Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Efficient Key Frame Extraction based on Adaptive Threshold and HOG for Video Summarization

Talib T. Al-Fatlawi[a], Salwa Shakir Baawi[b], Adil L. Albukhnefis[c]

[a]*Al-Qadisiyah University/College of Computer Science and Information Technology, Diwaniyah, Iraq, talib.turkey@qu.edu.iq*

[b]*Al-Qadisiyah University/College of Computer Science and Information Technology, Diwaniyah, Iraq, salwa.baawi@qu.edu.iq*

[c]*Al-Qadisiyah University/College of Computer Science and Information Technology, Diwaniyah, Iraq, adil.lateef@qu.edu.iq*

ARTICLE INFO

Article history:
Received: dd /mm/year
Rrevised form: dd /mm/year
Accepted : dd /mm/year
Available online: dd /mm/year

Keywords:

Video summarization, video compression, keyframe, adaptive threshold, Histogram of Oriented Gradient (HOG).

ABSTRACT

Key frame (KF) extraction process is considered as a crucial task in video structure analysis, it plays important role in video summarization, content analysis, video compression and so on. This process aimed to give a good video summarization by extracting the frame or set of frames that provide a comprehensive representation for the video sequence and removing the other frames that are considered redundant. In this paper, a new method has been proposed, the proposed method utilized Histogram of Oriented Gradient (HOG) as feature and adaptive thresholding technique, enabling the detection of substantial changes in visual content between consecutive frames in each video shot. By calculating the HOG differences and applying adaptive threshold, the method divides the shot's frames into groups, then from each groups a key frame with substantial features is selected, while other frames are considered as redundant and removed. Experimental evaluations on different videos, shows that the proposed methods able to extract key frames that accurately reflect the video's primary content and produce a good summarization. The results reduce the redundancy while preserving the essential idea of the video.

https://doi.org/10.29304/jqcsm.2025.17.22363

∗Corresponding author

Email addresses:

Communicated by ‘sub etitor’

Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

1. Introduction

In the era of digital media, where the process of capturing, generating, and sharing video content has become very easy, it has led to an exponential increase in the volume of video data generated daily. Many daily activities such as surveillance, sports, social media, education, scientific research, and entertainment, etc., have led to video being a ubiquitous form of information storage and communication [1]–[3]. Existing effective tools for analyzing, indexing, retrieving, and summarizing video have become very necessary. The process of extracting key frames is one of the fundamental steps in analyzing the structure of a video. Video is defined as a large data object with extensive information and high redundancies; it has a complex structure as shown in Fig. 1. Extracting key frames can be described as condensing or summarizing the video by selecting one or more frames that provide a complete picture of the video and capture most of its characteristics and content, while removing the remaining frames that are considered redundant. The process of extracting key frames can be defined as condensing or summarizing the video by extracting a frame or a group of frames that give a complete picture of the video and that have most of the characteristics and contents in the video, while deleting the remaining frames that can be considered redundant [1].

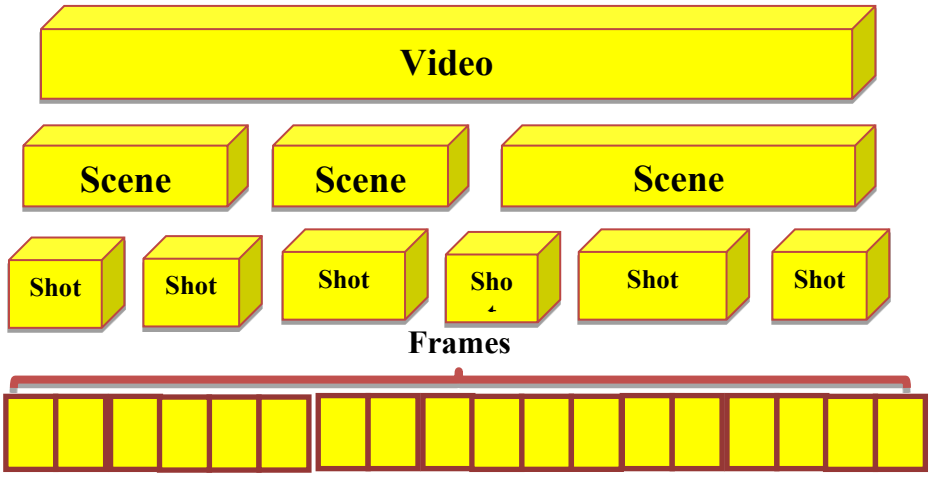

Fig. 1- Shows Video Structures

The selected key frames must enable users to understand the storyline of video without the need to view the entire video. Figure 2 shows the process of key frame extraction.

