Intelligent Video Surveillance System with Abnormal Behavior Recognition and Metadata Retrieval

Hyungtae Kim , Joongchol Shin , Seokmok Park , Joonki Paik *

Posted Date: 2 June 2023

doi: 10.20944/preprints202306.0147.v1

Keywords: Metadata generation; abnormal recognition; metadata retrieval; intelligent surveillance system

Article

Intelligent Video Surveillance System with Abnormal Behavior Recognition and Metadata Retrieval

Hyungtae Kim [1], Joongchol Shin [1], Seokmok Park [1] and Joonki Paik [1,2,*]

[1] Department of Image, Chung-Ang University, Seoul, 06974, Korea
[2] Department of Artificial Intelligence, Chung-Ang University, Seoul, 06974, Korea; hyungtae@ipis.cau.ac.kr (H.K.)
* Correspondence: paikj@cau.ac.kr (J.P.)

Abstract: Huge-scale video surveillance systems have become essential in crime prevention and situation recording. Traditional surveillance systems relied on human monitoring of video streams, which often led to errors and difficulties in understanding events. Furthermore, locating specific scenes within recorded videos required extensive human investigation. To overcome these challenges of inefficiency, inconvenience, and potential risks, we propose an intelligent analysis scheme that utilizes abnormal behavior recognition and metadata retrieval algorithms to replace human monitoring. The proposed method consists of three stages: i) metadata generation through object detection and tracking, ii) abnormal behavior recognition, and iii) SQL-based metadata retrieval. By incorporating specific information such as object color and aspect ratio, our technique enhances the usability of retrieval. Moreover, our abnormal behavior recognition module demonstrates robust classification capabilities for activities such as pushing, violence, falling, and crossing barriers. The proposed method can be seamlessly deployed on both edge cameras and analysis servers, making it adaptable to various surveillance setups. This approach revolutionizes the traditional surveillance paradigm, enabling more efficient, reliable, and secure video monitoring and analysis.

Keywords: metadata generation; abnormal recognition; metadata retrieval; intelligent surveillance system

1. Introduction

Video surveillance systems play a crucial role in preventing crimes, ensuring public safety, and capturing critical incidents. However, traditional surveillance systems heavily rely on human operators to monitor video streams, which can be prone to errors, inefficiencies, and limitations in real-time analysis. To overcome these challenges, intelligent video surveillance systems have emerged as a promising solution, leveraging advanced technologies such as object detection, tracking, abnormal behavior recognition, and metadata extraction and retrieval.

Intelligent video surveillance systems have witnessed remarkable advancements in object detection and tracking techniques, leading to automation and enhanced monitoring capabilities. State-of-the-art technologies have significantly improved the accuracy and efficiency of surveillance systems. Region-based methods, such as Faster R-CNN [1] and YOLO [2], have gained considerable attention for their ability to detect objects with high precision and real-time performance. Motion-based approaches, including optical flow [3] and Kalman filters [4], have been widely utilized to track objects in dynamic scenes with robustness and accuracy.

Additionally, other notable techniques in object detection and tracking are summarized as follows. Liu *et al*. proposed a real-time object detection method, called Single Shot MultiBox Detector (SSD), that achieves high accuracy by utilizing multiple layers with different resolutions [5]. Tan *et al*. proposed an efficient and scalable object detection framework called EfficientDet that achieves state-of-the-art performance by optimizing the trade-off between accuracy and computational efficiency [6]. Wojke *et al*. proposed a deep learning-based object tracking algorithm called DeepSORT that combines appearance information with motion cues for robust and accurate tracking [7]. Zhou *et al*. proposed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 June 2023 doi:10.20944/preprints202306.0147.v1

a tracking-by-detection approach called CenterTrack that incorporates object center estimation to improve tracking performance in crowded scenes [8].

Abnormal behavior recognition plays a vital role in intelligent video surveillance systems, enabling the identification and classification of unusual activities or events that deviate from normal patterns. Extensive research has been conducted in this field, utilizing machine learning algorithms, deep neural networks, and statistical models. Several notable works include the use of convolutional neural networks (CNNs) for abnormal event detection [9], anomaly detection based on spatiotemporal features [10], and behavior modeling using hidden Markov models (HMMs) [11].

