

Shot Classification for Human Behavioural Analysis in Video Surveillance Applications

R.Newlin Shebiah* and S.Arivazhagan

*Centre for Image Processing and Pattern Recognition, Department of Electronics and Communication Engineering
Mepco Schlenk Engineering College, Sivakasi - 626005, Tamilnadu, India*

Received 5 June, 2023; accepted 23 September 2023

Abstract

Human behavior analysis plays a vital role in ensuring security and safety of people in crowded public places against diverse contexts like theft detection, violence prevention, explosion anticipation etc. Analysing human behaviour by classifying of videos in to different shot types helps in extracting appropriate behavioural cues. Shots indicates the subject size within the frame and the basic camera shots include: the close-up, medium shot, and the long shot. If the video is categorised as Close-up shot type, investigating emotional displays helps in identifying criminal suspects by analysing the signs of aggressiveness and nervousness to prevent illegal acts. Mid shot can be used for analysing nonverbal communication like clothing, facial expressions, gestures and personal space. For long shot type, behavioural analysis is by extracting the cues from gait and atomic action displayed by the person. Here, the framework for shot scale analysis for video surveillance applications is by using Face pixel percentage and deep learning based method. Face Pixel ratio corresponds to the percentage of region occupied by the face region in a frame. The Face pixel Ratio is thresholded with predefined threshold values and grouped into Close-up shot, mid shot and long shot categories. Shot scale analysis based on transfer learning utilizes effective pre-trained models that includes AlexNet, VGG Net, GoogLeNet and ResNet. From experimentation, it is observed that, among the pre-trained models used for experimentation GoogLeNet tops with the accuracy of 94.61%.

Key Words: Shot Classification, Human Behaviour Analysis, Video Surveillance, Transfer Learning, Face Pixel Percentage.

Correspondence to: newlinshebiah@mepcoeng.ac.in

Recommended for acceptance by Angel D. Sappa

ELCVIA ISSN:1577-5097

Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

1 Introduction

A filmmaker envisions the narrative aspects of cinematic composition using shot, which is the ultimate element of visual language. Shot literally preserves semantics and affective states by increasing or decreasing the camera to subject distance to produce a deeper or shallower depth of field. Based on the distance of the focused object from the camera, video shots are broadly classified into Close-up, Medium and Long shot [1].

Close-up shot inhabits the screen with fractional portion of the subject such as a person's face, eyes etc. The close-up shot conceals the extraneous visual information and highlights the small tones of facial behavioral cues and emotion. With Close-up shot, measuring face attractiveness irrespective of the viewer is possible by analyzing shape, texture, symmetry and universal standard of beauty based on the golden ratio. Further, eye movements and gaze positions help to analyze human behaviour for interactive and diagnostic applications. Age estimation and Gender detection are potential surveillance applications that can be done with close-up shot. Medium Shot displays one or more characters above the waist and some surrounding areas. With mid-shot, facial expression can be interpreted to some extent. When more than one individual is included, medium shots can convey the intricacies of a relationship with the body posture and gaze exchange. The situation can be interpreted based on the interpersonal distance and the context. Long shots capture the subject or group of subjects from head to toe from a distance. The long shot displays the character's physical or emotional relationship to the environment and elements within it. The action and movement of the character are emphasized more rather than the emotional state. With long shot, natural characteristics such as height, body shape and certain personality traits can be analyzed. Postures are the most reliable cues about the actual attitude of people towards social situations that can be investigated. Further, the interpersonal distances between people like intimate, personal, socio-consultative and public spaces describe the nature of the relationship between them.

