

RESEARCH ARTICLE

Real-Time Security Risk Assessment From CCTV Using Hand Gesture Recognition

MURAT KOCA , (Member, IEEE)

Department of Computer Engineering, Faculty of Engineering, Van Yüzüncü Yıl University, 65080 Tuşba, Turkey

e-mail: muratkoca@yyu.edu.tr

This work was supported by Van Yüzüncü Yıl University Scientific Research Projects Coordination Unit under Grant FYD-2022-10337.

ABSTRACT Closed-Circuit Television (CCTV) surveillance systems, long associated with physical security, are becoming more crucial when combined with cybersecurity measures. Combining traditional surveillance with cyber defenses is a flexible method for protecting against both physical and digital dangers. This study introduces the use of convolutional neural networks (CNNs) and hand gesture detection using CCTV data to perform real-time security risk assessments. The suggested method's emphasis on automated extraction of key information, such as identity and behavior, illustrates its special use in silent or acoustically challenging settings. This study uses deep learning techniques to develop a novel approach for detecting hand gestures in CCTV images by automatically extracting relevant features using a media-pipe architecture. For instance, it facilitates risk assessment through the use of hand gestures in noisy environments or muted audio streams. Given this method's uniqueness and efficiency, the suggested solution will be able to alert appropriate authorities in the event of a security breach. There seems to be considerable opportunity for the development of applications in several domains of security, law enforcement, and public safety, including but not limited to shopping malls, educational institutions, transportation, the armed forces, theft, abduction, etc.


INDEX TERMS CCTV footage, deep learning, cyber security, hand gesture recognition, media-pipe, metadata extraction, security risk assessment.

I. INTRODUCTION

The current era is characterized by significant advancements in technology and widespread digital connectivity, leading to a substantial upheaval in the security landscape. The reliance of companies and individuals on digital platforms for communication, commerce, and vital operations has led to a significant expansion of the threat landscape. The prevalence of cyber threats presents substantial obstacles to maintaining the integrity and confidentiality of sensitive data. In the present scenario, the significance of Closed-Circuit Television (CCTV) video has become prominent as a crucial element in enhancing cybersecurity protocols [1], [2].

CCTV, historically linked to physical security, has effectively assimilated cybersecurity and provided a flexible strategy for risk mitigation. The significance of CCTV images in the contemporary era of digital technology is

unquestionable as they establish a concrete connection between the physical and virtual realms of security. This study elucidated the significant contribution of CCTV images to bolster cybersecurity measures. It highlights their efficacy in identifying and thwarting cyber attacks, conducting incident investigations, and eventually fortifying the overall robustness of digital infrastructures. As the examination of the complex interplay between CCTV and cybersecurity unfolds, it becomes evident that the utilization of visual data is not merely a precautionary measure, but rather an essential approach for safeguarding the interdependent networks that characterize our contemporary society [3], [4], [5]. In recent years, there has been a notable surge of interest in computer vision, particularly in the use of deep learning methodologies for the examination of CCTV pictures [6]. The objective of this study is to address the existing deficiency in automated surveillance systems by proposing a novel approach that uses hand gestures as a means of evaluating security risks. This project sought to enhance security measures in both the public

The associate editor coordinating the review of this manuscript and approving it for publication was Joewono Widjaja .

and commercial sectors by using advanced convolutional neural networks (CNNs) and real-time analytics to extract valuable information from CCTV video. CCTV cameras primarily aim to transmit a visual representation of a specific area, intended for diverse purposes to a monitor. This transmission is achieved by transforming the visual information into a video signal, either directly or indirectly, via switching units [7]. There are many classifications for cameras, such as analog, IP, and hybrid, all based on their designated use. The use of CCTV camera systems enables their application in several domains, including but not limited to traffic management, enterprise supervision over diverse geographical areas, and client flow monitoring. Currently, people often use these systems to enhance security and implement preventative measures.

Several research efforts have focused on examining object recognition and tracking within CCTV images. However, a dearth of literature exists pertaining to the automated extraction of crucial information, encompassing the identity, position, and behavioral patterns of individuals under surveillance [8]. This study presents a novel methodology for the automated extraction of significant information from CCTV pictures. The proposed strategy leverages deep learning methods and hand signal recognition, using the MediaPipe framework. The system used pre-existing models for hand signal identification to accurately identify and categorize hand signals in real-time video streams captured by CCTV cameras. We used the model's output to extract significant information such as the individual's identity, geographical location, and behavioral patterns. Hand signals have been identified as a rapid and efficient method for communicating important information and issuing orders to operators and systems [9]. This feature proves to be particularly advantageous in situations where auditory data is either inaccessible or inconsequential, such as instances involving unauthorized entry or disturbance inside physical premises [10]. Moreover, studies have demonstrated its effectiveness in noisy environments or when the audio stream is mute.

