ISSN 1870-4069

Crime Prediction Using Computer Vision: Data Selection, Balancing, and Representation

L. Enrique Morales-Márquez, J. Arturo Olvera-López, Ivan Olmos-Pineda

Autonomous University of Puebla, Faculty of Computer Science, Puebla, Pue., Mexico
luise.morales@viep.com.mx,
{jose.olvera,ivan.olmos}@correo.buap.mx

Abstract. Crime prediction and detection in video surveillance is a task currently under exploration. In various research papers, the ease of use of many novel video processing models and the focus on developing more powerful predictive architectures have relegated the task of data quality preprocessing to a small piece. These include everything from video set selection to analyzing the relationships between variables, tasks that impact model performance. This article presents preliminary results focused on the proper selection and processing of data from an ongoing crime prediction investigation. The paper presents the selection of the data set, the reasons for its selection, the representation of extracted data, data augmentation, and a method of data balancing that preserves data structure. It also presents a brief description of the method for determining dependencies between variables.

Keywords: Crime Prediction, Data Preprocessing, Computer Vision.

1 Introduction

Automatic crime prediction has become an important area of research in security and computer vision, with applications in public safety and crime prevention. However, current research faces significant challenges, including variability in environmental conditions such as lighting or camera angles; inherent imbalance in the ratio of normal events to crimes; and the movement of criminals, which can be very similar to that of a person with no malicious intent.

Recently, detection methods based on the analysis of human skeletons have emerged as a promising approach, as they offer a realistic representation of people's bodily behavior, at least visually. However, the quality of any investigation depends largely on preliminary steps that are often overlooked, such as the appropriate selection of a data set, the human pose extraction model, or class balancing. This article presents the preliminary stages of experimentation, focusing on data preprocessing. It is emphasized that early-stage decisions are just as crucial as the selection of a classification or prediction model or architecture.

This paper is structured as follows: Section 2 presents related work in computer vision-based crime prediction, subsequently, section 3 shows the methodology carried out so far and then, in section 4, the preliminary results are shown, which are discussed in section 5. Finally, section 6 contains the conclusions.

2 Related Work

Work on crime prediction in video has already been carried out; however, we often tend to focus on detailing the architecture or algorithm that will perform the detection, often neglecting to provide much detail in the papers or overlooking data preprocessing tasks. In [1], shoplifting prediction is performed using three-dimensional convolutional models (3DCNN) using the UCF–Crime Dataset [2] and obtaining an accuracy of 0.86. However, the preprocessing of the video clips consists of separating pre-crime timestamps, suspect behavior, and crime evidence, in addition to reducing the frame size to reduce computation time, but without taking other video characteristics into account.

Another similar treatment of video is found in [3], where frames from the CAVIAR dataset [4] are flagged for attention and converted to grayscale for feature extraction and subsequent detection of suspicious activity using VGG16 and Bidirectional Gated Recurrent Units (BiGRU) models. Accuracy, recall, and precision of 0.98, F1 of 0.97, and AUC of 0.99 are obtained, but the limitations and characteristics of the dataset are not taken into account, and efforts are focused on fine-tuning the data to feed intelligence models.

In [5], shoplifting is estimated using a fuzzy logic model that evaluates people's gait, crowd size, and degree of facial coverage of actors in videos from the UCF–Crime dataset. However, the facial coverage detection module is a retraining of the You Only Look Once (YOLO) model [6] that is fed with another dataset of helmets and face masks, to which a Gaussian filter is applied to remove artifacts within the frames, a single correction applied to the data. The fuzzy model achieves a precision of 0.77, a recall of 0.50, and an F1 of 0.60 value, performance that has room for improvement.

Currently, the convolutional model approach is commonly applied to crime prediction since these architectures are focused on image processing, it is common to find models that require minimal or no adjustment to the frames in order to be processed. However, an adequate selection and preprocessing steps have an impact on the prediction quality of the architecture being built.

3 Methodology

A methodological framework is essential to ensure the validity of any predictive model. Specifically, the process for processing the visual data in this work consists of dataset and Human Pose Estimation model, data augmentation for cases where a sufficient number of videos is unavailable. Subsequently, instance selection is performed to maintain class balance and, finally, determine the dependence or independence between variables for future training via Bayesian prediction models.