∗Corresponding author

Email addresses:

Communicated by 'sub etitor'

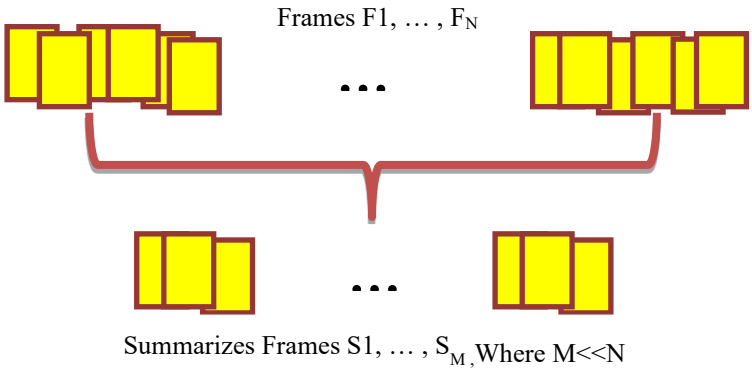

Fig. 2- Explains key frames extraction process

One of the major problems that arises when trying to extract the key frames is the determination of the optimal number of key frames (key frame set size); extracting few key frames can cause incomplete representation of the video narrative, missing the major moments and components in the scenes. On the other hand, extracting too many frames can lead to redundancy and cause computational cost and inefficient results. Therefore, the key frame number must satisfy the balance between comprehensiveness and efficiency. The size either determined priori as fixed number or posteriori based on methods and the visual changes [1], [4].

Several methods and features are used for key frame extraction. The most prominent techniques include: sequential comparison between frames, global comparison between frames, reference frame, curve simplification, clustering, object/event-based methods, and finally, panoramic frame [1], [4], [5].

In this paper, a new method is proposed. This method utilizes both HOG and adaptive techniques. First, the frames are divided into groups by comparing them with an adaptive threshold. Then, from each group, the frame with salient features is selected. In this method, the number of key frames is dynamic and depends on the changes in visual features in the video. Additionally, the proposed method takes into account.

The rest of the paper is organized as follows: Section 2 explains the HOG descriptor, ORB, and image enhancement. We introduced the proposed work in Section 3, Section explains the experimental results, and finally Section 4 is the conclusion.

2. Preliminaries

This section clarifies the important topics described in the following subsections.

2.1. Histogram of Oriented Gradient

Histogram of Orientated Gradients (HOG) is considered a robust feature descriptor widely used in image classification, computer vision, and pattern recognition. It works by dividing the image into small spatial regions called cells and computing the histogram of gradient orientations within each cell. HOG is particularly well suited for applications such as pedestrian detection, object recognition, and video analysis due to its sensitivity to edge structures while maintaining robustness to illumination changes [6]. Unlike many methods that rely on image refinement through Gaussian smoothing or similar filters, HOG operates directly on raw image gradients without pre-blurring, preserving fine edge details essential for accurate analysis. This balance between avoiding over-smoothing and retaining local structure has been emphasized in recent studies, especially in real-world contexts such as traffic sign recognition, image forgery detection, and deep learning-based restoration pipelines [7].

The major steps of HOG can be summarized as follows:

- Image preprocessing, such as converting the image to grayscale, resizing the image, or selecting a region of interest, preferably 64 * 128.
- Divide the region of interest into cells that are 8x8 in size.
- Compute the gradient for every cell, this process generates 64 gradient magnitudes.
- The computed magnitudes are binned into a 9-bin histogram (18-bin or 36-bin also accepted).
- Histograms are normalized to remove artifacts due to illumination changes.
- Finally, these histograms are collected into a single large histogram. The histogram based on the region of interest may be utilized in several applications, including SVM and machine learning.