Metadata extraction and retrieval techniques complement the intelligent video surveillance system by facilitating efficient data management and rapid scene retrieval. Metadata encompasses essential information such as object attributes, timestamps, and spatial coordinates, providing valuable context for analyzing and searching video content. Various approaches have been proposed for metadata extraction, including object-based methods [12], feature extraction algorithms [13], and deep learning-based techniques [14]. In terms of retrieval, SQL-based querying has proven effective in enabling fast and accurate retrieval of relevant scenes based on specific criteria.

The proposed intelligent video surveillance system integrates object detection and tracking, abnormal behavior recognition, and metadata extraction and retrieval modules as shown in Figure 1. These components collectively form a comprehensive framework that enhances surveillance capabilities and streamlines the analysis process. The system architecture encompasses a network of edge cameras for real-time monitoring and an analysis server for centralized processing and storage. This scalable and adaptable architecture ensures seamless integration with existing surveillance infrastructure, maximizing the system's efficiency and effectiveness.

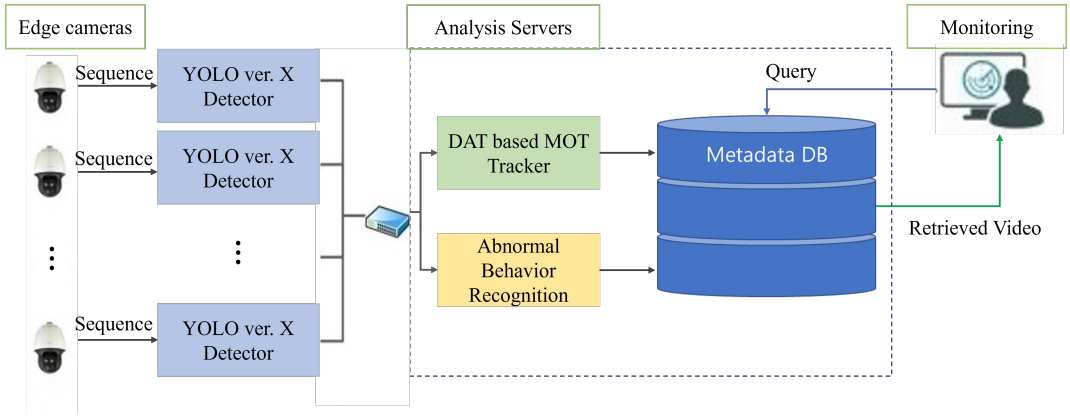

Figure 1. Overview of the proposed system. The proposed method is divided into three modules; i) edge cameras with detection, ii) Analsys servers including MOT tracker and Abnormal behavior recognition, and iii) Monitoring for video retrieval based on a query.

This paper is organized as follows. In Section 2, we introduce the proposed system and provide an overview of how representative basic metadata is generated. Section 3 focuses on event metadata generation through the recognition of abnormal behaviors. We explore the techniques and algorithms employed to detect and classify these abnormal behaviors, enabling the extraction of event-specific metadata. In Section 4, we delve into the design of the metadata structure and discuss the methodologies used for efficient video retrieval. In Section 5, we present the experimental results obtained through the implementation of our proposed system. Finally, in Section 6, we draw conclusions based on the findings presented throughout the paper.

2. Proposed System

In this section, we introduce the general procedures of the proposed surveillance systems, including object detection and tracking, with the aim of generating representative metadata for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 June 2023 doi:10.20944/preprints202306.0147.v1

video retrieval. As shown in Figure 1, the proposed system consists of three parts: i) edge cameras, ii) analysis servers, and iii) Monitoring. Each part plays a specific role as follows. Edge cameras are in charge of object detection when objects appear and generate low metadata which is a seed for tracking and simple query. In analysis servers, the tracker groups detection results and extracts trajectory and representative metadata which are common information about the object such as the color of clothes and size. Anomaly recognition detects an anomaly scene and classifies the cases of abnormal behaviors which are compartmentalized into normal and event metadata. In general, generated metadata is stored in the database. Exceptionally, when the query feed into the analysis servers, the retrieved video from the database is returned to the monitoring system.

2.1. Object Detection and Tracking

As aforementioned, there are diverse object detection and tracking methods. Since our framework focuses on video retrieval which qualifies as applicable to both edge cameras and analysis servers, when we design the proposed system, we consider the three conditions; i) the detector should detect not only normal size objects but also small objects with the lowest false positive, ii) the execution of the edge camera should be in real-time without hardware constraints, and iii) the tracker should associate between detected objects temporally and extract the metadata at the same time.