3.1 Dataset Selection

Dataset selection is a critical step in the development of prediction systems, as model generalization depends on it. As noted in [7], one of the factors attributed to the success of novel models is the availability of more and better data, and the effects of noisy or low-quality data must be taken into account when training algorithms.

Among the datasets available for crime detection, the UCF-Crime Dataset stands out for its widespread use in related studies. It contains approximately 1,900 video clips distributed across 13 categories that provide a variety of scenarios. However, limitations such as the maximum resolution of 320 x 240 pixels and poor image quality in almost all videos impact feature extraction.

To address these limitations, the Shoplifting Dataset 2022 [8] was selected, which provides certain advantages such as better resolution videos, balance with 92 shoplifting sequences versus 90 normal sequences, and controlled camera conditions at the cost of reducing the number of samples and limiting it to shoplifting. The general characteristics of both datasets are shown in Table 1.

Table 1. Characteristics of datasets for crime detection in video.

Dataset	Number of videos	Resolution	Crimes	Shooting angle	Frame Rate	Occlusion of actors
UCF– Crime	1,900	320x240p	Abuse, Burglary, Robbery, Stealing, Shooting, Shoplifting, Assault, Fighting, Arson, Vandalism	Zenith Street level High angle	30 frames per seconds	Partial, Total
Shoplifting 2022	182	640x480p 1920x1080p	Shoplifting	Street level	30 frames per second	None

While the UCF-Crime dataset contains a larger number of videos and a wider variety of crimes, the skeleton extraction models struggled to perform their task satisfactorily, a problem not encountered with higher-resolution videos. Furthermore, the absence of occlusions in the Shoplifting 2022 dataset allows for a clearer view of keypoint behavior, and the focus on a single crime allows for fine-tuning the predictive models for a single case to maximize their performance.

As shown in Figure 1, the improvement in quality allows MMPose [9] to consistently detect skeletons.

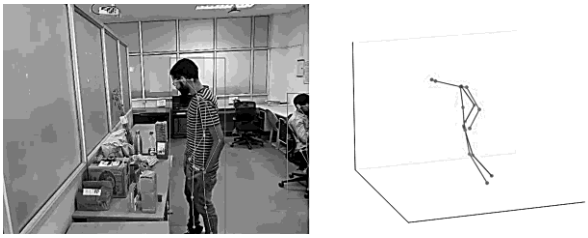

Fig. 1. Adequate human keypoint detection.

3.2 Representation of Joints

Since the two-dimensional representation of a person's skeleton may not be reliable due to scale or distance from the camera, it was decided to obtain the people's joints in three-dimensional space using MMPose 3D inference module. This module works by using 2D joints obtained using HRNet and processing them with a Meta Research model [10]. Keypoints are represented as a coordinate vector relative to an origin point located in the pelvis.

This architecture converts 2D keypoint sequences using residual blocks that capture temporal dependencies to generate maps for 3D skeletons and projects the results back into 2D for comparison with the initial skeleton. This process includes biomechanical constraints such as bone lengths or improbable positions when calculating 3D key-points. This process is shown in Figure 2.

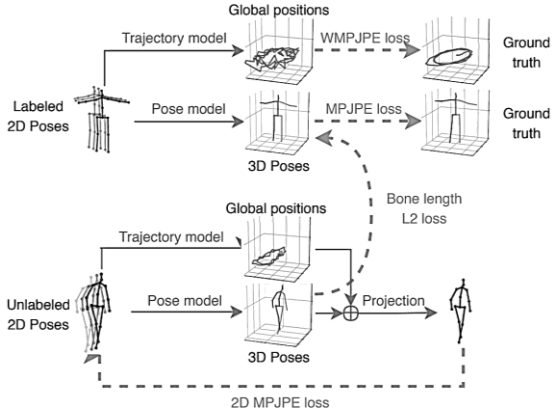

Fig. 2. Process of transforming 2D to 3D joints. Figure taken from [10].