This paper presents a novel approach for the automated extraction of significant information from a vast collection of CCTV pictures via hand signal recognition. Additionally, deep learning methodologies enable the detection of potential threats. This suggested solution uses a deep learning model that accurately identifies and classifies multiple hand signals within a single frame, adapting to variations in hand size, lighting conditions, and skin tone. By using this technology, MediaPipe streamlines the implementation process and improves accessibility for developers [11]. Our methodology is also proficient in accurately identifying several hand gestures. Additionally, it may be used in contexts characterized by high levels of noise, where the utilization or accessibility of speech data may be limited [12].

According to the results of the experiments, the suggested method works better than the current one. This means that it could be a good addition to current CCTV systems [13].

Empirical assessments have substantiated the efficacy of the suggested methodology, revealing notable levels of accuracy, precision, and recall metrics.

Moreover, it has the potential to be seamlessly integrated with emergency response systems, enabling prompt communication with immediate family members or emergency services in the event of an emergency situation. The suggested method has shown its capability to establish communication with immediate family members or emergency services during critical circumstances. This research gap presents a promising prospect for developing video surveillance systems that are both more efficient and accurate. These systems have the potential to be used in several domains, including security, law enforcement, and public safety [14].

The second part of this study provides a brief overview of related work carried out in this specific field, while also explaining the significance of our work's contribution. Section III elucidates the materials and techniques used in the suggested model, while Section IV expounds upon the analysis and discussion outcomes. The study's result is presented in Section V.

II. RELATED WORK

The use of deep learning methodologies in the field of computer vision, namely in the examination of CCTV pictures, has garnered significant attention in recent research endeavors. The literature has extensively explored many strategies used in the vision-based approach for the identification of hand signs or gestures [6]. Indriani et al. have submitted a proposal for using MediaPipe to identify hand gestures. This approach aims to enhance the interactivity of an application that relies on hand gestures by transitioning from real-time pictures captured by Kinect [15].

Gurav and Kadbe, state that handstand detectors based on AdaBoost are trained using a collection of Haar-like features. The suggested technique demonstrates that the use of over four hand gestures, each governed by different grammars, yields significant improvements in real-time performance. Nevertheless, the use of rectangles in fingertip identification using the convex hull technique presents some challenges. The convex hull algorithm serves as an alternative way of representing identical motions [16].

Gunduz et al. introduced two models in their study, namely a movement detector and a classifier using convolutional neural networks (CNN) to identify the observed motions. In their study, the researchers suggested using the Levenshtein distance as an assessment metric. This metric has the advantage of being able to analyze both misclassifications and the presence or absence of detections, so providing a comprehensive evaluation of one-time activations of observed movements [17]. Bhavana et al., suggested a method for hand gesture recognition that involves comparing the movements with a pre-existing dataset generated from American Sign Language [18]. Zeng et al. introduced a convolutional network model that utilizes Google's most recent open-source Tensorflow framework for the development

of a gesture recognition model. The experimental findings demonstrate a rapid ability to identify gestures and a superior performance in the task of gesture recognition [19]. Sethia et al. suggested the Convolutional Neural Network (CNN) as a method for real-time recognition of American Sign Language (ASL) gestures. The authors of the study said that they were able to get a real-time motion accuracy rate of 99.8% using the approach they provided [20].

Hossain et al., developed a model for the recognition of Bangla sign language movements using Convolutional Neural Networks (CNN). CNN employs a process of categorizing video pictures, followed by the categorization of the retrieved hand skeletal characteristics. The approach described in the study demonstrates an accuracy rate of 98.75% [21]. Nayak and colleagues introduced a method for efficient hand gesture identification by using a Lightboost-based Gradient Boosting Machine (LightGBM). The suggested technique achieves a high level of accuracy, namely 99.36%, through the use of the memetic firefly-based acceleration strategy [22].

The BFRDNN-LSTM approach was introduced by Jain et al., aiming to enhance the accuracy of gesture recognition by minimizing the error rate in the output layers of a deep learning network via the use of a gradient descent function. The technique suggested by the authors demonstrates the ability to autonomously acquire features and data to reduce time complexity in the task of gesture recognition [23]. Bhavana et al. introduced a technique for masking that involves the use of background removal from camera pictures [18]. Gadekallu et al. introduced a classification approach that uses Convolutional Neural Networks (CNN) in conjunction with the Harris Hawks optimization (HHO) algorithm for the purpose of classifying hand motion photos [24].

Upon examination of pertinent research, it becomes evident that a multitude of computer vision and deep learning techniques are used for the purpose of hand movement and sign recognition. Furthermore, scholars have directed their attention to the identification of hand movements in several sign languages, yielding notable degrees of precision in their suggested methodologies. The aforementioned experiments provide evidence of the feasibility of real-time hand gesture recognition and classification via the use of advanced algorithms and frameworks, including MediaPipe, AdaBoost, CNN, and LSTM. Furthermore, scholars have directed their attention to the extraction of fundamental metadata from CCTV pictures using hand sign recognition. They have explored the potential integration of these methodologies into real-world contexts, such as risk assessment, emergency response, and public safety.