In addition, practical surveillance cameras have a mount of kinds, where the cameras have differences in various perspectives such as field of view(FoV), resolution, computational power, frame rate, pan, tilt, zoom, and so on. These factors affect the performance of the detector. Hence, our detector is selected to satisfy our considerations mentioned above on those cameras. The most popular used object detection methods are categorized according to their base algorithm as RCNN-based, YOLO-based, and SSD-based. The RCNN-based algorithms have a two-stage framework and are slower than the others. The SSD-based method is faster than RCNN-based detection, but misdetection occurs when small objects exist in the scene. YOLO detection is also a one-stage method similar to the SSD method. it is an advantage to applying the surveillance system handled in our cases. Consequently, we adopt the YOLO-based detection method as our detector. the detailed version of YOLO is determined by evaluating the performance of the edge camera and practical monitoring scenarios. The YOLO versions 3, 5, and 7, and their modified submodules are available for use [15–17].

The detected object, denoted as O, is represented as follows:

$$O = \{O_{id}, x_l, y_t, w, h\}, \tag{1}$$

where O_{id}, x_l, y_t, w, b are object id, the top left coordinates, width, and height of the bounding box, respectively.

Edge camera encodes detection results as

$$C = \{C_{id}, t, fn, O_{id}, x_l, y_t, w, h\}, \tag{2}$$

where C_{id}, t and fn represent the camera id and time, frame number.

These encoded data are called low metadata and after the encoding, low metadata are fed into analysis servers with the video stream. The low metadata collected has redundant information because the camera captures the image at least 24 fps and up to 30 fps. Thus, the enhancement of both the representativeness of the metadata and the efficiency of retrieval is achieved by applying a proper object tracker.

As we mentioned at the head of this section, the objects tracking module in analysis servers associates among the low metadata and extracts the fine-grained metadata at the same time. Hence we select the scale-adaptive version of the distractor-aware tracker(DATs) as our tracker [18]. The DAT method uses the color histogram of the object,O, and its surrounding region,S. The scale adaptive ability of DAT is helpful to apply to our systems because most cameras have viewpoint changes that cause the scale variation of objects, and the proposed metadata includes the aspect ratio related to scale.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 June 2023 doi:10.20944/preprints202306.0147.v1

We define the basic tracking procedures for applying the DAT tracker. Tracking procedures are organized with i) birth of the tracker, ii) update, and iii) death of the tracker. When the low metadata is fed into the tracking module, initializing the tracker, is called tracker birth. The tracker birth,O_t, is defined with (1).

$$\begin{aligned} O_t &= \arg max(s_v(O)s_d(O)), \\ s_v(O) &= \sum_{x \in O} P(x \in O|b_x), \text{ and} \\ s_d(O) &= \sum_{x \in O} \exp(-\frac{\|x - c_{t-1}\|^2}{2\sigma^2}), \end{aligned} \tag{3}$$

where $s_v(\cdot)$ is the voting score and $s_d(\cdot)$ is the distance score from object center,c. b_x is the bin assigned to the color components of the detected object.

In the tracking process, the tracker is updated as follows

$$\begin{aligned} O_t^{new} &= \arg max(s_v(O_{t,i})s_d(O_{t,i})), \\ s_v(O_{t,i}) &= \sum_{x \in O_{t,i}} P_{1:t-1}(x \in O|b_x), \\ s_d(O_{t,i}) &= \sum_{x \in O_{t,i}} \exp(-\frac{\|x - c_{t-1}\|^2}{2\sigma^2}), \end{aligned} \tag{4}$$

where $O_{t,i}$ are detected objects as candidates to track. c_{t-1} is the center of the previous object.

The death of the tracker occurs when the tracked object disappears. In the Death of Tracker, low-level metadata are fine-grained and representative metadata are generated. The trajectory of the object,T_O, is estimated from a sequence of detected objects by the tracker and defined as;

$$T_O = \{x_s, y_s, x_{1/3}, y_{1/3}, x_{2/3}, y_{2/3}, x_e, y_e\}, \tag{5}$$

where x and y are horizontal and vertical coordinates. subscripts $s, 1/3, 2/3$ and e represent the start, $1/3$, $2/3$ and end points of trajectory, respectively. The trajectory becomes one of representative metadata for retrieval, and the bounding box information is stored separately. The precise algorithms are described in the following subsection.