2.2. Unsharp Mask Image Enhancement

Unsharp Mask (USM) is a classical image sharpening technique that enhances the visibility of edges by emphasizing high-frequency components. The method involves subtracting a blurred version of the original image to isolate edge information, then adding this difference back to the original image. The enhanced image I_enh is computed as:

$$I_{enh} = I_{orig} + \alpha(I_{orig} - I_{blur}) \quad (1)$$

Where α controls the sharpening strength. Gaussian blur is commonly used to create I_blur, with the blur radius enhancing the level of detail. This method preserves the overall image structure while boosting local contrast and edge sharpness.

Unsharp Mask is especially effective in preprocessing pipelines for feature detection tasks such as ORB or SIFT, where edge definition is critical. Applying USM to the luminance channel (e.g., V in HSV) often enhances keypoint detectability. Its low computational cost and interpretability make it suitable for real-time applications like video analysis and key frame extraction. Studies have demonstrated its effectiveness across domains, including medical imaging and object recognition [8]. Despite its simplicity, USM continues to be a powerful tool for improving visual quality in modern computer vision systems.

2.3. ORB

ORB (Oriented FAST and Rotated BRIEF) is a real-time feature detection and description algorithm that combines the FAST keypoint detector with the BRIEF descriptor, while enhancing both with rotation invariance and improved robustness. Unlike FAST, which lacks orientation data, ORB estimates the direction of each keypoint using the intensity centroid method, allowing it to adapt to image rotations. ORB ranks them using the Harris corner response to ensure that only the most stable keypoints are retained. This makes it especially useful in time-critical applications such as augmented reality, visual SLAM, and object recognition [9].

For descriptor construction, ORB modifies the original BRIEF by rotating the sampling pattern based on keypoint orientation, thus introducing rotation invariance. Additionally, a learning algorithm is used to select a subset of binary tests with low correlation and high variance, improving the descriptor's distinctiveness. These binary descriptors are matched using Hamming distance, which is computationally efficient compared to floating-point metrics used in SIFT and SURF. ORB strikes an optimal balance between speed and accuracy, making it highly suitable for mobile and embedded platforms [10], [11].

3. Proposed System

The proposed system uses HOG as features and an adaptive thresholding technique to identify the most representative frames in the video. The main goal is to capture frames that contain essential content and features while eliminating redundancy. An overview of the proposed system is shown in Figure 3.