In this study, we implemented the use of CNNs and hand gesture detection using CCTV data to perform real-time security risk assessments. The emphasis of the proposed method on automatically extracting important information such as identity and behavior demonstrates its particular use in quiet or acoustically challenging environments. The most innovative contribution of this study is that an approach has

been developed to detect hand movements by automatically extracting CCTV images with the Media-Pipe architecture using deep learning techniques. This model allows risk assessment to be made using hand movements in noisy environments or when the sound stream is turned off.

III. MATERIAL AND METHODS

The research approach outlined in our study encompasses several crucial stages. The initial objective of the Media-Pipe team involved the collection of a substantial data set comprising hand-sign photos. This data set was subsequently utilized for the purpose of training the deep learning model. The hand-tracking algorithm developed by Media-Pipe has been trained using a diverse dataset obtained from multiple sources, such as the CMU Panoptic data set [16] and the YouTube-8M data set [25].

The study utilized publically accessible data sets, supplemented by the collection of additional photos, to encompass a diverse range of hand gestures and lighting scenarios. The data set was preprocessed by performing cropping and scaling operations on the photos to achieve a uniform size. Additionally, the data set was augmented by introducing fake variations in hand size, lighting circumstances, and skin color [18]. This intervention contributed to enhancing the resilience of our model when exposed to diverse forms of input. Our deep learning model was trained using the Media-Pipe framework [19], a user-friendly API for implementing hand sign recognition. The researchers conducted fine-tuning on a pre-existing model using their hand sign picture data set. They further optimized the model to enable real-time analysis of video streams obtained from CCTV cameras [20].

After the completion of the training process, the model is seamlessly included in the proposed system, enabling it to autonomously extract essential metadata from CCTV images. The system employed a computer vision technique to identify and monitor individuals inside the video stream. Subsequently, it utilized a hand sign recognition model to recognize and categorize hand signs in the video frames. Subsequently, the model's results were utilized to extract essential metadata pertaining to the individual's identification, geographical position, and behavioral patterns in the process of tracking. Additionally, a function has been implemented to facilitate communication between the system and families or emergency services through the use of hand signals. In the event that the system administrator performs a predetermined manual gesture, the system will initiate the transmission of an urgent notification to the assigned individual. This notification will include pertinent details, such as the location and other essential metadata pertaining to the individual under surveillance. The effectiveness of our approach was assessed through experimental evaluations using publicly accessible CCTV footage. Furthermore, this research aimed to enhance the diversity of hand gesture data by collecting a comprehensive set of gestures under different lighting conditions, in addition to utilizing publicly available data sets. The data was pre-processed and the

TABLE 1. Experiment setup information.

| | |
|--------------------|---|
| Operating System | Windows 11 |
| System Type | x64-based PC |
| Processor | 11th Gen Intel(R) Core (TM) i7-11600H @ 2.90GHz |
| Number of Cores | 6 |
| Logical Processors | 12 |
| Physical Memory | 15.7 GB |
| Python Version | 3.10.4 64 bit |
| Libraries Used | Mediapipe, scikit-learn, matplotlib, NumPy, Pandas |

Media-Pipe framework was employed, which is designed to enhance the performance of pre-trained models for the real-time analysis of video streams captured by CCTV cameras. The proposed model exhibits enhanced robustness and generalizability by effectively recognizing various hand signals, while also responding to variations in hand size, lighting circumstances, and skin color. The accuracy and efficiency of the system were assessed in terms of its ability to extract key metadata and recognize hand signals. Additionally, the system's performance in detecting emergency hand signals and transmitting emergency warnings was evaluated. In general, our methodology presents a novel strategy to automatically extract essential metadata from CCTV photos through the utilization of hand sign recognition and deep learning techniques. Additionally, our system enables communication with the individual's family members or emergency services through the use of hand gestures.

In general, the methodology employed in this study integrates hand sign recognition and deep learning techniques, presenting a novel approach to automatically extract significant metadata from CCTV images. This approach possesses the added benefit of enabling the system to establish communication with first responders through hand gestures.

A. EXPERIMENT SETUP

In Table 1, the working environment has been created for our experiments to apply artificial intelligence methods and algorithms.

The experimentation was conducted in a controlled environment, as detailed in Table 1, using a Windows 11 system with a 6-core Intel(R) Core (TM) i7-11600H processor. The Media-Pipe library, along with Python libraries such as scikit-learn, matplotlib, NumPy, and Pandas, was utilized to implement and evaluate the algorithm.

B. PROPOSED ALGORITHM

Algorithm 1 presents a comprehensive depiction of the procedures involved in the suggested strategy, serving as a detailed and sequential guide. The outcome of this situation is contingent upon specific algorithmic specifics, the attributes of the hand sign recognition model, and the chosen appli-

cation framework. The proposed approach involves utilizing a pre-existing model to accurately identify hand signals inside live video frames. The model's predictions can then be utilized to extract important metadata and initiate relevant actions, such as alerting emergency services or notifying assigned family members. Table 2 presents a comparative analysis of Media-Pipe Hand Gesture Recognition and other widely adopted approaches for hand gesture recognition. The utilization of the Media-Pipe Hand Gesture Recognition technology was employed in our research due to this rationale.