2.2. Color and Aspect Ratio Metadata

In this subsection, the representative methods of generating color and ratio metadata are explained. To extract the representative color, we adopt the probabilistic latent semantic analysis(PLSA) based generative model [19]. PLSA-based model is trained with the Google images data set to generate the color distribution. The object color distribution is accumulated by the update process of the tracker. As a result, we can easily extract the representative color by comparing the object distributions with the reference distributions. The PLSA model-based color extraction is robust to changing illumination and viewpoint because the training of the reference distributions contains a brightness range from low to high. The number of proposed representative colors is 11. The colors are black, blue, brown, gray, green, orange, pink, purple, red, white, and yellow.

The aspect ratio is estimated by normalizing the integration of width and height in the tracking process. The aspect ratio,R_O, is denoted as;

$$R_O = \frac{H_O}{W_O} = \frac{\sum_{H_i \in O_{s,e}} H_i}{\sum_{W_i \in O_{s,e}} W_i}, \tag{6}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 June 2023 doi:10.20944/preprints202306.0147.v1

where H_O and W_O are normalized height and width of the object. H_i and W_i are height and width of $i-th$ which is contained in trajectories. The reason for using aspect ratio as the metadata is the aspect ratio has less effect than size by the change of camera resolution without preprocessing such as calibration. Furthermore, (6) implies the normalized aspect ratio. In both the denominator and numerator, the number of tracked frames is simplified.

The trajectory, representative color, and aspect ratio configure the basic metadata. These metadata are used for query-based retrieval. For instance, each of the metadata is respectively used to generate a single query, and also a combined metadata query is available. Single and combined query-based retrieval is generally used in searching for normal scenes. Since object detection and tracking are designed with the assumption that the appearance and behavior of objects follow common sense, this metadata-based retrieval is difficult to specify the abnormal scenes. Furthermore, the proportion between normal and abnormal is extremely dominated by normal. So, the metadata query-based abnormal search is inefficient. In order to solve this problem, we propose anomaly recognition to produce event metadata.

3. Event Metadata Generation by Abnormal Behavior Recognition

We propose the two-stream convolution neural networks based abnormal behavior recognition. The structure of the proposed abnormal recognition is demonstrated in Figure 2.

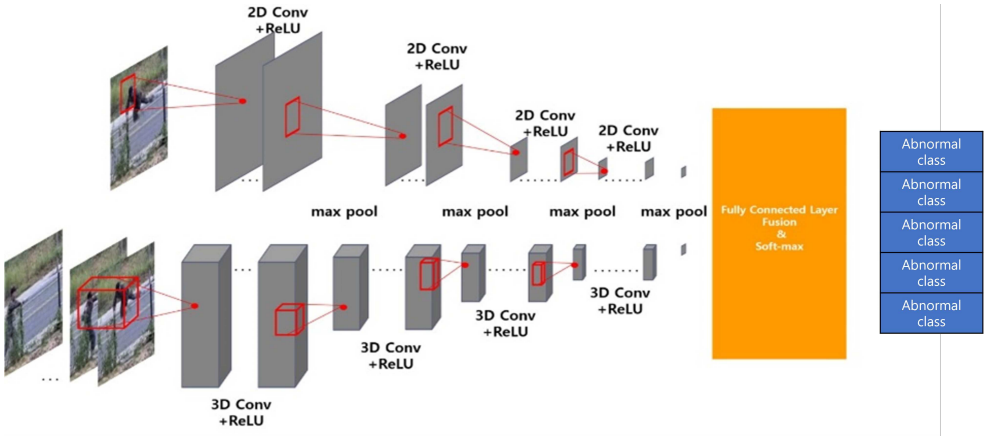

Figure 2. The network architecture of the proposed two-stream abnormal recognition; upper and lower streams are used for spatial and temporal, respectively.

The proposed network fuses the 2D-CNN [20,21] and 3D-CNN [22,23] architecture. In the 2D-CNN stream, the input video converts to the image by frame morphing, as shown in Figure 3. The spatial and temporal feature vectors are concatenated by using fully connected layer fusion.

Figure 3. The example of frame morphing.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 June 2023 doi:10.20944/preprints202306.0147.v1

For training the proposed network, we collect the large-scale behavior dataset. This dataset consists of 16,252 frames that are captured by surveillance cameras. the proposed anomaly recognition can classify the four categories; pushing(assault), falling, crossing wall, and normal.

Figure 4. The example frames of the large-scale abnormal recognition dataset.