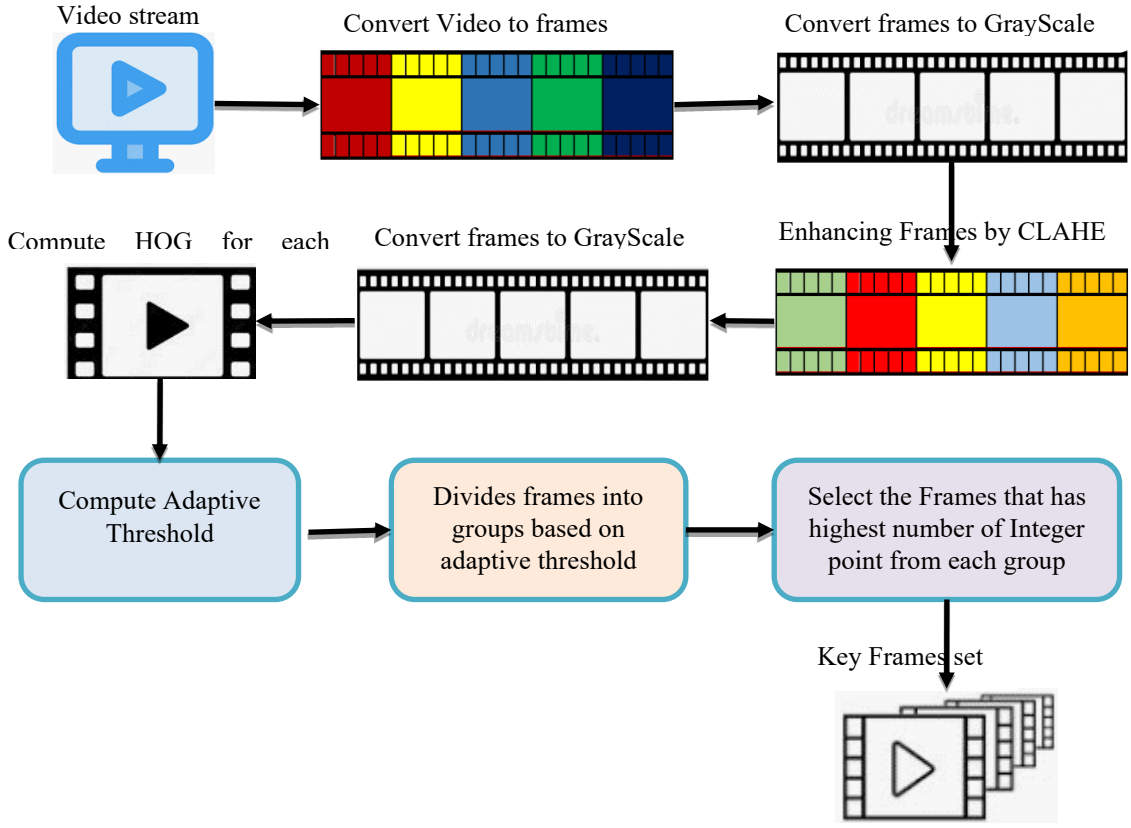

Fig. 3- Proposed System for Key Frames Extraction

The main steps of the proposed system can be summarized as follows:

1. Preprocessing: This step aims to prepare video frames for subsequent stages. It starts with converting the video into a sequence of frames and resizing each frame to a smaller size to reduce processing time and memory usage. Then, each frame is converted from RGB color space to HSV, where the processing focuses specifically on the V (Value) channel, which represents the luminance information. This channel contains the most important intensity features suitable for gradient-based descriptors like HOG. To improve the sharpness of the frame, the Unsharp Mask technique is applied to the V channel. This method enhances the edges by subtracting a blurred version of the image from the original and boosting the result, which makes the object boundaries more distinguishable. After that, the enhanced V channel is combined back with the original H and S channels to form an updated HSV image. Finally, the frame is converted back to BGR format and stored for later visualization, while the sharpened grayscale version of the V channel is extracted and preserved for HOG-based feature extraction. This preprocessing step ensures better contrast, sharper edges, and more robust structural information for the following stages.

2. Feature extraction: as it is mentioned previously, HOG is considered a powerful feature descriptor that can be used in many fields and applications. In this step, the HOG is constructed for each frame in the video frame sequence. Due to the nature of video, where the differences between the consecutive frames are very small, we adapt both the window size, block size and cell size to the full size of the image. This adaptation allows to analyze the gradient across the entire image and effectively compute the edge orientation and intensity.

3. Computing adaptive threshold: This step calculates the threshold based on statistical measurements of a set of HOGs. This threshold is computed and adjusted dynamically based on the HOGs distribution, providing a more context–sensitive boundary for decision-making process. In order to compute the threshold, first the

HOGs mean is computed, then it calculates the Euclidean distance between each HOG and this mean. After that, we compute the mean and the standard deviation of these distances. Later, the threshold (T) computed according to the following equation:

$$T = \mu + k \times \sigma \quad (2)$$

Where:

μ : is the mean of the differences between the HOGs mean and the HOG vectors

σ : is the standard deviation of the differences between the HOGs mean and the HOG vectors

k : is a constant number.

This method allows the threshold to adapt according to the variability in the visual contents and events in the scene, making it useful in scenarios where the changes in the data require a more flexible threshold rather than a fixed one. This leads to enhancing the accuracy of the subsequent steps.

4. Divides Frames: this step partitions the video frame sequence into coherent groups based on HOG similarity. This process compares the HOG of each frame with a HOG of the reference frame and groups frames that have similar HOG together.

   Initially, the HOG of first frame is marked as the reference, and then the HOGs of subsequent frames are compared to it by using the Euclidean distance (d). If d> T, the current group of frames are closed, and initializing a new group with the current frame as a new reference frame. This process continues until all frames have been processed. This procedure ensures that each group contains similar frames, and effectively divides the video into sections based on visual contents and events similarity.

5. Key Frame Selection: this step focused on selecting key frame from each group based on the number of ORB points. By analyzing the number of ORB points within each frame in a group, the system selects the frame that has the highest number of ORB points as a key frame for that group. This ensures that the extracted key frame from each group is more representative and informative and contain most important features within that section of video.

