Journal of Computer Science and Information Technology
https://jcsitech-upiyptk.org/ojs
2021 Volume 7 Issue 4 Hal: 94-104 e-ISSN: 2502-1486

Violence Detection in Ranches Using Computer Vision and Convolution Neural Network

Terungwa Simon Yange[1]*, Oluoha Onyekware[2], Charity Ojochogwu Egbunu[1], Malik Adeiza Rufai[3] and Comfort Eneya Godwin[1]

[1]Department of Computer Science, Joseph Sarwuan Tarka University, Makurdi, Nigeria
[2]Department of Computer Science, University of Nigeria, Nsukka, Nigeria
[3]Department of Computer Science, Federal University Lokoja, Lokoja, Nigeria
lordesty2k7@ymail.com

Abstract

This study engaged the convolutional neural network in curbing losses in terms of resources that farmers spends in treating animals where injuries must have emancipated from violence among other animals and in worst case scenario could eventually lead to death of animals. Animals in a ranch was the target and the study proposed a method that detects and reports activities of violence to ranchers such that farmers are relieved of the stress of close supervision and monitoring to avoid violence among animals. The scope of the study is limited to violence detection in cattle, goat, horse and sheep. Different machine learning models were built for each animal. The models yielded good results; the horse violence detection model had an outstanding performance of 93% accuracy, 93% accuracy for the sheep model, 88% accuracy for the goat model and 84% accuracy for the cattle model.

Keywords: Convolutional Neural Network, Computer Vision, Detection, Machine Learning, Animals.

1. Introduction

Man's understanding of his environment and how to make living in the world easier with less stress has led to amazing and evolutionary discoveries over the century. One of which can be called 'giving machine sense' (artificial intelligence). This research covers an area of artificial intelligence technology known as computer vision. Computer vision is a concept of technology that seeks to emulate not just the ability to view or capture images, objects but also to recognize or label the object like the human brain[1]. The human vision has two concepts; sensation and perception. Sensation means the activation of senses and perception is the process of understanding sensation. For a computer, image and video inputs are collected by devices such as cameras and received by a computer. This makes up its vision sensation. Its perception is the process of applying algorithms to determine relationships between the data and appropriate output[2]. Research on computer vision began in the 1950's where computers were trained to distinguish between circles and squares. And has now progressed to the point where object detection has reached up to 99%[3]. This technology has application in various spheres like automotive, healthcare, agriculture, banking, security etc. but here we will see how it can be applied in the field of agriculture[4].

A ranch is a large area of land used for rearing livestock. These livestock may include cattle, sheep, bison, horses etc. Keeping animals requires time and attention as there are certain things, they cannot do for themselves[5]. Even pet owners pay close attention to the welfare of their pets how much more a rancher who deals a number of animals depending on the size of his ranch. The bulk of the purpose of ranching is for profit (business) and so a successful business endeavour must make profits and minimize losses to the utmost minimum[6]. The kind of loss being discussed in this context refers to damages as a result of injuries, impairment and mutilation when animals fight. In the course of this study, we will see how such can be avoided even in a scenario where the ranchers are not readily available at the scene. Keeping watch after animals can be long and tasking. Just like humans, animals get into fights with one another for reasons which vary from competition for a food, mate, space or sometimes no reason at all[7]. These fight most of the time leads to serious damages ranging from bruises, cuts, bites, mutilation and even death. However, the extent of the damage, it incurs loss on the rancher.

The violence detection system is coined from machine learning. The system will be taught to identify a fight and alert the ranchers. It will as earlier stated receive input from CCTV cameras, process it and produce a corresponding output such as an alarm. This study therefore aims to develop a violence detection system that will notify ranchers in an event of a fight among ranch animals..

2. Related Works

The technology of computer vision provides a means where computer can perceive its environment and interpret the happenings around. It is a branch of artificial intelligence known as machine learning.

Accepted: 23-11-2021 | Revision: 20-10-2021 | Publication: 23-11-2021 | doi: 10.35134/jcsitech.v7i4.22

Machine learning can simply be put as using data to answer questions. This data is used to train a predictive model which can then be used to sever predictions on previously unseen data and answer those questions[8]. Computer vision allows robots and machines see the world around them. Once they can well perceive the world around them, thus interact with it better[9]. Computer vision seeks to emulate the human vision processing. It is amazing that humans and animals do this so effortlessly, while computer vision algorithms can be so error prone. And yet people who have not worked in the field often underestimate the difficulty of the problem[10]. And even as it seems very easy for humans to process vision, the human vision process still has not been fully grasp. There are underlying factors which makes it complex for machines to replicate/ emulate that standard of vision such as lighting, perspective, colour, sizes, texture of objects etc[11]. Computer vision is a broad field of study owing to the fact that it cuts across many other fields such as image processing, computer graphics, mathematics and statistics among others and there have been breakthroughs over the years in the application computer/machine vision in various sectors of economy medicine, agriculture, security, business, industry etc.

2.1 Violence Detection

Violence detection in any capacity requires the knowledge and understanding of the behaviour and responses of the object of study. So also, to detect violence in ranches, basic knowledge of the animal flight behaviour specifically ranch animals is very important. Animals would naturally flee to their place of habitat should they feel threatened in any way and each time there is a threat in the flight zone, the animal will react or move away indicating an impending threat. This also aids in the movement of animals or herds. The size of the flight zone is determined by the animal's tameness, fear or stress. Equids like horses are grazers and prey animals by nature. Their "fight or flight" reaction is prominent. Their senses are developed to rapidly detect changes in their environment[12]. Cattle higher in the dominance order have greater flight distance and normally intersperse themselves throughout the herd to avoid contact with each other. This means that less dominant cattle continually have to move around to avoid contact with them[13].