Behavior recognition reduces the challenging situation of tracking, such as assault. In an assault situation, at least two or more objects are located too close. Close objects contaminate the object model for tracking and lost the reliability of metadata. Not only assault but also other abnormal behaviors affect the tracking and retrieval, hence we designed the parallel structure to handle these weaknesses of metadata-based retrieval.

4. Metadata Structure Design and Video Retrieval

This section introduces the SQL-based metadata structure and the specific video retrieval according to the query. SQL is specialized in query-based programming. For this reason, we exploit SQL as our database management program. The proposed metadata structure has a hierarchy that consists of a common DB as a parent and two children DB including object DB and abnormal db. The separated hierarchical metadata structure is more efficient and faster than a single metadata structure. Because when we search the clip with a specific query, the separated structure can concentrate the query by blinding other data. Table 1 shows the parent structure of the proposed SQL-based metadata. As shown in Table 1, all detected objects' data are stored in the parent database as low-level data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 June 2023 doi:10.20944/preprints202306.0147.v1

Table 2 represents the trajectories generated by the tracker. The conciseness of the database in Table 2 is higher than the structure in Table 1. Table 3 shows the event metadata structure.

Table 1. The parent database structure of the proposed metadata.

Order	Cam_id	Fr_no	Obj_id	x	y	w	h	Color1	Color2	Color3
0	1	1	0	264	347	137	412	1	4	2
1	1	1	1	1485	446	357	178	1	4	2
2	1	1	2	352	476	239	150	1	2	4
3	1	2	0	275	344	129	416	1	2	4
4	1	2	1	1484	447	358	177	1	2	3
5	1	2	2	360	475	238	150	1	4	9
6	1	3	0	284	343	122	414	1	2	4
7	1	3	1	1484	446	356	178	1	2	3
8	1	3	2	381	470	197	163	4	1	7
9	1	4	0	295	341	123	414	1	4	2
10	1	4	1	1487	445	355	180	1	9	7
11	1	4	2	1310	361	277	174	4	2	3

Table 2. Configuration of object database structure in the proposed metadata.

Obj_id	Cam_id	Start_fr	End_fr	w	h	Color1	Color2	Color3
0	0	96	229	130	171	5	9	4
1	0	53	112	79	145	5	8	4
2	0	265	394	81	118	5	8	7

Table 3. Example of the proposed abnormal database structure.

Obj_id	Cam_id	Fr_no	x	y	w	h	Pushing	Falling	CrossingWall
1	0	321	1416	261	79	145	1	0	0
1	0	322	1392	270	78	144	1	0	0
1	0	323	1354	294	79	144	1	0	0
1	0	324	1362	294	78	144	1	0	0
1	0	325	1358	296	78	144	1	0	0
1	0	326	1354	296	78	144	1	0	0
1	0	327	1342	310	78	144	1	0	0
1	0	328	1338	298	78	144	1	0	0
1	0	329	1330	306	78	144	1	0	0
1	0	330	1270	332	78	144	1	0	0
1	0	331	1274	332	78	144	1	0	0
1	0	332	1282	318	78	144	1	0	0
1	0	333	1278	328	78	144	1	0	0
1	0	334	1272	328	78	144	1	0	0
1	0	335	1274	326	78	144	1	0	0
1	0	336	1278	320	78	144	1	0	0
1	0	337	1289	314	78	144	1	0	0

Figure 5 represents our retrieval system configuration. The combination of queries is activated by checking the check button. The detailed query option is able to fill out the box or select the color. After the query is set up, the search results are listed in the bottom left. When selecting the video path, the summarized video is played on the right panel.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 June 2023 doi:10.20944/preprints202306.0147.v1

