Automated Video Summarization for Suspicious Event Detection By Making Pipeline

Syed Muhammad Hussain
School of Science and Engineering
Habib University
Karachi, Pakistan
sh06892@st.habib.edu.pk

Muhammad Azeem Haider
School of Science and Engineering
Habib University
Karachi, Pakistan
mh06858@st.habib.edu.pk

Affan Habib
School of Science and Engineering
Habib University
Karachi, Pakistan
sh06892@st.habib.edu.pk

***Abstract*—With the widespread use of surveillance cameras in public places, there is an increasing need for automated methods to quickly analyze video footage and detect suspicious events. In this paper, we propose a novel approach for automated video summarization for suspicious event detection by using multi-modal. Our proposed method first extracts visual features from the input video as an event then classifies these events and generates a summary of the events. To evaluate the proposed approach, we conduct experiments on a publicly available dataset of surveillance videos. The results demonstrate that our method outperforms existing state-of-the-art techniques in terms of both summarization quality and suspicious event detection accuracy. Our approach can be useful for a variety of real-world applications, such as security surveillance, public safety, and law enforcement.**

***Index Terms*—OpenCV, Multimodal, CNN, Event Detection, Video Summarization**

I. Introduction

Surveillance through CCTV cameras has been in the works since the 1940s. Surveillance cameras have proven extremely useful in stopping crime due to the fear of getting caught. However, the sheer amount of data these cameras generate is overwhelming and requires much manual labor to analyze. This is where computer vision comes in. Computer vision is the science of making computers see and understand the world. It is a field that has been proliferating in the past few years.

With the advent of computer vision technology and algorithms, this manual labor has changed and has been made easier for humanity. Suspicious activity is now tracked through computer vision algorithms that notify the stakeholders when suspicious activity is recognized. Detecting and recognizing suspicious activities to stop them in time is a challenging task. However, with the help of computer vision, this task has become easier by instilling fear in the hearts of the culprits and making catching the culprits easier.

With the crime index of Pakistan being among the highest in the world, especially Karachi standing at 55.0 crime index, suspicious activity detection models have become the need of the hour in the country. Especially completely automated detection models that detect and recognize suspicious activities without human intervention. Summarizing 24-hour CCTV footage is another crucial task for the models to carry out so individuals do not have to skim through the footage to find suspicious activity.

While many models have been proposed in the past, they have yet to be able to achieve the desired results. This is due to various reasons, such as the need for a proper dataset, lack of proper training, and lack of proper testing. In this paper, we propose a novel approach for automated video summarization for suspicious event detection by using multi-modal. Our proposed method first extracts visual features from the input video as an event then classifies these events and generates a summary of the events through the activity labels. We have conducted experiments on a publicly available dataset of surveillance videos.

Automated security surveillance has become a need of the hour for the people of Pakistan and Karachi. Our model can be a real game changer since it can be used to detect suspicious activities in the city. In addition, the model can also be used to summarize 24-hour CCTV footage so that individuals do not have to skim through entire footage to find suspicious activity.

II. Literature Review

There has been a lot of related work in the field of object recognition through CCTV surveillance footage.

Social force model was used in this article for object recognition [1]. The authors use a dataset of surveillance videos of crowds and simulate the movement of individuals within the crowd using the social force model. They then analyze the movement patterns and detect anomalies using a clustering approach.

The method proposed involved two paths - slow and fast - and integrates both audio and visual data. [2] The authors compare their method with other state-of-the-art approaches and achieve superior performance in various datasets, including Kinetics, Charades, and AVA. The proposed approach shows promising results in recognizing complex video events with both visual and auditory cues, indicating the potential of audiovisual networks in video recognition tasks.