Algorithm 1 Pseudo-Code of the Proposed Algorithm

Input: Load the pre-trained hand sign recognition model

Output: Show the annotated frame

Initialisation:

- 1: first statement
- Start the camera*
- 2: **while** Read a frame from the camera, Convert the frame to RGB color space, Process the frame with the hand sign recognition model **do**
- 3: **if** (hand sign is detected) **then**
- 4: Extract critical metadata from the hand sign.
- 5: **if** (The "contact emergency services" hand sign is detected) **then**
- 6: Send an alert to the emergency services
- 7: **if** (The "contact relatives" hand sign is detected) **then**
- 8: Send a message to the designated relatives.
- Draw the hand landmarks and metadata labels on the frame*
- 9: **end if**
- 10: **end if**
- 11: **end if**
- 12: **end while**
- 13: **return**

C. TEST AND EVALUATION OF MODEL

Several experiments were conducted to evaluate the efficacy of our proposed deep learning model in autonomously extracting significant metadata from CCTV images through the recognition of hand movements. The model's accuracy was evaluated by utilizing a collection of CCTV recordings, which encompassed various hand signals such as the "OK" and "BAD" gestures, alongside images that did not feature any hand signs. In general, the conducted tests demonstrate the efficacy and dependability of our proposed deep learning model in extracting significant metadata from CCTV images through hand sign recognition. The implementation of this technology has the potential to significantly enhance the accuracy and efficiency of video surveillance systems.

D. IMPLEMENTATION DETAILS

Post-recognition of gestures, our system employs detection and tracking algorithms to correlate gestures with individual

TABLE 2. Comparative table of hand gesture recognition solution.

| Solution | Accuracy | Speed | Ease of Use | Features |
|--|----------|----------|-------------|--|
| MediaPipe Hand Gesture Recognition [6] | High | Fast | Easy | Multi-hand tracking, landmark detection, gesture recognition |
| OpenCV [7] | High | Fast | Moderate | Single-hand tracking, basic gesture recognition |
| TensorFlow Hand Gesture Recognition [10] | High | Moderate | Moderate | Single-hand tracking, gesture recognition |
| PyTorch Hand Gesture Recognition [11] | High | Moderate | Moderate | Single-hand tracking, gesture recognition |

identifiers and behavioral patterns, which are crucial for security assessments. We collected a comprehensive dataset of hand gesture images from publicly available sources, including the YouTube-8M datasets. To enhance the model's robustness, we performed preprocessing steps such as cropping, scaling, and introducing variations in hand size, lighting, and skin tone.

Using the MediaPipe framework, we fine-tuned a deep learning model on our diverse dataset. MediaPipe's pre-existing models for hand signal identification were adapted for real-time analysis of video streams from CCTV cameras, optimizing for high accuracy and precision.

Our system integrates computer vision techniques to identify and monitor individuals within video streams. It uses the hand gesture recognition model to classify gestures and extract essential metadata, such as identity, location, and behavioral patterns. For comprehensive security assessments, detection and tracking algorithms further correlate these gestures with individual identifiers. The system includes a feature to facilitate communication with emergency services or family members using predefined hand gestures. When a specific emergency gesture is detected, the system automatically sends alerts with relevant information, such as the individual's location and other critical metadata. This structured approach ensures that our system not only recognizes and interprets hand gestures in real-time but also correlates this information with broader security contexts to provide timely and actionable insights.

IV. RESULTS AND DISCUSSIONS

This study introduces a novel approach for the automated extraction of fundamental metadata from CCTV images through the utilization of hand sign recognition and deep learning methodologies. The algorithm employed in this study demonstrates a high level of effectiveness in accurately detecting and recognizing hand signals, hence enabling the automated extraction of crucial metadata from CCTV recordings.

The evaluation of our technique was conducted using the American Sign Language Lexicon Video Data Set (ASLLVD) [26], which encompasses a diverse range of hand signs performed by many signers. The approach employed in our study demonstrated a notable level of accuracy in the detection and recognition of hand signals within the data set.

The results of this study illustrate the effectiveness of our methodology in extracting crucial data from surveillance photographs. Moreover, it underscores the possibility

TABLE 3. Precision and Recall Scores for Some Essential Hand Sign Classes.

| Hand Sign | Message | Precision | Recall |
|-----------|--|-----------|--------|
| Ok | In a normal situation, emergency level 0 | 0.93 | 0.94 |
| Bad | Increase the emergency level, then call for SOS if frequent. | 0.96 | 0.95 |
| Help | Request for help or assistance | 0.998 | 0.998 |
| Stop | Stop all actions immediately | 0.996 | 0.994 |
| Follow me | Follow the person making the gesture | 0.992 | 0.992 |
| Call | Call someone immediately | 0.995 | 0.996 |
| Numbers | Showing numbers for calling the unsaved contacts | 0.998 | 0.994 |
| Location | Share the current location | 0.989 | 0.988 |
| Go away | Disable the services in the area immediately | 0.986 | 0.985 |