As previously mentioned, tracking points within space may not be reliable; therefore, angles formed by angles formed by groups of three joints are calculated, taking one of them as the pivot point to which the calculated angle belongs, thus avoiding potential problems related to the distance between joints, their relative position in the plane, and considering perspective, scale, the person size, and proximity to camera. Angles in degrees are calculated by solving the dot product of the line segments formed by two points A, C with a common pivot point B, obeying Formula 1:

$$\theta = \arccos\left(\frac{\overrightarrow{AB}\cdot\overrightarrow{BC}}{\|\overrightarrow{AB}\|\|\overrightarrow{BC}\|}\right). \quad (1)$$

3.3 Data Augmentation

In our experimentation, we consider a set of 15 angles formed by different joints of a person's skeleton in a single frame as an instance. For greater control over augmentation

and instance selection, the dataset was divided into separate subclasses, one for each actor appearing on camera, as well as non-mutually exclusive classes related to the camera view or the actor's behavior. Some of the subclasses have considerably fewer instances than others; furthermore, for reasons explained in section 3.5, at least 1000 instances are required for each subclass. To accomplish this task, data augmentation was performed by modifying the videos with classic transformations that can be seen in Figure 3.

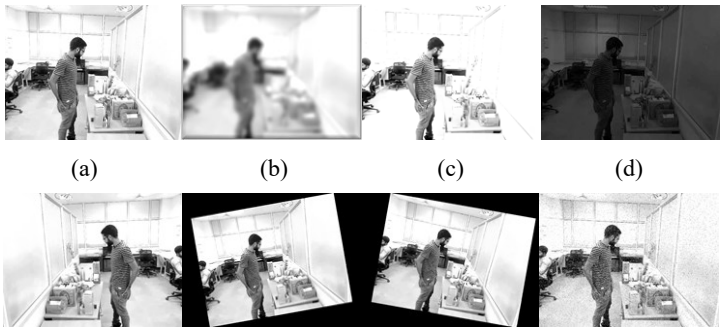

Fig. 3. a) Original image, b) Gaussian blur, $\sigma = 2.5$, c) Brightness increase to 150%, d) Brightness decrease to 10%, e) Flip relative to the y-axis, f) 10° rotation, g) -10° rotation, h) Salt and pepper noise at 5%.

The parameters for Gaussian blur, brightness modification and rotation were chosen based on the experimental results obtained in some studies reported in [11]. On the other hand, the salt and pepper noise was adjusted with a saturation of 5% to maintain sufficient similarity based on the results obtained in [12].

Gaussian blur can simulate out-of-focus frames caused by rapid movement or focus errors. Similarly, slight rotations simulate natural movements or positions because, in the real world, people and objects are rarely perfectly aligned, but they do not exhibit exaggerated inclinations. On the other hand, variations in lighting conditions describe how lighting conditions can change in videos, allowing models to detect a person's structural patterns. Furthermore, the mentioned transformations are computationally inexpensive, so augmenting the dataset does not require a lot of computational time.

Salt and pepper noise simulates unfavorable conditions that can occur in videos, such as dead pixels or interference. This is useful because it encourages intelligence models to avoid relying on data from specific regions and focus on global or contextual patterns.

These variations were carried out in order to obtain the largest possible transformation while respecting realistic variations and ensuring that the MMPose model correctly extracted keypoints. The number of instances before and after augmentation is shown in Tables 2 and 3.

3.4 Instance Selection

At this point, we have reached a problem with unbalanced classes. Note that in Tables 1 and 2, there are classes with more than 100,000 instances, while the minority classes contain 1,120 and 4,056 elements. A random selection of the same number of instances in each subclass does not guarantee that the new sample will contain members that reflect the behavior of the subclass.

Some instance selection methods based on clustering have shown satisfactory results, preserving the structure and behavior of the cluster, as in [13]. Based on this idea, instance selection is carried out using the k-Means algorithm, where k corresponds to the number of instances of the minority class; this applies to all classes.

Table 2. Data augmentation for Shoplifting class.

Subclass	Original amount	Amount after augmentation	Subclass	Original amount	Amount after augmentation
Actor 1	1,614	12,912	Backpack on	591	4,728
Actor 2	806	6,448	Backpack in hand	4,359	34,872
Actor 3	462	3,696	Remove backpack	1,448	11,584
Actor 4	704	5,632	Without backpack	3,862	30,896
Actor 5	204	1,632	Back view	140	1,120
Actor 6	1,823	14,584	Front view	1,357	10,856
Actor 7	1,455	11,640	Diagonal view from the front	6,986	55,888
Actor 8	1,061	8,488	Diagonal view from the back	2,045	16,360
Actor 9	529	4,232	Side view	6,706	53,648
Actor 10	321	2,568	Diagonal entry	1,057	8,456
Actor 11	577	4,616	Front entry	1,430	11,440
Actor 12	315	2,520	Side entry	1,196	9,568
Actor 13	201	1,608	In position	6,553	52,424
Actor 14	174	1,392			

Table 3. Data augmentation for Normal class.