4. Experimental Results

The evaluation of the proposed key frame extraction system was conducted on a diverse set of ten video shots, each representing a different content type, including sports, documentaries, urban scenes, nature landscapes, and cartoons. Importantly, each video clip contained a single continuous shot, ensuring that content variation within the clip arises primarily from object motion and visual complexity, rather than explicit scene cuts. This experimental setup enables an in-depth assessment of the algorithm's sensitivity to intra-shot dynamics and content structure. Table 1 shows the description of the videos.

Table 1. Description of the videos

Video No.	Video size (MB)	FPS	Duration (sec)	No of Frames	Frame Resolution
1	23.49	25	13.84	436	1280 × 720
2	6.14	25	12	300	1280 × 720
3	41.75	25	30	755	1280 × 720
4	21.43	25	17.68	442	1280 × 720
5	4.50	25	12.64	316	1280 × 720
6	40.16	25	29.40	735	1280 × 720
7	114.01	25	65.48	1637	1280 × 720

8	64.02	25	41.52	1038	1280 × 720
9	12.98	25	23.52	588	1280 × 720
10	44.92	25	25.60	640	1280 × 720

Table 2 shows the number of extracted keyframes varied significantly with the value of the adaptive threshold parameter k. Lower values (e.g., k=0.5) consistently produced a higher number of key frames across all content types, indicating greater sensitivity to subtle variations in visual features. Conversely, higher values of k (e.g., 3 and 4) yielded fewer key frames, reflecting stricter thresholding and more selective segmentation. This behavior is evident in high-motion videos such as sports clips (e.g., Video7), where the number of key frames dropped from 51 at k = 0.5 to only 3 at k = 4. Similarly, animated content and nature videos with dynamic backgrounds (e.g., Video6, Video8, Video9) showed a high count of key frames at low k values due to frequent pixel-level variations. Figure 4. shows the result of the proposed systems.

Table 2 - The extracted key frames

Video No.	Number of extracted key frames				
	K=0.5	K=1	K=2	K=3	K=4
1	4	4	2	2	2
2	8	7	5	2	1
3	11	7	5	5	2
4	8	6	4	4	2
5	5	5	4	3	3
6	16	14	6	6	4
7	51	30	15	7	3
8	17	12	7	3	3
9	25	9	9	7	4
10	9	6	3	3	1

In contrast, low-motion or visually homogeneous videos—such as city tours or documentary segments (e.g., Video1, Video5) produced fewer key frames and showed more stable behavior across all k values. This consistency reflects the robustness of the method in suppressing redundancy when the visual content is temporally stable. The experimental results demonstrate that the integration of HOG features with adaptive thresholding successfully captures meaningful variations in gradient structure, even within a single shot, while minimizing unnecessary frame selection in static regions.

Moreover, the ability to control the granularity of segmentation through a single parameter (k) provides flexibility and ease of tuning, which is particularly advantageous for real-world applications. The algorithm adapts well to various content types without the need for handcrafted rules or prior knowledge about the video. Overall, the proposed system exhibits a favorable trade-off between sensitivity and redundancy control, validating its suitability for content-aware video summarization and indexing tasks, particularly in scenarios where scene boundaries are absent and decisions must rely solely on visual dynamics.

K =0.5

K= 1

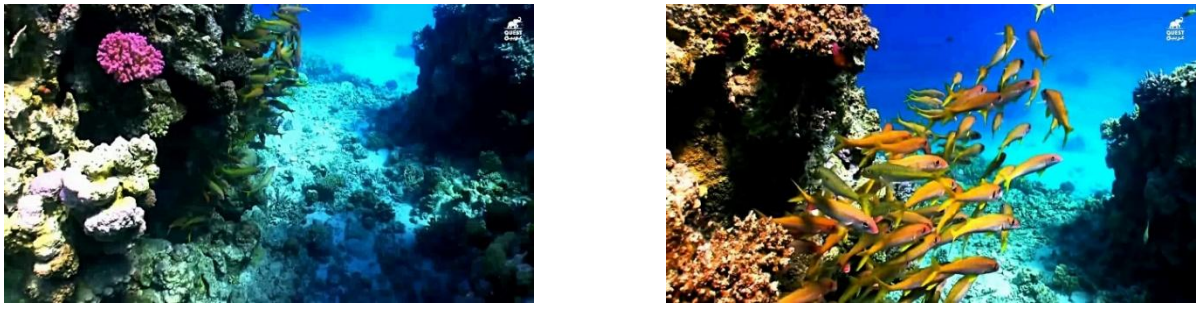

Fig. 4- Shows The Results of the Proposed System at Different K.