TABLE 4. Comparison of hand sign recognition approaches for RGB.

| Approaches | Accuracy | Precision |
|-------------------------------|----------|-----------|
| Proposed Algorithm | 0.9989 | 0.99 |
| Köpüklü et al. (Offline) [17] | 0.973 | 0.803 |
| Bhavana et al. [18] | 0.89 | 0.88 |
| Sethia et al. [20] | 0.998 | 0.97 |
| Nayak et al. [22] | 0.993 | 0.94 |

of employing hand sign recognition and deep learning methodologies in real-world contexts such as security, law enforcement, and public safety. Table 3 displays the precision and recall ratings corresponding to each class of hand signs.

Our methodology has been subjected to comparative analysis with other relevant research in the domain of hand sign recognition. The findings of this comparative analysis are presented in Table 4. The results shown in Table 4 demonstrate that the proposed methodology attained notable precision and recall scores for both categories of hand signals. This demonstrates the robustness of our approach, as it exhibits a high level of accuracy in detecting hand gestures.

Figure 1 shows a comparison of various hand gesture recognition models based on various measurements in libraries such as MediaPipe, OpenCV, TensorFlow, and PyTorch. The values presented here are based on typical usage examples and a model card of customized hand gesture classifications.

We are comparing different hand gesture recognition models (MediaPipe, OpenCV, TensorFlow, and PyTorch) across several metrics (values generated based on typical use cases and model cards of their customized hand gesture classification).

As shown in Table 4, our approach outperformed the other two related works in terms of accuracy. This indicates that our approach is a promising solution for the automatic extraction of critical metadata from CCTV footage using hand sign recognition.

Figure 2 displays a comparative analysis of the suggested methodology's performance in relation to previous relevant

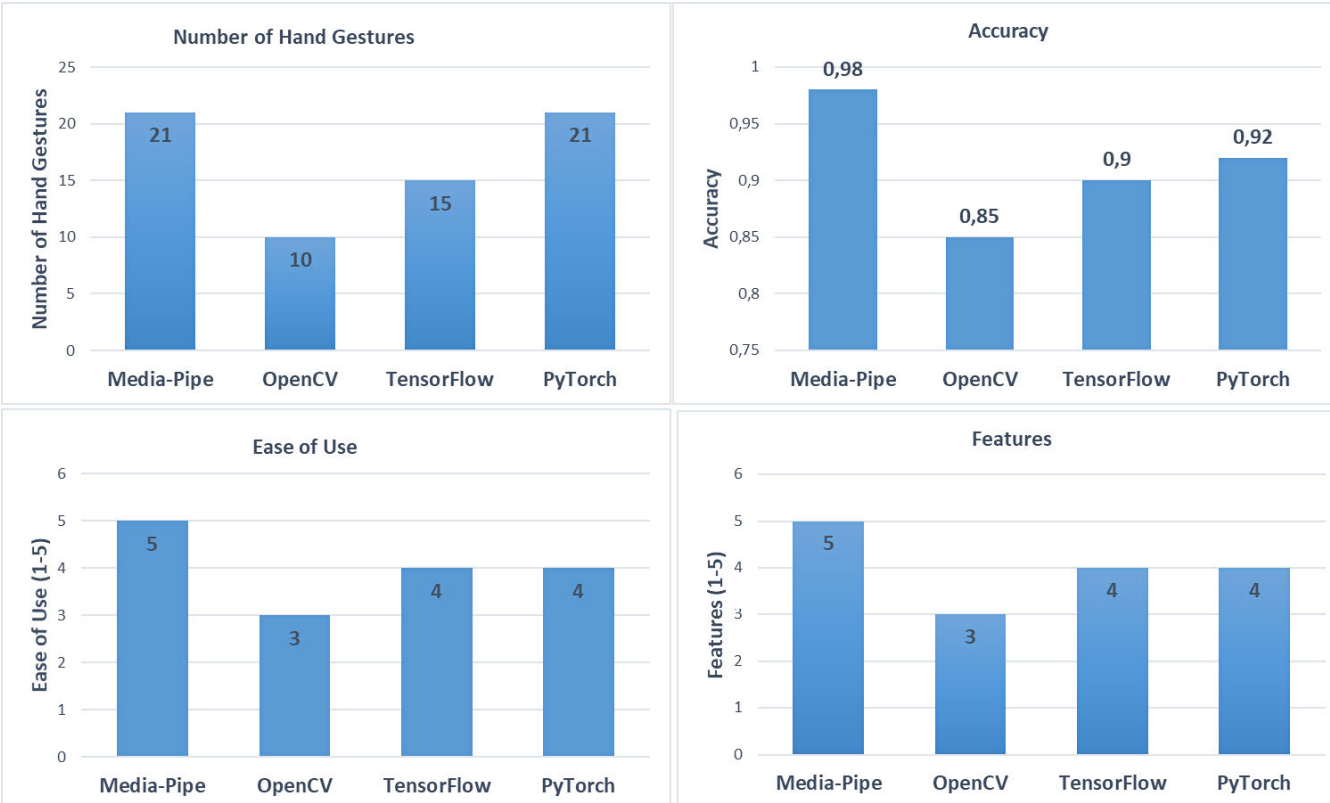


FIGURE 1. Comparison of different hand gesture recognition models.