Subclass	Original amount	Amount after augmentation	Subclass	Original amount	Amount after augmentation
Actor 1	5,081	40,648	Backpack on	2,785	22,280
Actor 2	1,004	8,032	Without backpack	17,134	137,072
Actor 3	5,923	47,384	Back view	2,438	19,504
Actor 4	1,858	14,864	Front view	5,580	44,640
Actor 5	793	6,344	Diagonal view from the front	16,865	134,920
Actor 6	510	4,080	Diagonal view from the back	7,278	58,224
Actor 7	838	6,704	Side view	15,281	122,248
Actor 8	1,578	12,624	Diagonal entry	1,732	13,856
Actor 9	1,215	9,720	Front entry	4,009	32,072
Actor 10	507	4,056	Side entry	1,980	15,840
Actor 11	621	4,968	In position	12,447	99,576

Starting from an initial data set, each point is relocated within a cluster whose center is the closest by calculating the average of the points within the cluster. The relocation process is repeated until a stopping criterion is met, typically a number of iterations [14] or until centroids have no changes.

An element is then selected from each cluster, resulting in a new selected set containing data that preserves the structure of the entire class but with a size that respects a perfect balance between classes. The element selected as the cluster representative is the instance closest to the centroid, this is because the k-Means algorithm could generate synthetic centroids that do not correspond to any instance and therefore are not valid elements for training.

3.5 Dependences Between Variables

One of the most common ways to determine dependence between variables is through the construction of Directed Acyclic Graphs (DAGs). However, this approach typically requires evaluating a large number of graphs, which requires a significant amount of time and computing resources when using combinatorial algorithms [15]. In this case, the *Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning* (NOTEARS) algorithm [15] is used, which allows obtaining a DAG that shows the dependence between variables.

The NOTEARS algorithm proposes an objective function that minimizes the Mean Square Error (MSE) of fit, subject to an acyclicity constraint such that:

$$h(W) = tr(e^{WoW}) - d = 0. \quad (2)$$

Where W is the matrix containing the weights of the causal relationships and d is the number of variables. This formulation employs Lagrangians to solve the optimization problem efficiently and provides a valid DAG.

Values close to zero in the weight matrix indicate a small dependence that may not be considered depending on the research purpose. Larger values suggest a strong influence and should therefore be considered. A positive value on the edges indicates a direct relationship; that is, an increase in one variable leads to an increase in the other variable. Conversely, a negative value indicates an inverse relationship, where an increase in one variable leads to decrease in the other one. It is important to mention that authors suggest using at least 1,000 instances to run the algorithm with reliable results.

4 Preliminary Results

In this section, the result of applying the NOTEARS algorithm to a data set is briefly shown, and finally, the similarity of the DAGs of the complete data sets with respect to the graphs obtained using the balanced sets is evaluated.

4.1 Dependency Graphs

NOTEARS algorithm returns a weight matrix containing all possible combinations of variable pairs and the dependency values between them. In practice, some of the

weights are 0, since this means that these edges do not exist and the acyclicity restriction is met. However, some edges obtain values close to zero. The variables included in these edges can be considered independent; however, the threshold at which this decision is made is not defined and is therefore subject to the conditions of the study being carried out.

Figure 4 shows an example of a dependency DAG for the Actor7 subclass of the Shoplifting videos, where only dependencies with a value greater than 1.5 are placed. The triple: *θKeypoint: Reference1(Side), Reference2(Side)* indicates the pivot joint for which the angle is calculated relative to the points *Reference1(Side)* and *Reference2(Side)*. Labels *L (Left)* and *R (Right)* indicate the side of the body where the joint is located.