5. Conclusion and Future Work

In this paper, a content-aware key frame extraction system was proposed, designed specifically to handle single-shot video segments of varying content types, including sports, documentaries, natural landscapes, urban scenes, and animations. The system integrates Histogram of Oriented Gradients (HOG) as a global structural feature, combined with a statistically adaptive thresholding technique to segment video frames based on visual similarity. Additionally, the use of Unsharp Masking enhanced edge definition, contributing to improved keypoint detection using the ORB algorithm in the final selection stage. Experimental results across ten diverse videos confirmed the system's ability to balance redundancy reduction and content preservation, with a consistent correlation between the value of the threshold parameter kk and the number of extracted key frames. The approach demonstrated robustness across high-motion and low-motion videos, adaptability to content complexity, and practicality for video summarization tasks where no explicit shot boundaries exist.

Although the results are promising, several directions for future work can further enhance the system’s capabilities. One promising direction is to incorporate deep learning techniques such as Convolutional Neural Networks (CNNs)

as feature extractors instead of, or alongside, handcrafted descriptors like HOG and ORB. CNN-based features can capture higher-level semantic information, making the system more effective in complex and content-rich videos. Another important enhancement involves automating the selection of the threshold parameter k. Instead of relying on manual tuning, future versions of the system can leverage metaheuristic optimization algorithms such as Genetic Algorithms, Particle Swarm Optimization, or Ant Colony Optimization to determine the most appropriate k value based on the video's motion and structural properties.

Moreover, combining multiple types of features—such as gradient-based, texture-based, and deep semantic features—could improve the robustness and accuracy of both frame segmentation and key frame selection. This multimodal approach would allow the system to better handle a wide range of video genres, content complexities, and visual variations. Collectively, these future directions aim to make the system more adaptive, scalable, and suitable for real-world applications in video summarization, indexing, and intelligent multimedia retrieval.

References

[1] I. H. Ali and T. T. AL-Fatlawi, "Key Frame Extraction Methods," International Journal of Pure and Applied Mathematics, vol. 119, no. 10, pp. 485–490, 2018, doi: 10.1201/b19318-9.
[2] S. Jadon and M. Jasim, "Unsupervised video summarization framework using keyframe extraction and video skimming," 2020 IEEE 5th International Conference on Computing Communication and Automation, ICCCA 2020, pp. 140–145, 2020, doi: 10.1109/ICCCA49541.2020.9250764.
[3] H. B. Taher and A. H. Awadh, "Video Summarization for Surveillance System Using key-frame Extraction based on Cluster," Journal of Education for Pure Science- University of Thi-Qar, vol. 11, no. 1, pp. 54–65, 2021, doi: 10.32792/jeps.v11i1.91.
[4] G. Gao, "Key Frame Extraction," in Video Cataloguing, 1st Editio., 2015, pp. 57–64. doi: 10.1201/b19318-9.
[5] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, vol. 41, no. 6, pp. 797–819, 2011, doi: 10.1109/TSMCC.2011.2109710.
[6] S. Krig, "Interest Point Detector and Feature Descriptor Survey," Computer Vision Metrics, no. 1, pp. 187–246, 2016, doi: 10.1007/978-3-319-33762-3_6.
[7] I. Sedeeq, "Image Forgery Detection Using Histogram-Oriented Gradients (HOG)," Iraqi Journal of Science, vol. 66, no. 5, pp. 2048–2058, 2025, doi: 10.24996/ijs.2025.66.5.22.
[8] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 4th Editio. 330 Hudson Street, New York, NY 10013, 2018. doi: 10.1201/9781003002826-2.
[9] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in IEEE International Conference on Computer Vision, 2011, pp. 2564–2571.
[10] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in European conference on computer vision. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 778–792. doi: 10.1007/978-3-642-15561-1_56.
[11] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147–1163, 2015, doi: 10.1109/TRO.2015.2463671.