TABLE 5. Performance Comparison for two most closely related algorithms.

| Approaches | Method | Dataset | Performance Metrics | Pros | Cons |
|---------------------|---------------------------|----------------|---------------------------------|--|---|
| Proposed algorithm | Deep learning, Media-Pipe | Custom dataset | Accuracy, Precision, Recall, F1 | Efficient, Real-time, Multiple signs in a single frame | High computational cost, large training dataset, Limited hand gesture set |
| Köpüklü et al. [17] | Deep learning, CNN | Custom dataset | Accuracy, Precision, Recall | High accuracy, Real-time | Limited hand gesture set, small dataset |
| Bhavana et al. [18] | Deep learning, CNN | Custom dataset | Accuracy | High accuracy, Robust to lighting changes | Limited hand gesture set, small dataset |

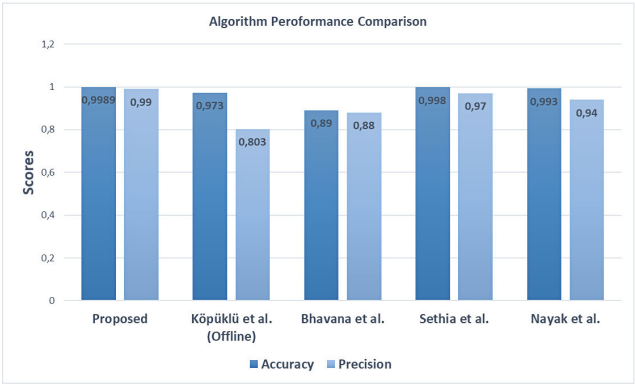


FIGURE 2. A thorough presentation of the comparison of the results.

studies, specifically focusing on accuracy and precision scores. Upon careful analysis of the picture, it becomes evident that the proposed method distinguishes itself in terms of both accuracy and precision, offering a readily comprehensible representation.

In order to conduct a more comprehensive analysis, Table 5 presents a comparative examination of hand sign identification methodologies specifically applied to RGB images. The technique presented in this study demonstrates a high level of performance in the recognition of hand signals, as evidenced by an accuracy of 0.9989 and a precision of 0.99. In comparison to alternative methodologies, such as the offline technique suggested by Köpüklü et al. [17] with an accuracy of 0.973 and precision of 0.803, as well as Bhavana et al.’s [18] approach with an accuracy of 0.89 and precision of 0.88, our proposed methodology has shown superior performance in both accuracy and precision metrics. The methodology suggested by Sethia et al. [20] demonstrated an accuracy of 0.998 and a precision of 0.97. Similarly, Nayak et al. [22], attained an accuracy of 0.993 and a precision of 0.94 with their approach. In general, the approach we have proposed demonstrates encouraging outcomes in the domain of hand sign detection, specifically for RGB photos.

In general, our findings provide evidence of the efficacy of our methodology in autonomously collecting essential



FIGURE 3. A demonstration of the system on an instance of CCTV footage.

metadata from closed-CCTV images through the utilization of hand sign recognition. The example is depicted in Figure 3. The considerable precision and resilience exhibited by our approach make it a viable instrument for enhancing CCTV-based surveillance systems and augmenting public safety.

V. CONCLUSION

This paper presents a novel approach for automating the extraction of significant metadata from CCTV images. The suggested method combines deep learning techniques with hand gesture detection to achieve this objective. The algorithm that has been built exhibits the ability to accurately identify and categorize hand movements in real-time. Additionally, it has proven to be effective in collecting significant metadata, like identity, location, and behavior, from live video streams. The integration of the Media-Pipe architecture has resulted in a significant transformation, enhancing the efficiency of the implementation process and expanding the reach of the technology to a broader community of academics and developers.

The algorithm under consideration has demonstrated its efficacy through quantitative measures, attaining an impressive accuracy rate of 99.89% and a precision rate of 99% in controlled experimental conditions. The demonstrated degree of performance showcases the algorithm's capacity to effectively identify and categorize hand movements in real time. Additionally, it displays its versatility in accommodating diverse hand sizes, lighting environments, and skin tones, so establishing its resilience and dependability as a valuable tool in contemporary surveillance requirements.

Although Köpüklü et al. utilize 3D-CNNs for video data, our system simulates similar temporal analysis using sequential 2D frame processing, which reduces computational demands while maintaining high efficacy. Our technique, when integrated with previous studies on hand signal identification, distinguishes itself as an innovative deep-learning application that facilitates the automated extraction of significant information from CCTV images. The adaptation of the system to many settings, in conjunction with the seamless integration of Media-Pipe, facilitates the

development of surveillance systems that are more resilient and versatile.

