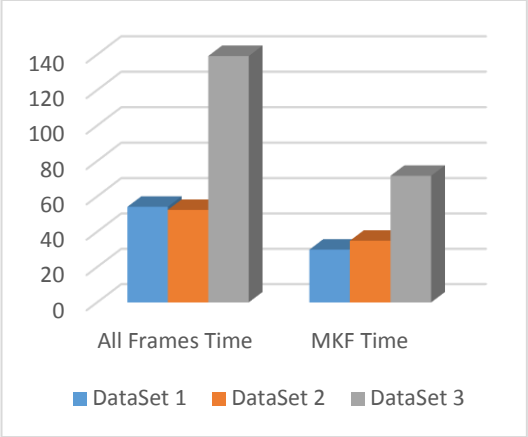


Fig. 10. Comparison of processing speed

## V. Conclusion

A new preprocessing technique is developed for large scale video analysis. The proposed MKF technique allows for capturing the small changes in the scene, such that it decreases the miss rate and improves computation time for wide area surveillance applications. From the experimental results, it is evident that our MKF approach has potential applications in analyzing big data to improve computation time without losing important contextual information.

## References

[1] C. Panagiotakis, A. Doulamis and G. Tziritas, "Equivalent Key Frames Selection Based on Iso-Content Distance and Iso-Distortion Principles," *8th International Workshop on Image Analysis for Multimedia Interactive Services* (WIAMIS'07), pp. 29, 2007.

[2] H.-C. Lee and S.-D. Kim, "Iterative key frame selection in the rate-constraint environment," *Signal Processing: Image Communication*, vol. 18, pp. 1-15, 2003.

[3] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proc. IEEE Int. Conf. Image Process.*, pp. 866-870, Oct. 1998,

[4] J. Vermaak, P. Perez, and M. Gangnet, "Rapid summarization and browsing of video sequences," in *British Machine Vision Conf.*, pp. 424-433, 2002.

[5] W. Wolf, "key frame selection by motion analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1228-1231, 1996.

[6] P. Gresele and T. Huang, "Gisting of video documents: A key frames selection algorithm using relative activity measure," *The 2nd International Conference on Visual Info. System*, 1997.

[7] A. Nagasaka and Y. Tanaka, "Automatic video indexing and full-video search for object appearances," *Proc. of the IFIP TC2/WG 2.6 Second Working Conference on Visual Database Systems II*, pp. 113-127, 1992.

[8] K. Khurana and M. B. Chandak, "Key Frame Extraction Methodology for Video Annotation," *International Journal of Computer Engineering and Technology (IJCET)*, vol.4, Issue 2, pp. 221-228, 2013.