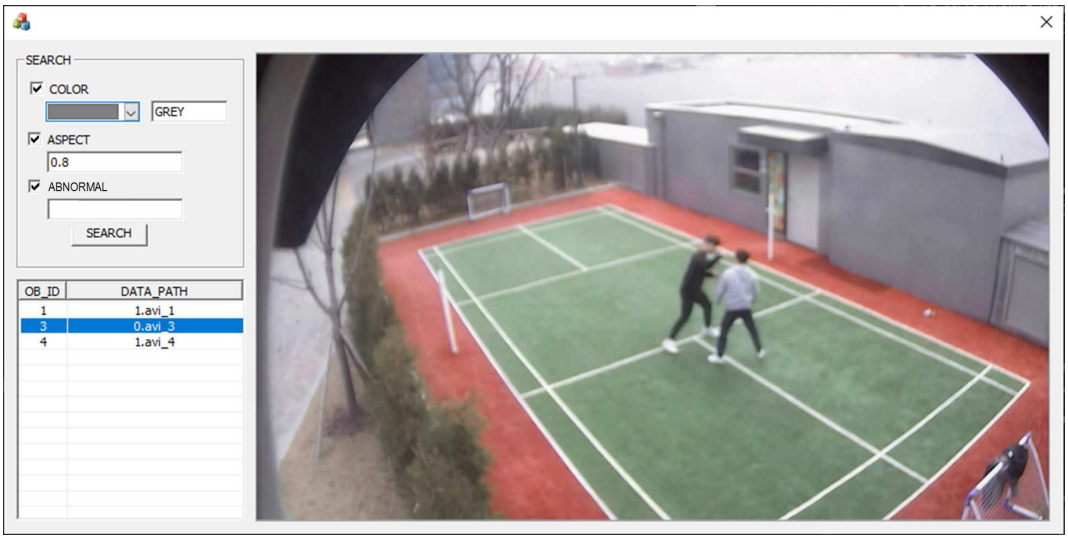

Figure 5. The example of the proposed retrieval system.

5. Experimental Results

In this section, the performance of the proposed method is demonstrated in either a qualitative or quantitative manner. Figure 6 shows the configuration of the validation test images for the color extraction. As shown in Figure 6, for the validation, we use the Frankfurt DB which is organized by gathering the images captured in the Frankfurt marathon. The arbitrary eight points are selected on the representative color in an object. This experiment shows the robustness of the color extraction in illumination changes. Selected eight points have variants in not only illumination conditions but also color space. Figure 7 presents the color extraction results with up to rank 3. In Figure 7, even though the white experiments include the looks like pink, the representative color is correctly extracted. Figure 7 demonstrates that even if looks like there are no color changes, the actual color has diverse variants.

The quantitative result is produced in Figure 8. The accuracy of color-based retrieval is calculated as;

$$Accuracy = \sum \frac{Results}{GT} \times 100, \tag{7}$$

where the results are extracted results. In Figure 8, T, F, and A are true, false, and accuracy, respectively. Colors 1, 2, and 3 indicate the rank of extraction. The proposed color-based method achieves 86.23%.

	Black	Blue	Brown	Gray	Green	Orange	Pink	White	Red	Purple	Yellow
Color											
Height (px)	461	543	300	246	463	319	146	495	506	480	450
Width (px)	184	218	109	150	185	127	365	198	202	189	117

Figure 6. The configuration and information of validation images for the representative color extraction.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 June 2023 doi:10.20944/preprints202306.0147.v1