2.2 Empirical Study on Violence Detection Methods

There are several and diverse methods of detecting violence and over the years, research has been progressive. Below are some works/research by several authors in detecting violence in the field of computer science:

Violence detection was classified into three different methods and these were based on visual, audio and the hybrid method[14]. Various state of art Violence detection methods can be classified into four groups namely Optical flow based, Space time interest point, Motion binary pattern and Grey level co-occurrence matrix. Their work was only a theoretical review of violence detection methods.

Recorded ECG signals. Next, ECG and 3D acceleration signals were segmented according to violent and nonviolent events based on the video recording, such that ECG signal length was six seconds in each case[15]. Feature extraction was done using Bivariate Empirical Mode Decomposition (BEMD) and Recurrence Quantification Analysis (RQA). BEMD resulted in 5-6 intrinsic mode functions (IMFs) for each ECG signal. During experiments using RQA, it was found that the dimension parameter could not be larger than ten in order not to degrade the performance. The time-delay parameter, however, had a small effect and the search was limited from 5 to 25 samples with 5 steps. A sensor was strapped on 12 pupils; this can be unrealistic to have these devices strapped on animals much less humans in real life scenarios for the purpose of reading their ECG signals.

Research used optical flow history as the main parameter of his work. The learning method used is the random forest algorithm. With random forest, optical flow features obtained were captured in a histogram[16]. Multiple Instance Learning was used to detect anomaly in surveillance videos introduced a fight dataset and used two of the best action recognition methods currently available (STIP) and MoSIFT) to assess the performance in the fight detection problem[17]. Space-Time Interest Points (STIP) is an extension of the Harris corner detection operator to space-time. The detected interest points are characterized by a high variation of the intensity in space, and non-constant motion in time. These salient points are detected at multiple spatial and temporal scales. Then, HOG (Histograms of Oriented Gradients), HOF (Histograms of Optical Flow) and a combination of HOG and HOF termed HNF feature vectors were extracted for 3D video patches in the neighbourhood of the detected STIPs. These features can be used for recognizing motion events with high performance and they are robust to scale, frequency and velocity variations of the pattern. Their work spanned anomaly detection in humans alone.

Proposed a novel system for Violent Scenes Detection, which is based on the combination of visual and audio features with machine learning at segment-level[18]. Multiple Kernel Learning is applied so that multimodality of videos can be maximized. In particular, Mid-level Violence Clustering is proposed in order for mid-level concepts to be implicitly learned, without using manually tagged annotations. Finally, a violence-score for each shot was calculated. The whole system was trained on a dataset from MediaEval 2013 Affect Task and evaluated by its official metric.

Though their system achieved promising results, the results are not high enough. The first point to be considered here is that their feature vectors might not be distinct enough. Although they have used only Trajectory-based features for visual information, they can be easily affected by camera motion.

Considered normal and anomalous videos as bags and video segments as instances in multiple instance learning (MIL), and automatically learn a deep anomaly ranking model that predicts high anomaly scores for anomalous video segments[19]. Furthermore, sparsity and temporal smoothness constraints in the ranking loss function to better localize anomaly during training was introduced. They also introduced a new large-scale first of its kind dataset of 128 hours of videos. It consists of 1900 long and untrimmed real-world surveillance videos, with 13 realistic anomalies such as fighting, road accident, burglary, robbery etc. as well as normal activities. This dataset could be used for two tasks; the first for general anomaly detection considering all anomalies in one group and all normal activities in another group and the second, for recognizing each of the 13 anomalous activities. Their experimental results show that the MIL method for anomaly detection achieved significant improvement on anomaly detection performance as compared to some other approaches. They also provided the results of several deep learning baselines on anomalous activity recognition. The low recognition performance of these baselines reveals that their dataset is very challenging and opens more opportunities for future work.

Method to detect violent scenes in movies. The process considered two views (video and audio). For the audio view, a supervised method based on HCRFs is exploited to improve the classification performance[20]. For the video view, they detect violent shots by locating the violent areas. Finally, the classifiers from the audio and video were combined and co-trained. The results from several movies with violent contents confirmed the methods' effectiveness. Furthermore, the adoption of a weakly supervised method reduced the complexity in computation. The results even though it was not considered good enough must have stemmed from the fact that though the HCFR prediction of audio view got a pass mark for time dimension information, the classification performance was still affected by parameter settings like the length of testing sequence and the number of hidden states. Other reasons may be that the annotation and scaling rules are still not the optimal solutions, and the representative features and the blocks search pattern still have important influence on the performance.

Explored a different methodology of violence detection, which relies upon two deep neural network (DNNs) frameworks to learn spatiotemporal information on video clips under different scenarios subjective and conceptual-based[21]. Each concept was represented using deep feature representations after which they were aggregated and trained as a binary classification problem in description of violence using a shallowed neural network. Finally, they show that using more specific concepts is an intuitive and effective solution, besides being complementary to form a more robust definition of violence. When they compare the accuracies between the testing and validation set, it was found that the C3D model had a smaller difference from the validation to the testing set. They interpreted this as a higher robustness of this model, in comparison with the significant higher accuracy for the validation set in the CNN-LSTM model. The fusion network also shows its relevance when they compared its results with the average accuracy of all the individual concepts. For example, in the testing set for CNN-LSTM, the average accuracy of the networks for individual concepts was 58%, while the accuracy of the respective fusion network was 61%. A network trained directly for the violence detection problem – i.e., without relying on individual concepts and their fusion – yields 56% accuracy for both C3D and CNN-LSTM compared to the 63% and 61% of the proposed method in each network, respectively. Using two networks intended to detect features related to the passage of time and movement yielded better results in concepts that were closely related to movement, such as fights and explosions, while more static concepts such as blood and cold arms followed close behind in terms of accuracy. Combining features learned in these networks with another network, more suited for the detection of objects in still images, can lead one step further in solving this difficult problem.

Criminal Aggression Recognition Engine (CARE) was proposed as a method which detects violence from pre-recorded videos as well as from live video.

