I. Introduction and Literature Review

Fire is a natural disaster that can cause extensive damage to property and endanger human lives. Conventional fire detection systems, which are primarily based on environmental factors such as room temperature, air concentration, and particle movements in the room, often struggle to detect fires in their early stages or in larger more open areas [1]. Additionally, the installation of such systems in existing buildings can be costly and complex based on the existing building infrastructure.

A more practical and cost-effective alternative lies in visual-based fire detection systems that utilize existing Close Circuit Television (CCTV) already present in most buildings, using such systems that use computer vision, a more robust fire detection system can be built [2].

In the realm of visual fire detection, fire color and shape properties are vital for determining possible Regions of Interest (ROIs). Traditional rule-based approaches identify fires based on color, often employing color spaces like RGB and YCbCr [3]. A study by A. f. Mutar [4] found HSV color space particularly effective for this purpose. In addition, While rule-based approaches can be effective in certain scenarios, they tend to struggle in environments with varied lighting conditions. Therefore, machine learning-based approaches, which uses large datasets for model training and refinement, have gained traction to solve this problem [5] [6] [7]. An example includes the approach by Ryu and Kwak, they combined computer vision techniques with a machine learning model as a pre-processing steps to improve detections. In the study, a HSV mask was first applied to filter out pixels with fire-like color. Subsequently, Harris Corner Detection was employed on the masked image, and only the top-facing corners were selected to define the ROIs. The ROIs was then classified by a CNN model, yielding a high F1-score of 97.4% [6].

The use of deep learning models, like Vision Transformer (ViT), for classifying extracted ROIs is another crucial component in modern fire detection systems. ViT models are known for their ability to extract meaningful features from images by utilizing self-attention mechanisms, capturing both global and local dependencies in an image [8] [9]. In the study by Zhang et al. [10], ViT was employed to detect fire occurrences in power plants. When compared to other classification models such as AlexNet, ResNet50, VGG16, VGG19, DenseNet121, and DenseNet169, the ViT model demonstrated superior performance with an F1-score of 93.12%. In another study by Shahid et al. [11], four out of six ViT models proposed in the work by Dosovitskiy et al. [8] were compared to Squeezenet and Inception-V1 for fire detection. It was concluded that the ViT-B/32 model achieved the highest score with an accuracy of 94.03%.

This research aims to contribute to the field of fire detection by introducing and evaluating a fire detection system built with a combination of HSV-based Harris Corner Region Extraction with Vision Transformer classification. This unique combination consisted of two parts, object detection with HSV color conversion and Harris Corner Detector, and object classification using a Vision Transformer model. With this research we hope our findings could provide insight and be an inspiration for the development of alternative methodologies in solving fire detection problems.