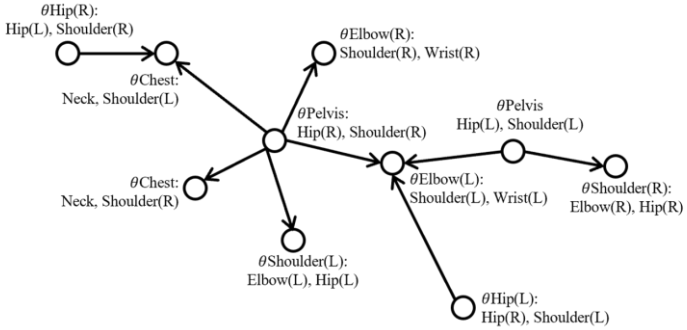

Fig. 4. Example of a dependency DAG for Actor7 in Shoplifting class. *θKeypoint* refers to the angle with vertex at the mentioned keypoint taking *Reference1(Side)* and *Reference2(Side)* as *reference.*

The frequencies of values of all the edges for the subclass Actor7 can also be seen in Figure 5. For this particular work, the weight values are bounded by the interval [-3, 3] and most of them are found in values close to 0, so the threshold could be adjusted close to 1 to discard such dependencies and consider the variables involved as independent. The rest of the dependencies that are considered strong should be considered in the next phases of the investigation.

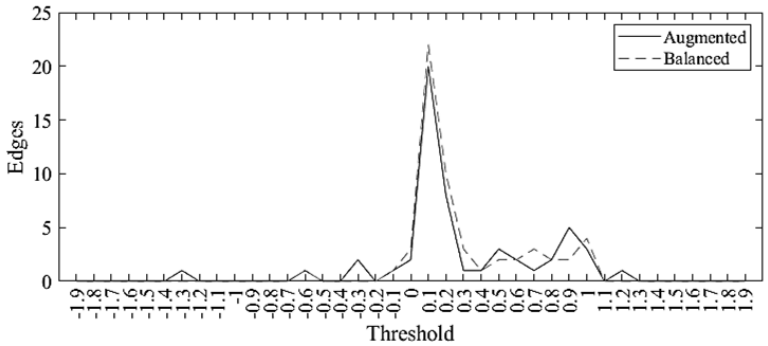

Fig. 5. Comparison of frequency of edge weights between original and balanced.

4.2 Quality of Dependency Graphs via Subsampling

In order to determine the quality of the DAG obtained using balanced classes after instance selection, the MSE is obtained between the edges of the graph with the total data and the graph with selected instances.

The minimum and maximum differences found between the edges of the graph, as well as the MSE, the average, and the standard deviation of the latter, are reported in Tables 4 and 5. Frequencies of the edges compared to original dataset are depicted in Figure 4.

The data preprocessing results shown may appear somewhat more extensive than what is typically published, but meticulous preprocessing is essential, especially for computer vision tasks. The methodology used here includes dataset selection, augmentation, instance selection, and determination of dependencies between variables; steps that have a significant impact on the quality of the results of predictive models.

The MSE values between the full dataset and the balanced subsets are low, with values of 0.0778 and 0.0479 for the Shoplifting and Normal classes, respectively, considering maximum differences of up to 3 units and almost 2.5 units. Similarly, the standard deviations remain low, with values of 0.0332 and 0.0379, demonstrating that when all balanced subclasses are considered, the dependency graphs maintain sufficient similarities to continuing the experiments.

To discard variables that could be considered dependent, an initial experiment takes into account the frequency of the edges in each interval, all the behaviors of each subclass of both classes are overlapped. Having a behavior similar to that of a normal distribution, those variables that are within the 68.3% corresponding to a standard deviation on each side of the bell are preserved, see Figure 6.

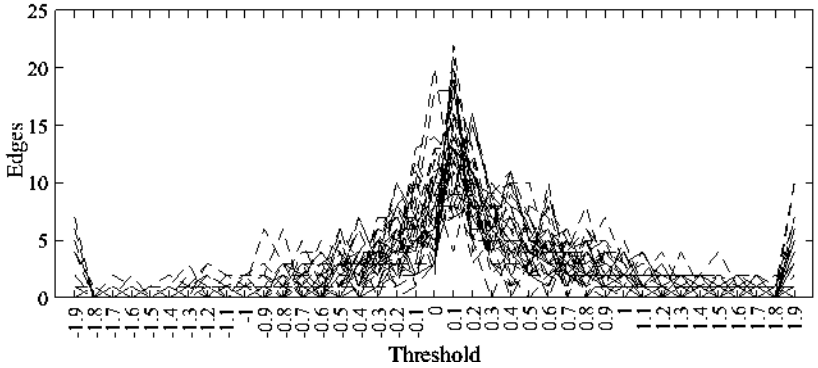

Fig. 6. Overlapping the behavior of the edges of all subclasses.