The model is trained with clips performing a number of actions, violent and non-violent, from varying datasets. The results achieved for live video were encouraging with almost 80% accuracy. There are a number of ways that CARE could be improved to become more useful, robust, and applicable to within real-life situations[22]. In order to recognize violent and nonviolent images, adopted the BoW model and integrated the SPM scheme and soft voting strategy[23]. Four different feature representations are tested in the experiments. Among the four basic procedures in the BoW model, they paid the most attention to feature representation so as to evaluate the effectiveness of different features in classifying violent and non-violent images. Four commonly used features are chosen as a comparison: SIFT, HOG, LBP and colour histogram. Among the four different feature representations, LBP has achieved the best classification performance followed by SIFT and HOG. Colour histogram turns out to be the least discriminative in classifying violence and non-violence images.

Unlike standard approaches that only use temporal motion information, a descriptor exploits the spatio-temporal characteristic of substantial derivative. In particular, the spatial and temporal motion patterns are captured by respectively the convective and local accelerations. After estimating the convective and local field from the optic flow, followed the standard bag-of-word procedure for each motion pattern separately, and they concatenated the two resulting histograms to form the final descriptor[23]. They extensively evaluated the effectiveness of the proposed method on five benchmarks, including three standard datasets (Violence in Movies, Violence in Crowd, and BEHAVE), and two new video-surveillance sequences downloaded from YouTube. The results of their method indicated the importance of spatial information to reveal complex pedestrian's dynamics in crowded scenarios. They demonstrated that the combination of the spatial and temporal motion patterns mostly has a significant effect on the performance of the classifiers.

CASSANDRA system aimed to detect human aggression in a complex real-world environment. A distinguishing aspect of CASSANDRA is the exploitation of complementary audio and video cues to disambiguate scene activity in real life environments[24]. From the video side, the system used overlapping cameras to track persons in 3D and to extract features regarding the limb motion relative to the torso. From the audio side, it classified instances of speech, screaming, singing, and kicking object. The audio and video cues are fused with contextual cues (interaction, auxiliary objects); a Dynamic Bayesian Network (DBN) produced an estimate of the ambient aggression level. It showed the benefit of combining the various cues, and the use of person-specific visual features derived from 3D person tracking. The CASSANDRA system is quite complex and ambiguous to perform the task.

Method to detect violence sequences. Firstly, the motion regions were segmented according to the distribution of optical flow fields. Secondly, they proposed to extract two kinds of low-level features to represent the appearance and dynamics for violent behaviours from the motion regions[25]. The proposed low-level features are the Local Histogram of Oriented Gradient (LHOG) descriptor extracted from RGB images and the Local Histogram of Optical Flow (LHOF) descriptor extracted from optical flow images. Thirdly, the extracted features were coded using Bag of Words (BoW) model to eliminate redundant information and a specific-length vector was obtained for each video clip. At last, the video-level vectors were classified by Support Vector Machine (SVM). In face of the noisy moving scenes, a new method was proposed to segment the motion regions according to the distribution of optical flow fields. The segmentation of the motion regions played an important role in simplifying the features and decreasing the noises. In the motion regions, they proposed to extract two kinds of low-level features: Local Histogram of Oriented Gradient (LHOG) and Local Histogram of Optical Flow (LHOF) to represent the video-based activities spatio-temporally. LHOG could capture the appearance information and LHOF obtained the dynamic information of the objects. Considering the different class spaces for different kinds of low-level features, they adopted the late-fusion strategy. That is to say, LHOG and LHOF features were processed respectively under the framework of BoW model, and then the two kinds of vectors were combined into new vectors, followed by the SVM classifier. Compared with the previous methods, the proposed method achieved better performance on the three challenging datasets (hockey fights, crowd violence and BEHAVE datasets). Experimental results could practically demonstrate the effectiveness of the proposed approach for both general violence and crowd violence sequences. Despite the results of their research, there still exist that vacuum in animal farms.

Violent content in movies can influence viewers' perception of the society. For example, frequent depictions of certain demographics as perpetrators or victims of abuse can shape stereotyped attitudes[26]. They propose to characterize aspects of violent content in movies solely from the language used in the scripts. This made their method applicable to a movie in the earlier stages of content creation even before it is produced. This they said is complementary to previous works which rely on audio or video post-production. Their approach was based on a broad range of features designed to capture lexical, semantic, sentiment and abusive language characteristics. They used these features to learn a vector representation for complete movie, and for an act in the movie. The former representation is used to train a movie-level classification model and the latter, to train deep-learning sequence classifiers that make use of context. They tested their models on a dataset of 732 Hollywood scripts annotated by experts for violent content. Their performance evaluation suggests that linguistic features are a good indicator for violent content. To date (as at the publishing of their work), they are the first to show that language used in movie scripts is a strong indicator of violent content. The work is the first to study how linguistic features can be used to predict violence in movies both at utterance- and movie-level. This comes with certain limitations. For example, their approach do not account for modifications in postproduction (e.g., an actor delivering a line with a threatening tone). Also, their results suggest that sentiment-related features were the most informative among those considered.

An approach for detecting violence in videos where Discriminative Slow Feature Analysis (D-SFA) was introduced to learn slow feature functions from dense trajectories derived from videos[27]. Afterwards, with the learnt slow feature functions, the Accumulated

Squared Derivative (ASD) features are extracted to represent videos. Finally, a linear support vector machine (SVM) is trained for classification. They also constructed a Violence Video (VV) dataset which includes 200 violence samples and 200 non-violence samples collected from Internet and movies. The results have shown that their approach achieved a promising performance. There is need to refine and enlarge our violence video database by collecting more diverse and representative violence videos, so that the violence detector can be evaluated more fairly.