"SlowFast" a new video recognition algorithm was introduced in this paper which has a slow pathway with a low frame rate and high spatial resolution, and a fast pathway with a high frame rate and low spatial resolution. The authors argue

that this design allows the network to capture both temporal and spatial information effectively. [3]

The author of this research paper proposes a method for detecting video surveillance in order to protect privacy. The authors employ a VGG19 convolutional neural network (CNN) to analyze the content of the video and detect if it is from a surveillance camera. They pre-processed the dataset, which is called Snatch 1.01, by dividing it into frames and extracting features using the VGG19 algorithm. The results showed that the proposed system outperformed the state-of-the-art methods with 81% accuracy and a detection time of 0.025 frames per second. The authors conclude that the VGG19 CNN is effective in detecting video surveillance, which can be used to protect privacy. [4]

The author proposes a novel approach for detecting anomalies in surveillance videos. The approach involves using a two-stream spatiotemporal auto-encoder that learns to reconstruct normal behavior from the input video stream. The authors also introduce a new dataset, named UCF-Crime, which consists of real-world surveillance videos with anomalies. [5]

The findings of the review are presented in the form of various techniques and methods used to solve particular research problems, along with their strengths and weaknesses. The authors also highlight the scope for future work in this area. Overall, this research paper provides a critical review of various intelligent video surveillance techniques for suspicious activity detection, highlighting their strengths, weaknesses, and future potential for improving public security. They aim to understand the methodologies used for detecting abnormal human behavior, tracing abandoned objects, unattended baggage, and other suspicious activities. This analysis leads to an extensive comparison between various proposed methods, with many technologies based on intelligent techniques like neural systems, fuzzy logic, support vector machines, genetic algorithms, etc., emerging as a basis for intelligence in such systems. [6]

This paper proposes developing an intelligent video surveillance system that can actively monitor in real-time without human input. In solving the problems of the existing video surveillance system, deep learning technology will be carried through the data processing model design to visualize data for crime detection after building an artificial intelligence server and video surveillance camera. This design proposes an intelligent surveillance system to quickly and effectively detect crimes by sending a video image and notification message to the web through real-time processing. The model uses the COCO dataset for training. [7]

A comparison between audio and video analysis is drawn in an attempt to classify violence detection in real-time streams in this article. This study, which followed the CRISP-DM methodology, made use of several models available through PyTorch in order to test a diverse set of models and achieve robust results. The results obtained proved why video analysis has such prevalence, with the video classification handily outperforming its audio classification counterpart. [8]

This paper aims at studying and analyzing deep learning techniques for video-based anomalous activity detection. The focus has been given to various anomaly detection frameworks having deep learning techniques as their core methodology. Deep learning approaches from both the perspectives of accuracy-oriented anomaly detection and real-time processing-oriented anomaly detection are compared. [9] This paper also sheds light on research issues and challenges, application domains, benchmark datasets, and future directions in the domain of deep learning-based anomaly detection.

In this paper, the authors used a supervised learning approach and Convolutional Neural Network to detect unusual activities. The datasets consist of violence, robbery, and fire. The training module used to train the images with CNN and videos is the inception v2 model. The paper provided no model accuracy, although images were provided with an accuracy of detection on them which varied depending from image to image. [10]

The paper consists of two Neural Networks, CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network). [11] CNN is used for the purpose of extracting high-level features from the images so that the complexity of the input can be reduced. RNN is used for classification purposes, which is well suited for processing video streams. The system is using a pre-trained model called VGG-16 (Visual Geometry Group) which is trained on the Image Net Dataset. The research deals with suspicious activities inside a school. Such as students using mobile phones on campus is considered a suspicious activity and so are students fighting. They got the video through CCTV cameras and the videos are converted into frames in order to train the model.

This paper involved a novel activity recognition and detection framework utilizing the YOLOv4 version and the 3D CNN. Fine-tune convolution neural network architectures for better object recognition accuracy by incorporating object spatial and temporal information were also an important part of the research. [12] Region of interest (ROI) was detected using the fine-tuned version of Yolo-v4. A sequence of 16 frames was generated and ROI is passed through a sequence of frames into the 3D CNN for classification. The model had a 96.2% training accuracy and 94.2% validation accuracy.

From the papers, we read, and the research we carried out, we found no papers on a two-model system for the detection and recognition of activity which summarized the activities carried out over a specific amount of time.

III. Dataset

The DCSASS (Distributed Camera System for Anomaly and Suspicious Spatio-Temporal event detection) dataset available on Kaggle is a video dataset that consists of 3,305 video clips of anomalous and normal events captured by a distributed camera system. This dataset was built based on a dataset created by Sultani et al. (2018) and contains additional annotations and labels.