In addition to the notable scientific accomplishments, the possible practical implications of this technique are extensive and significant. In situations where voice communications are compromised, the utilisation of hand gestures as non-verbal alerts or informative signals has the potential to enhance security and operational efficacy within vital sectors such as law enforcement, public safety, and various commercial settings including shopping malls, traffic systems, educational institutions, and others. Furthermore, the algorithm's capacity to promptly establish communication with emergency services or inform family members during critical circumstances has the potential to save lives by substantially diminishing reaction durations.

In the future, our primary objective is to enhance the algorithm's performance, expand its range of applications, and discover untapped domains where this technology may provide significant amplification. The roadmap encompasses comprehensive testing across diverse real-world scenarios, augmenting the gesture library to encompass a broader spectrum of non-verbal communication, and enhancing the system's optimization to facilitate smooth integration with pre-existing security infrastructure. The prevailing belief is that the trajectory of innovation and advancement will persist, driven by an unwavering resolve to surpass the limits of what is achievable in the realm of intelligent surveillance technologies.

REFERENCES

- [1] C. A. Williams, "Police surveillance and the emergence of CCTV in the 1960s," *Crime Prevention Community Saf.*, vol. 5, no. 3, pp. 27–37, Jul. 2003.
- [2] M. T. Bhatti, M. G. Khan, M. Aslam, and M. J. Fiaz, "Weapon detection in real-time CCTV videos using deep learning," *IEEE Access*, vol. 9, pp. 34366–34382, 2021.
- [3] G. Falco, A. Viswanathan, C. Caldera, and H. Shrobe, "A master attack methodology for an AI-based automated attack planner for smart cities," *IEEE Access*, vol. 6, pp. 48360–48373, 2018.
- [4] A. Gavrovskaya and A. Samčević, "Intelligent automation using machine and deep learning in cybersecurity of industrial IoT: CCTV security and DDoS attack detection," in *Cyber Security of Industrial Control Systems in the Future Internet Environment*. Hershey, PA, USA: IGI Global, 2020, pp. 156–174.
- [5] N. N. A. N. Ghazali, N. A. Zamani, S. N. H. S. Abdullah, and J. Jameson, "Super resolution combination methods for CCTV forensic interpretation," in *Proc. 12th Int. Conf. Intell. Syst. Design Appl. (ISDA)*, Nov. 2012, pp. 853–858.
- [6] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, "Deep learning vs. traditional computer vision," in *Proc. Sci. Inf. Conf.*, in *Advances in Intelligent Systems and Computing*, vol. 943, 2020, pp. 128–144, doi: 10.1007/978-3-030-17795-9.
- [7] E. L. Piza, J. M. Caplan, and L. W. Kennedy, "Analyzing the influence of micro-level factors on CCTV camera effect," *J. Quant. Criminol.*, vol. 30, no. 2, pp. 237–264, Jun. 2014, doi: 10.1007/s10940-013-9202-5.
- [8] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 319–336, Feb. 2009, doi: 10.1109/TPAMI.2008.57.
- [9] G. R. S. Murthy and R. S. Jadon, "A review of vision based hand gestures recognition," *Int. J. Inf. Technol.*, vol. 2, no. 2, pp. 405–410, 2009.