Newtonian mechanics inspired methods such as Social Force Model have been successfully applied for anomaly detection in crowd scenes. However, several socio-psychology studies have shown that SFM-based methods may not be capable of explaining behaviours in complex crowd scenarios[28] This, an alternative approach consists in describing the cognitive processes giving rise to the behavioural patterns observed in crowd using heuristics. Inspired by these studies, they proposed a new hybrid framework to detect violent events in crowd scenario. Three behavioural heuristics are proposed (H1: An individual chooses the direction that allows the most direct path to a destination point, adopting his/her moving direction according to the possible presence of obstacles. H2: In crowd situations, the movement of an individual is influenced by his/her physical body contacts with surrounding persons. H3: In violent scenes, an individual mainly moves towards his/her opponents to display violent actions). Each heuristic is first formulated and estimated from video (heuristics estimation) and the Bag-of-Words paradigm is used for the representation of the estimated heuristics (heuristics representation) then all the forces are concatenated to shape the final descriptor, named, Visual Information Processing Signature (VIPS). Their work covers violence patterns in crowd scenarios (for humans) only.

Improved Fisher Vectors (IFV) for videos which allowed the representation of a video using both local features and their spatio-temporal positions[29]. Then, the popular sliding window approach for violence detection, and re-formulate the Improved Fisher Vectors and use the summed area table data structure to speed up the approach. The proposed extension has shown (according to their results) that their approach was shown to significantly outperform the existing techniques on three violence recognition datasets (all centred on human violence).

The performance of ViF in violence datasets varying the optic flow algorithm; they used Iterative Reweighted Least Squares (IRLS), Horn-Schunck and Lucas-Kanade. As datasets, a Crowded and Hockey datasets were used, also they built a new dataset with videos taken from surveillance cameras, which they named Surveillance Videos (SV)[30]. Their evaluation concluded that the ViF's accuracy with the IRLS optic flow algorithm had better results, but in the case of Hockey dataset, ViF's with Horn-Schunck was better. The computational cost of the optical flow algorithms was evaluated the top performer (Horn- Schunck) with only 0.25 seconds to process two frames, compared to 16.95 and 7.80 seconds of Lucas-Kanade and IRLS respectively. Thus, the use of ViF with Horn-Schunck is highly acceptable due to its computational cost and better results for certain datasets such as Hockey enabling its use in real time yet not in ranch surveillance.

To detect violence in a video description method is to apply local spatio-temporal description on the query video. Then, the low-level description is further summarized onto the high-level feature based on Bag-of-Words (BoW) model[31]. However, traditional spatio-temporal descriptors are not discriminative enough. Moreover, the BoW model roughly assigns each feature vector to only one visual word, therefore inevitably causing quantization error. To tackle these constraints, their work employed Motion SIFT (MoSIFT) algorithm to extract the low-level description of a query video. To eliminate the feature noise, Kernel Density Estimation (KDE) was exploited for feature selection on the MoSIFT descriptor. In order to obtain the highly discriminative video feature, sparse coding scheme instead of BoW model was adopted to further process the selected MoSIFTs. Their test cases includes hockey fight and crowd violence which still leaves a vacuum for work in animal violence.

Machine learning classifiers: Support Vector Machines (SVM) and Neural Networks (NN). He used a balanced subset of the existing ACCEDE database of movie excerpts containing 880 movie excerpts manually tagged as violent or non-violent[32]. During an early experimental stage, using the features originally included in the ACCEDE database, tested the use of audio features alone, video features alone and combinations of audio and video features. These results provided the baseline for further experiments using alternate audio features, extracted using available toolkits, and alternate video features, extracted using his own methods. He found that audio features could be easily extracted using existing tools and have a strong impact in the system performance and in terms of video features, features related with motion and shot transitions on a scene seem to have a better impact when compared with features related with colour or luminance; the best results are achieved by combining audio and video features.

2.3 Summary of the Review

Violence detection in ranches is new as existing works focuses on anomaly detection in humans but not one on animals. Whereas, animal tussle does exist and losses are incurred from those events seeing they are living things with temperaments which resulting from their personality behaviours exhibited as they react to

externals factors. Ranching can be made less demanding with a system in place such as the one we propose; a system to curtail animal violence by detecting anomaly behaviours in ranch animals without having to be on the scene.

Different techniques were proposed for anomaly detection from the video and few in real time scenarios. These techniques can be classified into three categories based on the classifier used:

1. Violence detection technique using machine learning
2. Violence detection technique using SVM.
3. Violence detection technique using deep learning.

The accuracy of these methods was influenced by the technique employed in object recognition, features extraction and classification along with the datasets on which they are evaluated[33].

3. Materials and Methods

The CRISP-DM development method was employed in this research. This is because compared to KDD (Knowledge Discovery in Databases) and SEMMA (Sample, Explore, Modify, Model, and Assess), its flexibility allows for iterations which could help build stronger models.

3.1 Existing System

The existing ranch system involves a man (herdsman/cowboy) monitoring the herd in a vehicle or on back of a horse. This activity requires a number of manpower depending on the size of the herd they are looking after. They sit out with the herd and as anomaly activities are perceived, the herdsmen/cowboys come in the scene with their various animal restraint techniques to resolve the crisis. Managing a ranch system is hard work. It requires investment of time and energy and lots of it because of its nature. Herdsmen/farmers have to be with the animals for very long hours and in various weather conditions which can be wearisome but very necessary. The ranch system being a very old today lacks of skilled labor in the field. It is also capital intensive and requires a high cost of maintenance.

3.2 Data Collection

The data used in this work was sourced from Kaggle online data source and retrieved in their native form. It includes violent and non-violent scenes of the various animals under study namely: violent and non-violent scenes in cattle, violent and non-violent scenes in horses, violent and non-violent scenes in goats and violent and non-violent scenes in sheep. The data was in image format. The images were transformed from their native format into a jpg format whether as an image or video. Each pixel of the images has one channel. Grayscale images were more accurate than black and white and they were easier to process than other multiple channels like RGB. Also, the image data will be of 90 by 160.CNN requires large amount of data, thus, in this work, we got a total of 4910 capacity set.