After this process, a Random Forest model is trained to classify each video frame as an image belonging to a normal sequence or one in which there is a behavior prior to the theft, this experiment obtained an accuracy of 0.74, precision, recall and F1 of 0.61, values that show a wide margin of improvement for future experiments that consider a different selection of threshold of dependency values and automatic selection of

hyperparameters for the random with the objective of improving the classification results. Confusion matrix for this experiment is presented in Figure 7.

Table 4. MSE for Shoplifting videos.

Subclass	MSE	Maximum Difference	Subclass	MSE	Maximum Difference
Actor 1	0.0895	2.2702	Backpack on	0.0771	1.5785
Actor 2	0.0393	0.9323	Backpack in hand	0.0959	1.8076
Actor 3	0.1035	2.6415	Remove backpack	0.0602	1.7603
Actor 4	0.0523	1.6001	Without backpack	0.0824	1.5439
Actor 5	0.0762	1.7344	Back view	0.1025	3.4630
Actor 6	0.0453	1.5792	Front view	0.0720	1.1674
Actor 7	0.0529	1.6119	Diagonal view from the front	0.1026	1.6953
Actor 8	0.0397	0.7361	Diagonal view from the back	0.0801	1.0507
Actor 9	0.0121	0.7015	Side view	0.1031	1.5206
Actor 10	0.0872	1.6404	Diagonal entry	0.0342	0.7173
Actor 11	0.1329	2.9279	Front entry	0.1093	2.2266
Actor 12	0.0765	1.3339	Side entry	0.1320	1.8082
Actor 13	0.1498	3.1309	In position	0.0278	0.7353
Actor 14	0.0632	1.2718	**Mean (all subclasses)**	0.0778	
			St. Deviation (all subclasses)	0.0332	

Table 5. MSE for Normal videos.

Subclass	MSE	Maximum Difference	Subclass	MSE	Maximum Difference
Actor 1	0.1039	2.4128	Without backpack	0.0737	1.8061
Actor 2	0.0156	0.5990	Back view	0.0388	1.7170
Actor 3	0.0651	1.9416	Front view	0.0814	1.3664
Actor 4	0.0936	3.1229	Diagonal view from the front	0.0908	1.7113
Actor 5	0.0061	0.4405	Diagonal view from the back	0.0383	2.0397
Actor 6	0.0015	0.3816	Side view	0.0867	1.8049
Actor 7	0.0058	0.3944	Diagonal entry	0.0528	1.2931
Actor 8	0.0654	1.5684	Front entry	0.1294	1.6946
Actor 9	0.0393	1.1893	Side entry	0.0124	0.6764
Actor 10	4.9967e-06	0.0123	In position	0.0400	1.3990
Actor 11	0.0035	0.3790	**Mean (all subclasses)**	0.0479	
Backpack on	0.0102	0.5600	**St. Deviation (all subclasses)**	0.0379	

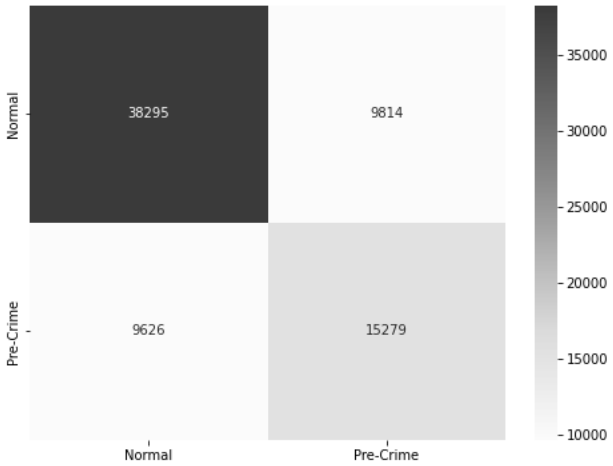

Fig. 7. Confusion matrix for random forest.