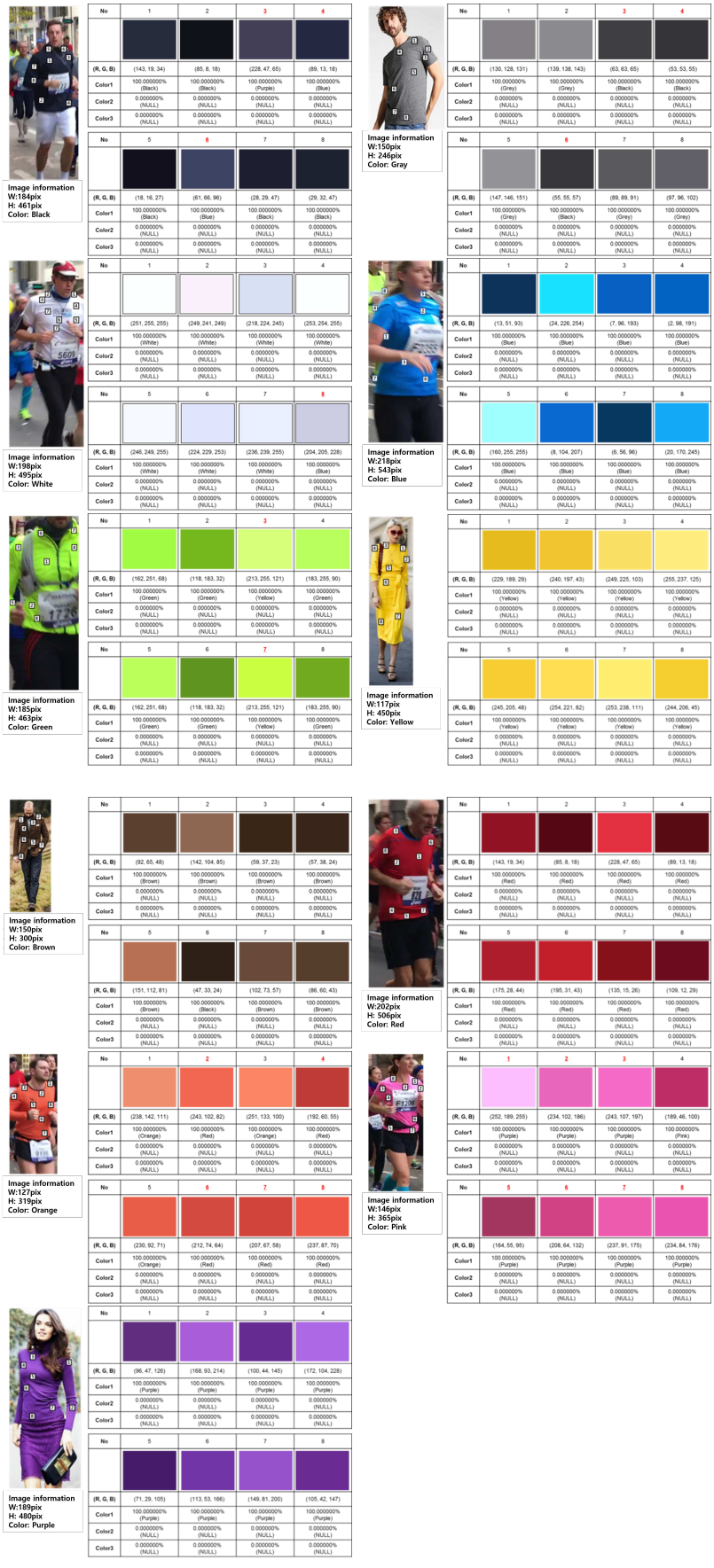

Figure 7. The results of the representative eleven-color estimation in arbitrary eight points in each image.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 June 2023 doi:10.20944/preprints202306.0147.v1

	Black			Blue			Brown			Gray			Green			Orange		
	T	F	A(%)	T	F	A(%)	T	F	A(%)	T	F	A(%)	T	F	A(%)	T	F	A(%)
COLOR1	39	2	93.1	46	31	59.7	0	0	0	9	15	37.5	44	20	68.7	3	17	15.0
COLOR2	40	1	97.5	69	8	86.6	0	0	0	11	13	45.8	58	6	90.6	6	14	30.0
COLOR3	40	1	97.5	74	3	96.1	0	0	0	22	2	91.6	59	5	92.1	7	13	35.0

	Pink			Purple			Red			White			Yellow			Total Accuracy
	T	F	A(%)	T	F	A(%)	T	F	A(%)	T	F	A(%)	T	F	A(%)	
COLOR1	6	23	20.6	1	1	50.0	34	33	50.7	0	3	0	0	0	0	
COLOR2	18	11	62.0	2	0	100	50	17	74.6	0	3	0	0	0	0	86.23%
COLOR3	22	7	75.8	2	0	100	56	11	83.5	0	3	0	0	0	0	

Figure 8. The Quantitative results of the color retrieval.

Figure 9 shows the results of abnormal behavior recognition. The percentage implies the confidence of the classified class. Three cases are correctly recognized.

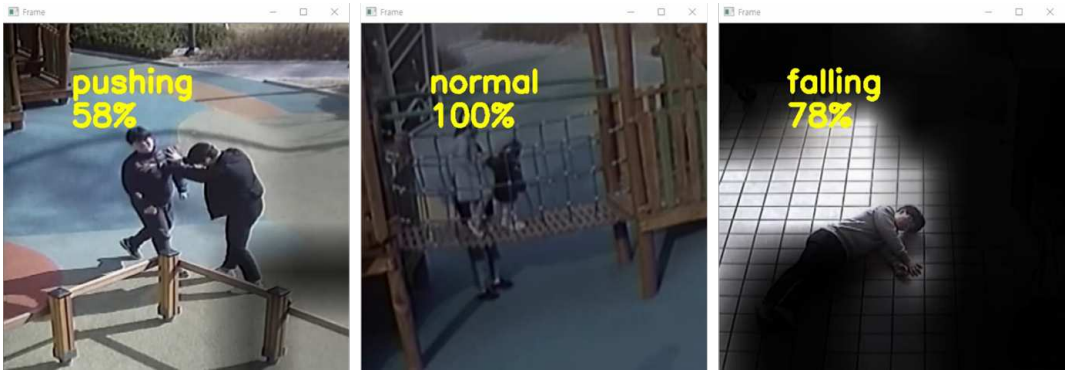

Figure 9. The information of validation images.