3.3 Data Preparation

With the variety of data collected, it is difficult to have it in a perfect form. Thus, this stage involves extraction of video frames, image preprocessing, image labeling collected into the various classes they belong. Then finally, we split them into training and validation set.

Frame extraction: Data collected as videos will be converted into images by extracting their frames. The technique employed here is done by extracting the frames sequentially. This involves extracting one frame per second.

Image Preprocessing: The images and frames extracted are in different format, sizes and resolutions. The extracted frames will be resized to 90 by 160 to give them a uniform dimension. Also, they will be transformed to greyscale. This is done to reduce computation complexity. The images would then be sharpened using the unsharp masking technique.

Unsharp Masking: This is a sharpening technique where an unsharp/smoothened version of an image is subtracted from the original image. The smoothening is carried out using Gaussian blur. This can be considered as a convolution operation on an image with a kernel mask that is a two-dimensional Gaussian function g(x,y) as defined below:

$$g(x,y)=1/(\sigma\sqrt{2\pi})\ e^{(-(x^2+y^2)/\ [\![2\sigma]\!]\ ^2)} \quad (1)$$

The size of the Gaussian kernel mask is a function of the parameter σ, and the size of the kernel mask determines the range of frequencies that are removed by the Gaussian filter. The Standard Deviation slider determines the value of σ in pixels. The unsharp masking is then subtracted from the original image according to equation.2:

$$F(x,y)=c/(2c-1)\ I(x,y)-(1-c)/(2c-1)\ U(x,y) \quad (2)$$

In equation 3.2, F(x,y) represents the brightness value of the pixel at the coordinate (x,y) in the filtered image and I(x,y) and U(x,y) represents the brightness values of the corresponding pixels in the original and unsharp (blurred) images respectively. The constant c controls the relative weightings of the original and blurred images in the difference equation. The equation presented above demonstrates that an unsharp mask filter operates by subtracting weighted parts of the unsharp mask from the original image.

3.4 Modeling

In this phase, we design the model that will detect violence in ranch animals and report it.

1. Proposed System

We propose an intelligent system that will watch, monitor the activities of its environment (a ranch farm) and report violent activities. The system monitors the herd and interprets the activities of the herd and in this case as either violent or nonviolent. The system will receive input from its environment from a live feed such as a CCTV and send them to the system which will then process the raw data so it is readable by the system to describe/interpret the activity on the farm.

2. Algorithm

The following is a pseudo code describing the classification process and a corresponding flowchart (figure 2).
Start
Initialize camera
Extract one frame per second
Image preprocessing (frame):
Convert video frame to greyscale
Resize video frame (90,160)
Apply unsharp masking
Return frame
Classify image:
Apply pre trained model to preprocessed frame for classification
Return classification result
Decode result
If result is cattle violence || goat violence || horse violence || sheep violence
{
Output result
}
Read next frame
Stop

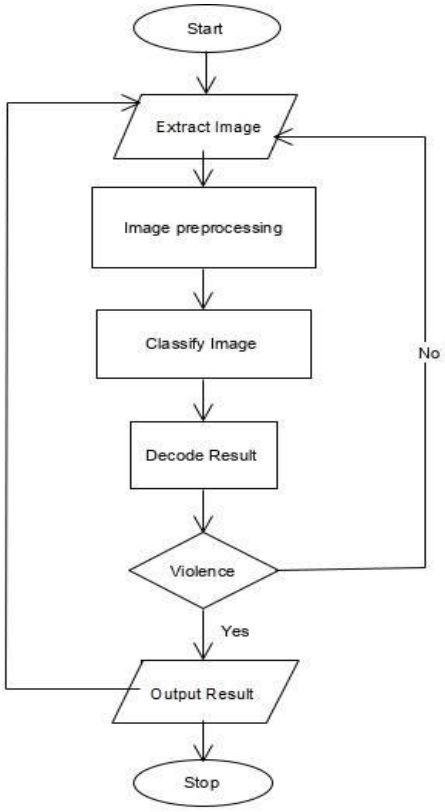

Figure 1: System Flowchart

3. Network Overview

The first part of the system will extract input (the live recording of the scene) frame after frame for the neural network. Each frame will be preprocessed so that it is compatible and readable by the network. The CNN then receives each frame as input into its layers (where the output of each layer feeds into the next as input) will be classified into violent or nonviolent classes.

The layers of a CNN are so called or classified by reason of the function they perform. The first layer of the CNN is the input layer which receives the images to be processed. The next layer, the convolution and pooling (sub sampling) layers extract features from the data and processes the data reducing the learnable features. Once the features are extracted by the convolution layers and down sampled by the pooling layers, the fully connected layer (the final layer of the CNN) maps them by a sub-set of fully connected layers to the final outputs of the network. The final fully connected layer typically is where classification is done and it has the same number of output nodes as the number of classes of classification[34].

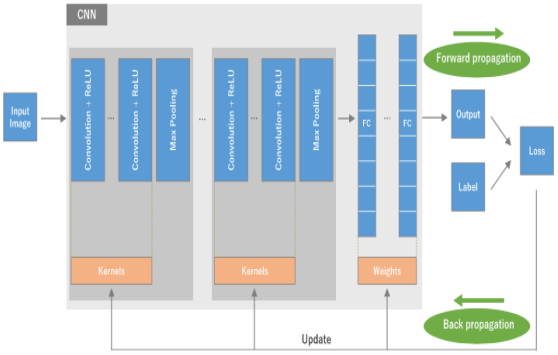

Figure 2: Network Architecture and Training Process

The input layer consists of 90 × 160 sensing nodes receiving the original image. It has no learnable parameter. The output of the input layer is the size of the input image. The calculation of the output shape of the convolution and sub-sampling is stated as follows:

$$\text{output} = \frac{input - kernel\ size + 2 * padding}{stride} + 1 \qquad (3)$$

The convolution layers are implemented using a 3 × 3 kernel size, stride of 1 and no padding at all. The subsampling layers use a kernel size of 2 × 2, stride of 2 and no padding as well. The final fully connected/dense layer produced an output vector of 2 classes for each animal violence detection model namely: no violence and cattle violence for cattle, no violence and horse violence for horses, no violence and sheep violence sheep and no violence and goat violence for goats.