5 Conclusions

This paper explores the data preprocessing stages in crime prediction using computer vision, focusing on data selection, class balancing, and three-dimensional representation of human poses. Important decisions such as dataset selection have a significant impact on the results. Opting to use the Shoplifting Dataset 2022 allowed for more accurate extraction of human keypoints due to the higher resolution and quality of the video clips compared to the UCF – Crime Dataset.

The use of data augmentation techniques made it possible to meet the instance count restriction required by the NOTEARS algorithm, and k-Means instance selection ensured class balance while maintaining the structure and behavior of the original data. Furthermore, NOTEARS provided dependency graphs in less time and with lower resource consumption than traditional combinatorial methods. These preliminary results show that DAGS calculated with balanced data maintain comparable quality to those generated with the entire data set, validating the proposed approach.

Detection and prediction tasks can be extended to other crimes such as theft or assault. In addition, in the future, a greater number of angles between joints can be considered, along with the inclusion of temporal statistical data obtained using existing angles and other preprocessing tasks such as data filtering to reduce erratic behavior or instability. Of course, the exclusive use of angles from human keypoints suggests limitations such as not utilizing more contextual information or, at this point, interactions with other people. However, the preliminary result can be used as a starting point for improvements, taking into account the previously suggested expansions that can be carried out.

This paper lays the groundwork for future research on crime prediction using computer vision, which has the potential to contribute to public safety.

Acknowledgments

This research work is supported by SECIHTI through national scholarships for post-graduate students number 846099.

References

1. Martínez-Mascorro, G.A., Abreu-Pederzini, J.R., Ortiz-Bayliss, J.C., Garcia-Collantes, A., Terashima-Marín, H.: Criminal Intention Detection at Early Stages of Shoplifting Cases by Using 3D Convolutional Neural Networks. Computation **9**(24), (2021)
2. Sultani, W., Chen, C., Shah, M.: Real-World anomaly detection in surveillance videos. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6479 – 6488. IEEE, Salt Lake City, UT, USA (2018)
3. Gandapur, M.: E2E-VSDL: End-to-end video surveillance-based deep learning model to detect and prevent criminal activities. Image And Vision Computing **123**, 104467–104476 (2022)
4. Caviar dataset, http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/, last accessed 2025/04/30
5. Pouyan, S., Charmi, M., Azarpeyvand, A., Hassanpoor, H.: Propounding First Artificial Intelligence Approach for Predicting Robbery Behavior Potential in an Indoor Security Camera. IEEE Access **11**, 60471–60489 (2023)
6. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779 – 788. IEEE, Las Vegas, NV, USA (2016)
7. Karimi, D., Dou, H., Warfield, S. K., Gholipour, A.: Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical Image Analysis **65**, 101759–101789 (2020)
8. Shoplifting Dataset (2022) - CV Laboratory MNNIT Allahabad, https://mmpose.readthedocs.io/en/latest/, last accessed 2025/04/30
9. Welcome To MMPose's Documentation, https://mmpose.readthedocs.io/en/latest/, last accessed 2025/04/30
10. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training. In: 2022 IEEE/CVF Conference On Computer Vision And Pattern Recognition (CVPR), pp. 7745 – 7754. IEEE, Salt New Orleans,LA, USA (2022)
11. Yan, W.Q., Nguyen, M., Stommel, M.: Image and Vision Computing. (2023). https://doi.org/10.1007/978-3-031-25825-1.
12. Bindal, N., Garg, B.: Novel three stage range sensitive filter for denoising high density salt & pepper noise. Multimedia Tools And Applications. 81, 21279-21294 (2022)
13. Olvera-López, J. A., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F.: A new fast prototype selection method based on clustering. Pattern Analysis And Applications **13**(2), 131–141 (2009)
14. Jin, X., Han, J.: K-Means clustering. In: Encyclopedia of Machine Learning. pp. 563-564 (2010)
15. Zheng, X., Aragam, B., Ravikumar, P., Xing, E. P.: DAGs with NO TEARS: continuous optimization for structure learning. In: NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 9492 – 9503. Curran Associates Inc, Montreal, Canada (2018)