Table 4 presents the recognition accuracy of the proposed method. Accuracy is estimated using (7). In Table 4, the first row indicates the recognition result, and the first column presents the video categories. Table 4 shows that the proposed method achieves 88.55%.

Table 4. The accuracy of abnormal behavior recognition.

Results	Pusshing	Falling	CrossingWall	Noraml	Count
Pushing	120	3	1	12	136
Falling	0	71	0	9	80
CrossingWall	0	0	72	8	80
Accuracy					88.85%

6. Conclusions

This paper introduced an intelligent surveillance system incorporating anomaly recognition and metadata-based retrieval, aimed at efficiently and accurately retrieving specific objects. A key highlight of the proposed system is its modular design, which improves portability and adaptability. The color and aspect ratio extraction method showcased in this work demonstrates its versatility in both detection and tracking tasks. Furthermore, the standalone abnormal recognition module provides additional flexibility.

Moving forward, future research could focus on exploring instance segmentation as a potential replacement for object detection, thereby improving the accuracy of metadata extraction. Additionally,

 doi:10.20944/preprints202306.0147.v1

a comprehensive survey of tracking methods could be undertaken to consolidate tracking and recognition networks, leading to enhanced system performance. By further advancing these areas, the proposed intelligent surveillance system can be strengthened, offering increased precision and reliability for object retrieval applications.

Overall, this paper has laid the foundation for an intelligent surveillance system that combines anomaly recognition, metadata-based retrieval, and modular design. The suggested future directions open up exciting opportunities for further advancements, contributing to the development of even more robust and effective surveillance systems in the field.

Author Contributions: Conceptualization, Kim, H. and Park, S.; methodology, Kim, H.; software, Shin, J.; validation, Shin, J. and Park. S; formal analysis, Kim, H.; investigation, Paik, J.; resources, Paik, J.; data curation, Shin J. and Kim. H.; writing—original draft preparation, Kim H.; writing—review and editing, Paik, J.; visualization, Kim, H.; supervision, Paik, J.; project administration, Paik, J.; funding acquisition, Paik, J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT). [No.2014-0-00077, Development of global multi-target tracking and event prediction techniques based on real-time large-scale video analysis and 2021-0-01341, Artificial Intelligence Graduate School Program(Chung-Ang University)].

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to patent issue.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **2015**, *28*.
2. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
3. Horn, B.K.; Schunck, B.G. Determining optical flow. *Artificial intelligence* **1981**, *17*, 185–203.
4. Kalman, R.E. A new approach to linear filtering and prediction problems **1960**.
5. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, 2016, pp. 21–37.
6. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10781–10790.
7. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. 2017 IEEE international conference on image processing (ICIP). IEEE, 2017, pp. 3645–3649.
8. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv preprint arXiv:1904.07850* **2019**.
9. Sakurada, M.; Yairi, T. Anomaly detection using autoencoders with nonlinear dimensionality reduction. Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis, 2014, pp. 4–11.
10. Mahadevan, V.; Li, W.; Bhalodia, V.; Vasconcelos, N. Anomaly detection in crowded scenes. 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, 2010, pp. 1975–1981.
11. Chan, A.B.; Vasconcelos, N. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE transactions on pattern analysis and machine intelligence* **2008**, *30*, 909–926.
12. Sivic, J.; Zisserman, A. Video Google: Efficient visual search of videos. *Toward category-level object recognition* **2006**, pp. 127–144.
13. Laptev, I. On space-time interest points. *International journal of computer vision* **2005**, *64*, 107–123.
14. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
15. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* **2018**.
16. Glenn-Jocher, Y.F.L. 3181. *Ultralytics: Github; https://github. com/ultralytics/yolov5/discussions/3181m1* **2021**.
17. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696* **2022**.

18. Possegger, H.; Mauthner, T.; Bischof, H. In defense of color-based model-free tracking. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2113–2120.
19. Van De Weijer, J.; Schmid, C.; Verbeek, J.; Larlus, D. Learning color names for real-world applications. *IEEE Transactions on Image Processing* **2009**, *18*, 1512–1523.
20. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
21. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; others. Imagenet large scale visual recognition challenge. *International journal of computer vision* **2015**, *115*, 211–252.
22. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.
23. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.