4. Results and Discussion

The classification model for the violence cases were developed separately. That is cattle violence detection, horse violence detection, goat violence detection and sheep violence detection. This is because separately, they perform better than collectively. The models were built with 4 convolution and subsampling layers and an output layer with 2 vectors indicating animal violence and no violence for that animal. We used ReLU and sigmoid activation for the convolution and dense layers respectively. Table 1 shows the distribution of the dataset. 75% of each data was used for training and 25% for testing. The results from the test cases are read from a confusion matrix. Below are the test cases and their results.

Experiment 1

The system was tested with violent and non-violent scenes in cattle. This was done with 599 samples. Figure 5 shows that the system correctly classified 504 violent scenes as violence and wrongly classified 95 non-violent scenes as violent. It records a misclassification rate of 0.16% and an accuracy of 84%.

Experiment 2

Here, the system was tested with 278 samples of goat violence and no violence in goat. Figure 6 shows that it correctly classifies 246 as no violence and wrongly classified 32 samples of goat violence as no violence. This also means that the system has a misclassification rate of 0.11% and accuracy of 89%.

Experiment 3

The system was also tested with 144 samples of violence and no violence in horses. Figure 7 shows it correctly classified 134 samples of horse violence and wrongly classified 10 samples of no violence as violence. This gives rise to about 0.07% of misclassification rate and accuracy of 93%.

Experiment 4

Finally, the system was tested on 207 samples of sheep violence and no violence in sheep. Figure 8 shows that the sheep model correctly classified192 samples of no violence and wrongly classified 15 violence samples as no violence. This amounts to 0.07% misclassification rate and accuracy of 93%. Table 2 shows the results of the models after 5 epochs. The loss is a penalty for wrong prediction, if the model was perfect the loss would be zero; it gives the summation of errors made on the training and test sets. Accuracy is measured in percentage; it is the percentage of the data classified correctly whereas loss is a score and the closer it is to zero the better. The horse model outperformed the others having the lowest loss and a high accuracy score as well. The cattle violence detection model performed least having the highest loss and least accuracy on both the training and test data. Figure 9 shows the performance of the cattle model through 5 epochs. It started with a training accuracy of 79% and rose to 84% right after the first epoch. The test accuracy started off at 84% and remained stable through the 5 epochs. The model converged at 2nd epoch as shown in the loss graph in figure 9. Convergence is a state the loss settles within an error range. This means that further training will not improve the model. The cattle model converges at a loss of 0.44 on both training and testing.

The graph in figure 10 reveals that model was good. The goat model started with a training accuracy 81% and stabilizes after the first epoch. It also started with a test accuracy of 88% and maintained it through the 5 epochs. It started with a training and test loss of 0.66 and 0.62 respectively and converged after the 3rd epoch at a train and test loss of 0.37 and 0.36 respectively. The horse model which outperformed the others started at training and loss of test 0.63 and 0.50 converged faster than any other model at a loss of 0.25 and 0.23 respectively as shown in figure 11.

The horse model starts off with a training accuracy of 78% and converges at 93% after the 2nd epoch as described in the loss graph in figure 11. It attained a test accuracy of 93% from the 1st epoch and was stables throughout the training process. The performance of the sheep model is quite spectacular; it had a consistent and stable training and test accuracy of 93% through 5 epochs it converged after the 4th epoch with training and test loss of 0.27 each. The model shows slight symptoms of over fitting at the 3rd epoch where on the loss graph we see a slight increase in the loss greater than the train loss but was reverted in the next epoch as shown in figure 12.

Data remains the core for any machine learning project. The performance of the models varied for a number of reasons. In this case, our highest performing model which is the horse model had the least volume of data. One of the reasons is the fact that horses have a most distinctive way of showing violent behaviour. Sheep feeding side by side or goats in a cluster jesting can easily be mistaken for violence thereby their performance. Cattle have the most data available online. This is because bull fight is widely practiced around the world.

Despite the quantity, the quality of the data does not quite suit this work as most of it is filled with noise. It is filled with people (crowds) watching and cheering the competitors. Initially we had over 5000 data on ranch animals. They included very blurry images, lots of people on the scene (crowd), shadows, hardly was there any data that is consistent with a ranch structure/organisation of herd. Most of the data found

on goat are non-violent, a reason being that goat fights are uncommon and do not usually last. There exists a number of novel works in violence detection but in humans. One of such was done by Bruno *et al.*, (2019) in their work Towards Subjective Violence detection.

They also employed CNN in violence detection and achieved an accuracy of 63%. Also, Waqas *et al.*, (2015) employed the use of 3D convolutional network in the research Real-world Anomaly Detection in Surveillance Videos, attained an accuracy of 23% Even as they were good attempts, they cannot be employed in a ranching system. This is because animals and humans behave and react differently in threatening situations.