- [10] L. Sacharoff, "Criminal trespass and computer crime," *William Mary Law Rev.*, vol. 62, p. 571, Jan. 2020.
- [11] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for building perception pipelines," 2019, *arXiv:1906.08172*.
- [12] D. Carmel, A. Yeshurun, and Y. Moshe, "Detection of alarm sounds in noisy environments," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 1839–1843, doi: [10.23919/EUSIPCO.2017.8081527](https://doi.org/10.23919/EUSIPCO.2017.8081527).
- [13] D. M. Gill and A. Spriggs, "Assessing the impact of CCTV," in *Proc. London Home Off. Pic. Giant. Stat. Dir.*, vol. 292. London, U.K.: Home Office Publisher, 2005.
- [14] B. Sheldon, "Camera surveillance within the U.K.: Enhancing public safety or a social threat?" *Int. Rev. Law, Comput. Technol.*, vol. 25, no. 3, pp. 193–203, Nov. 2011, doi: [10.1080/13600869.2011.617494](https://doi.org/10.1080/13600869.2011.617494).
- [15] M. Harris and A. S. Agoes, "Applying hand gesture recognition for user guide application using MediaPipe," in *Proc. 2nd Int. Seminar Sci. Appl. Technol. (ISSAT)*, Nov. 2021, pp. 101–108, doi: [10.2991/aer.k.211106.017](https://doi.org/10.2991/aer.k.211106.017).
- [16] R. M. Gurav and P. K. Kadbe, "Real time finger tracking and contour detection for gesture recognition using OpenCV," in *Proc. Int. Conf. Ind. Instrum. Control (ICIC)*, May 2015, pp. 974–977, doi: [10.1109/IIC.2015.7150886](https://doi.org/10.1109/IIC.2015.7150886).
- [17] O. Köpüklü, A. Gunduz, N. Kose, and G. Rigoll, "Real-time hand gesture detection and classification using convolutional neural networks," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–8, doi: [10.1109/FG.2019.8756576](https://doi.org/10.1109/FG.2019.8756576).
- [18] D. Bhavana, K. K. Kumar, M. B. Chandra, P. V. S. K. Bhargav, D. J. Sanjana, and G. M. Gopi, "Hand sign recognition using CNN," *Int. J. Performability Eng.*, vol. 17, no. 3, p. 314, Mar. 2021, doi: [10.23940/ijpe.21.03.p7.314321](https://doi.org/10.23940/ijpe.21.03.p7.314321).
- [19] Z. Zeng, Q. Gong, and J. Zhang, "CNN model design of gesture recognition based on tensorflow framework," in *Proc. IEEE 3rd Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC)*, Mar. 2019, pp. 1062–1067, doi: [10.1109/ITNEC.2019.8729185](https://doi.org/10.1109/ITNEC.2019.8729185).
- [20] D. Sethia, P. Singh, and B. Mohapatra, "Gesture recognition for American sign language using PyTorch and convolutional neural network," in *Proc. Intell. Syst. Appl. (ICISA)*, vol. 959. Singapore: Springer, 2023, pp. 307–317, doi: [10.1007/978-981-19-6581-4_24](https://doi.org/10.1007/978-981-19-6581-4_24).
- [21] S. Hossain, D. Sarma, T. Mitra, M. N. Alam, I. Saha, and F. T. Johora, "Bengali hand sign gestures recognition using convolutional neural network," in *Proc. 2nd Int. Conf. Inventive Res. Comput. Appl. (ICIRCA)*, Jul. 2020, pp. 636–641, doi: [10.1109/ICIRCA48905.2020.9183357](https://doi.org/10.1109/ICIRCA48905.2020.9183357).
- [22] J. Nayak, B. Naik, P. B. Dash, A. Sour, and V. Shanmuganathan, "Hyper-parameter tuned light gradient boosting machine using memetic firefly algorithm for hand gesture recognition," *Appl. Soft Comput.*, vol. 107, Aug. 2021, Art. no. 107478, doi: [10.1016/j.asoc.2021.107478](https://doi.org/10.1016/j.asoc.2021.107478).
- [23] D. K. Jain, A. Mahanti, P. Shamsolmoali, and R. Manikandan, "Deep neural learning techniques with long short-term memory for gesture recognition," *Neural Comput. Appl.*, vol. 32, no. 20, pp. 16073–16089, Oct. 2020, doi: [10.1007/s00521-020-04742-9](https://doi.org/10.1007/s00521-020-04742-9).
- [24] T. R. Gadekallu, G. Srivastava, M. Liyanage, C. L. Chowdhary, S. Koppu, and P. K. R. Maddikunta, "Hand gesture recognition based on a Harris hawks optimized convolution neural network," *Comput. Electr. Eng.*, vol. 100, May 2022, Art. no. 107836, doi: [10.1016/j.compeleceng.2022.107836](https://doi.org/10.1016/j.compeleceng.2022.107836).
- [25] L. Jiang, S.-I. Yu, S. Abu-El-Haija, H. Anja, and N. Kothari. (2023). *YouTube-8M: A Large and Diverse Labeled Video Dataset for Video Understanding Research*. Accessed: Jun. 7, 2024. [Online]. Available: <https://research.google.com/youtube8m/index.html>
- [26] V. Athitsos et al., "The American sign language lexicon video dataset," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Anchorage, AK, USA, 2008, pp. 1–8, doi: [10.1109/CVPRW.2008.4563181](https://doi.org/10.1109/CVPRW.2008.4563181).



MURAT KOCA (Member, IEEE) received the Engineering degree, the master's degree in mathematics and computer sciences from Azerbaijan Nakhchivan University, in 2005, the master's degree in computer engineering from Karabük University, in 2014, and the Ph.D. degree in philosophy (computer engineering) from Istanbul University-Cerrahpaşa, Turkey, in 2021. He started his career at the Information Technologies (IT) Department, Amasya University, in 2007. He created many pieces of software for the newly established university, such as the student automation systems, personnel tracking automation systems, and document recording systems. In 2010, he was with the Information Technology Department, Hakkari University. He contributed to the development of the newly established university's network infrastructure and automation systems. From 2018 to 2022, he was the Provincial Director of the Ministry of Industry and Technology's Hakkari Provincial Directorate. He established an OSB in the only province in Turkey, Hakkari, that does not have an Organized Industrial Zone (OSB). He is currently with the Faculty of Engineering, Van Yüzüncü Yıl University, with a focus on computer engineering. He created and executed numerous IT projects. His research interests include cyber security, big data, deep learning, data mining, the Internet of Things, and network security.

• • •