The videos in this dataset were captured at different locations and times, such as airports, train stations, and university campuses. The videos are in the MP4 format and

have a resolution of 640x360 pixels. The videos are divided into two classes: normal and anomalous events. The normal events include people walking, standing, and talking, while the anomalous events include actions like fighting, stealing, and vandalism.

This dataset contains videos based on the following 13 classes: Abuse, Arrest, Arson, Assault, Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism. Each video is labeled as normal (0) or abnormal (1) according to its content.

The DCSASS dataset includes a subset of surveillance camera videos that contain both normal and anomalous behaviors. Here's a brief overview of this subset:

- **Size** The subset contains a total of 15 videos, each with a duration of approximately 5 minutes.
- **Content** The videos show various scenes, such as arrest, burglary and shooting. The videos contain both normal and anomalous behaviors.
- **Source** The videos were collected from public surveillance cameras in different locations.
- **Usage** The videos can be used for a variety of purposes, such as training and testing of video-based anomaly detection algorithms, research on behavior analysis and crowd detection, and evaluation of surveillance systems.

The dataset includes annotations for each video clip, including the start and end time of the event, the type of event (normal or anomalous), and a description of the event. Additionally, the dataset includes precomputed optical flow features for each video clip, which can be used for machine learning and computer vision algorithms.

This dataset can be used for research on anomaly detection in surveillance videos and can be used to develop and evaluate algorithms for detecting abnormal events in real-world scenarios.

Overall, the surveillance camera video subset of the DCSASS dataset provides a useful resource for researchers and practitioners in the field of surveillance and security. It provides a realistic and diverse set of videos for testing and evaluation of surveillance systems and anomaly detection algorithms.

All in all, there is a total of 16853 videos, where 9676 videos are labeled as Normal and 7177 as abnormal.

IV. Methodology

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . ."

D. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. **??**", even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only

the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] Mehran, R., Oyama, A., & Shah, M. (2009, June). Abnormal crowd behavior detection using social force model. In 2009 IEEE conference on computer vision and pattern recognition (pp. 935-942). IEEE.

[2] Xiao, F., Lee, Y. J., Grauman, K., Malik, J., & Feichtenhofer, C. (2020). Audiovisual slowfast networks for video recognition. arXiv preprint arXiv:2001.08740.

[3] Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6202-6211).

[4] Butt, U. M., Letchmunan, S., Hassan, F. H., Zia, S., & Baqir, A. (2020). Detecting video surveillance using VGG19 convolutional neural networks. International Journal of Advanced Computer Science and Applications, 11(2).

[5] Sultani, W., Chen, C., & Shah, M. (2018). Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6479-6488).

[6] G. Mathur and M. Bundele, "Research on Intelligent Video Surveillance techniques for suspicious activity detection critical review," 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, India, 2016, pp. 1-8, doi: 10.1109/ICRAIE.2016.7939467.

[7] Sung, C. S., & Park, J. Y. (2021). Design of an intelligent video surveillance system for crime prevention: applying deep learning technology. Multimedia Tools and Applications, 1-13.

[8] Reinolds, F., Neto, C., & Machado, J. (2022). Deep learning for activity recognition using audio and video. Electronics, 11(5), 782.

[9] Pawar, K., & Attar, V. (2019). Deep learning approaches for video-based anomalous activity detection. World Wide Web, 22(2), 571-601.

[10] Tripathi, R. K., Jalal, A. S., & Agrawal, S. C. (2018). Suspicious human activity recognition: a review. Artificial Intelligence Review, 50, 283-339.

[11] Amrutha, C. V., Jyotsna, C., & Amudha, J. (2020, March). Deep learning approach for suspicious activity detection from surveillance video. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 335-339). IEEE.

[12] Rehman, A., Saba, T., Khan, M. Z., Damaševičius, R., & Bahaj, S. A. (2022). Internet-of-Things-Based Suspicious Activity Recognition Using Multi modalities of Computer Vision for Smart City Security. Security and Communication Networks, 2022.