Table 1: Data Distribution

Animal Data	Volume	Training Set	Test Set
Cattle	2,394	1,975	599
Goat	1,111	833	278
Horse	576	432	144
Sheep	829	622	207

Table 2: Models' Performance after 5 epochs

Model	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
Cattle	0.44	84%	0.44	84%
Goat	0.37	88%	0.36	88%
Horse	0.25	93%	0.23	93%
Sheep	0.26	93%	0.26	93%

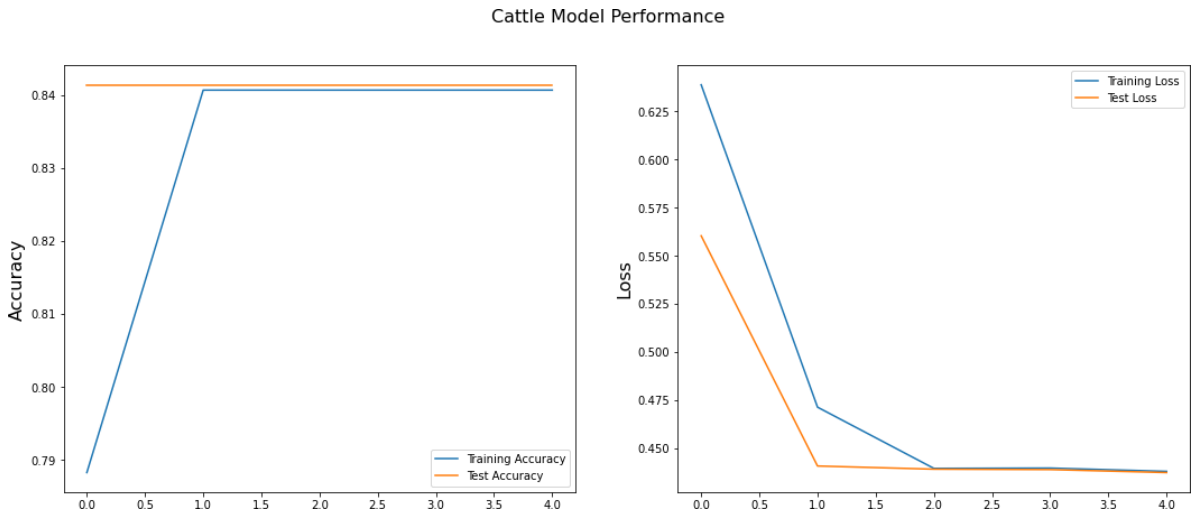

Figure 9: Graph Showing Cattle Model Performance

5. Conclusion

In this study, we developed a system that can detect violence activity in ranch animals; cattle, goats, horses and sheep as our case study. To achieve this, we built a deep learning model on an image dataset containing 4,910 images of violence and non-violence scenarios in cattle, goat, horse and sheep. We adopted the CRISP-DM development method to achieve this carefully defining each stage of the process as we go. The dataset was resized to 90 by 160, changed to greyscale and sharpened to reduce computational cost. We trained four (4) convolutional neural networks on 70% data of each animal. The different models showed promising performance with the horse model having the most outstanding performance and the cattle model performing least having the highest loss and least accuracy.

Ranching remains hard work but with a system such as this in place effectively performing its role, a significant workload is removed from the rancher even at such a time as this where skilful labour is on short supply. With this, ranchers can be rest assured that their herd are in good hands. Also, the work force required is significantly cut down no matter how large the herd is. Ranchers are encouraged to automate their business like every other business in the 21st century. This will help reduce cost and improve business processes. Also, we recommend that ranchers keep data on the daily activities and routines in their ranch. This will make providing data driven solutions effortless in the nearest future. There remains a vast vacuum even in this ranch system where artificial intelligence can play significant roles like object (animal) detection to take inventory of the animals in a matter of seconds, violence detection in other ranch animals such as poultry and even disease detection in the animals as well.

References

[1] Camacho Czenne. (2017). *A look at computer vision* [Video] Retrieved from https://m.youtube.com/watch?v=WkDsqeq0sGA

[2] Campbell Carolyn. (2019). *Computer vision - tech talk* [Video]. Retrieved from https://m.youtube.com/watch?v=6Hm-dqtjNjk

[3] Bruno, P., Bahram, L., Joao P., Sandra, A., Zanoni, D. and Anderson R. (2019). Toward Subjective Violence Detection in Videos. *IEEE*. 978-1-5386-4658-8/18/. https://ieeexplore.ieee.org/document/8682833

[4] Dong, W., Zhang, Z., Wei, W., Liang, W. and Tienui, T. (2012). Baseline Results for Violence Detection in Still Images. *IEEE*. DOI 10.1109/AVSS.2012.16. https://ieeexplore.ieee.org

[5] Hany, F., Liang, Y., Tian, H., Zhu, Z., Guobing, S., Tuija, H., and Esko, A. (2017). Violence Detection from ECG Signals: A Preliminary Study. Retrieved from www.jprr.org

[6] Hijazi, S., Rishi, K. and Chris, R. (2015). Using Convolutional Neural Networks for Image Recognition. www.cadence.com

[7] Phillips, C. (2002). *Cattle Behaviour and Welfare* (2nd ed.). United States: Iowa State University.

[8] Anuja, J. N. and Gopalakrishna, M. T. (2016). *Violence Detection in Surveillance Video-A survey*. International Journal of Latest Research in Engineering and Technology (IJLRET). Retrieved from https://www.ijlret.com/

[9] Azevedo, A., Santos, M. F., (2008). KDD, SEMMA and CRISP-DM: A parallel overview. Presented at IADIS European Conference on Data Mining 2008, Amsterdam, The Netherlands. Retrieved from: https://www.researchgate.net/publication/220969845

[10] Bermejo, E., Deniz, O., Bueno, G., and Sukthankar, R. (2011). Violence Detection in Video Using Computer Vision Techniques. Retrieved from http://visilab.etsii.uclm.es/

[11] Centre for Food Security and Public Health. (2010). *Animal Behaviour and Restraint: Cattle*. United States: College of Veterinary medicine, Iowa State University.

[12] Centre for Food Security and Public Health. (2014). *Animal Behaviour and Restraint: Equine*. United States: College of Veterinary medicine, Iowa State University.

[13] Gregorio, L. (2016). Detecting Violent Excerpts in Movies using Audio and Video Features. *Lisboa: ISCTE-UTL, 2016*. http://hdl.handle.net/10071/12995

[14] Guo, Y. (2017). *7 steps to machine learning* [Video]. Retrieved from https://m.youtube.com/watch?v=dTRsI8KNTW0/

[15] Hardeman, K. (2018). Spatio-temporal Classification of Aggression in Video Surveillance using Optical Flow History. Retrieved from https://www.dspace.library.uu.nl/

[16] Dan, F. (2014). A Harder Outlook for Today's Cowboy. *New York Times*. Retrieved from https://www.nytimes.com/2014/01/26/us/a-harder-outlook-for-todays-cowboy

[17] Dockter, R. (2018). *Lecture1: Introduction to Computer Vision* [PDF]. United States: The University of Minnesota.

[18] Eneim, M. (2016). An Intelligent Method for Violence Detection in Live Video Feeds. Retrieved from http://fau.digital.flvc.org/islandora/object/fau%3A33912/datastream/OBJ/view/

[19] Jian, L., Yi, S., and Weiqiang, W. (2010). *Violence Detection in Movies with Auditory and Visual Cues*. Paper presented at the International Conference on Computational Intelligence and Security. Retrieved from https://www.jdl.ac.cn/

[20] Kaiye, W., Zhang, Z. and Liang, W. (2012). Violence Video Detection by Discriminative Slow Feature Analysis. Liu, L., Zhang, C., and Wang L. (Eds.): *Pattern Recognition.Communications in Computer in Information Sciense*, *32*(1) Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3642-33506-8_18

[21] King Ranch. (2015). *Encyclopædia Britannica Ultimate Reference Suite*. Chicago: Encyclopædia Britannica.

[22] Kooija, J.F.P. Liema, M.C. Krijndersb, J.D. Andringab, T.C. and Gavrila D.M. (2015). Computer Vision and Image Understanding. http://dx.doi.org/10.1016/j.cviu.2015.06.009

[23] Long, X., Chen, G., Jie, Y., Qiang, W., and Lixiu, Y. (2014). Violent Video Detection Based On Mosift Feature and Sparse Coding. *IEEE*. https://ieeexplore.ieee.org

[24] Manasson, A. Ph.D. (2019). Why using CRISP-DM will make you a better Data Scientist Case Study: *Comparison of Los Angeles and New York Airbnb listings and trends using CRISP-DM*. Retrieved from https://towardsdatascience.com/why-using-crisp-dm-will-make-you-a-better-data-scientist-66efe5b72686

[25] Muhammad, R., Adnan, A., Hikmat, U., Shahid, M., Amina, I., Muzamil, A., Mahwish, I. and Ahsan M. (2019). A Review on State-of-the-Art Violence Detection Techniques. DOI: 10.1109/ACCESS.2019.2932114.

[26] Piotr, B. and Francois, B. (2016). Human Violence Recognition and Detection in Surveillance Videos. *IEEE International Conference on Advance Video and Signal Based Surveillance*. Retrieved from https://ieeexplore.ieee.org

[27] Shinichi, G., and Terumasa, A. (2014). Violent Scenes Detection Using Mid-Level Violence Clustering. In: David C. Wyld *et al.* (Eds) : CCSIT, SIPP, AISC, PDCTA, NLP – 2014 pp. 283–296, 2014. DOI: 10.5121/csit.2014.4224

[28] Yamashita, R., Mizuho N., Richard, K. and Kaori, T. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, *9*(4), 611–629. https://doi.org/10.1007/s13244-018-0630-9

[29] Zhou P, Ding Q, Luo H. and Hou X. (2018). Violence detection in surveillance video using low-level features. PLoS ONE 13(10): e0203668. https://doi.org/10.1371/journal.pone.0203668.

[30] Zhou, P., Qinghai, D., Haibo, L. and Xinglin, H. (2017). Violence Interaction Detection in Video Based on Deep Learning. *Journal of Physics: Conference Series*, *84*(4), 012-044. DOI: 10.1088/1742-6596/844/1/012044.

[31] Szeliski, R. (2010). *Computer Vision: Algorithms and Applications*. Retrieved from http://szeliski.org/Book/.

[32] Vicente, M., Karla, F. and Juan, C. (2016). Real Time Violence Detection in Video with ViF and Horn-Schunck. *LACCEI International Multi-Conference for Engineering, Education, and Technology: Engineering Innovations for Global Sustainability*. DOI: http://dx.doi.org/10.18687/LACCEI2016.1.1.122

[33] Victor, R., Krishna, S., Karan, S., Anil, R., Yalda, T. and Shrikanth, N. (2019). Violence Rating Prediction from Movie Scripts. Retrieved from www.aaai.org.

[34] Waqas, S., Chen, C. and Mubarak, S. (2015). Real-world Anomaly Detection in Surveillance Videos. Retrieved from https://openaccess.thecvf.com

[35] Warren, G., Doyle, G., and James B. *Understanding horse behaviour*. (PB1654). United States: Animal Science Department, University of Tennessee.

[36] Ranch**.** (2015). *Encyclopædia Britannica Ultimate Reference Suite*. Chicago: Encyclopædia Britannica.

[37] Sadegh, M., Hamed, K., Alessandro, P. and Vittorio, M. (2015). Violence Detection in Crowded Scenes using Substantial Derivative. *IEEE International Conference on Advance Video and Signal Based Surveillance*. Retrieved from http://www.hamedkiani.com/uploads/5/1//8/51882963/camera_ready_approved_final.pdf

[38] Sadegh, M., Hamed, K., Alessandro, P. and Vittorio, M. (2016). Angry Crowds: Detecting Violent Events in Videos. In: Leibe, b., Matas, J., Sebe, N., Welling, M. (eds) *Computer Vision – ECCV2016*. ECCV 2016. Lecture Notes in Computer Science, *99*(11). Springer, Cham. https://doi.org/10.1007/978-3-319-4647-7_1

[39] World Health Organisation. (2002). *World report on violence and health: summary*. Genever: Author

[40] Zhou, X. (2018). *Understanding the Convolutional Neural Networks with GradientDescent and Backpropagation*. Presented at Journal of Physics: Conference Series 1004 (2018) 012028. Retrieved from https://doi:10.1088/1742-6596/1004/1/012028