The state-of-the-art methods for shot classification rely on learning based techniques with statistical or deep features. Chauhan et al [2] proposed a simple shot classification method with Edge Pixel Ratio for basketball videos and classified it into two categories such as close-up and long views. Ekin et al [3] projected a statistical analysis using features like dominant field color, Grass pixel ratio, and color histogram for soccer video summarization. Sigari et al [4] used rule-based heuristics and SVM to classify the shots into far, medium, close-up, and out-field views. Tong et al.,[5] used features like color, texture, shot length, motion patterns, motion entropy, action regions, field shape properties, shot pace etc. and classified the shots using different supervised learning algorithms. Papachristou and Tefas [6] presented Linear Discriminant Analysis (LDA) based representation utilizing features like HSV / Disparity Histogram, Auto-correlogram, RGB moments, Gabor wavelet moments and Wavelet transform moments. To meter the camera distance of every shot, Shot Level Motion-Based Descriptors, Normalized shot duration, Stationarity percentage, Smoothness percentage, Shot Level Attention-Based Descriptor are fed into a probabilistic SVM classifier. Wang and Cheong [7], Chudasama and Patel [8] proposed a unified framework for cricket video shots classification as Field, Pitch, Boundary, Close-Up, Crowd, Fielders' gathering and Sky. The feature vector consists of Pixel Ratios of Grass, Pitch, Motion, Field, Skin, Edge and Sky trained with Multi-Perception Neural Network.

Xu et al [9] incorporated information density and geometric information in addition with saliency map for accurate visual attention. Further, the saliency map is combined with colour and texture features for SVM to classify the shot types into Close-up, Two Shot, Over Shoulder, Cut-In, Mid Shot, Wide Shot and Cut-a-way. Canini et al [10] built a model with an ensemble of features like local colour intensity distribution of frames, Motion activity maps, 2D geometry of the scene, face dimension and spectral amplitude. The model was trained using decision trees and Support Vector Machines. Cherif [11] set forth a qualitative description based on the human body information (i.e., head height and position) and classified the shots into seven categories and reported 90.91 percentage accuracy. Wei et al [12] proposed a probabilistic fusion model with effective features from the layer-wise output of a pre-trained network extracted from the deep convolutional neural network (CNN). Minhas et al [13] prove the efficiency of deep features in comparison with the handcrafted features for field sports videos. Savardi et al [14] used pre-trained deep network architecture and classified the shots from complete filmographies by six different directors into Close-up, medium and long shots.

2 Proposed Shot Scale Analysis System

The proposed methodology for shot scale analysis includes simple quantitative measures and deep learning-based method. The quantitative method is by analysing the percentage of Face Pixels in the frame. Further, the transfer learning based method that automatically learns the discriminative features and classifies it in to varying shot types is also presented.

2.1 Quantitative Method - Shot Classification based on Percentage of Face

The quantitative method of shot classification is by detecting the face region of the image using Viola Jones Face detector. The percentage of face pixels in the image is the ratio of area of face region to the total area of the image.

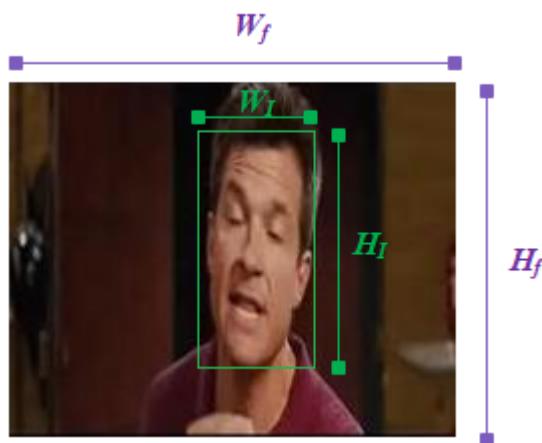


Figure 1: Close Shot depicting Face Region

Figure 1 depicts the face region and the entire image region of a close shot category.

$$\text{Face Pixel Percentage} = \left(\frac{W_f \times H_f}{W_I \times H_I} \right) \times 100 \quad (1)$$

where, W_f and H_f are the width and height of the face region of the image, while W_I and H_I are the width and height of the image. The Face pixel Ratio is thresholded with predefined threshold values and grouped into Close-up shot, Mid shot and Long shot categories.

The Viola-Jones algorithm captures distinguishing elements of face by Haar-like feature and accelerates the process by integral image technique. The usage of boosting scheme in Viola-Jones face detector effectively prunes the irrelevant features during the training stage. For further reducing the false positive rate and increasing the accuracy of detection, the face-detection algorithm is built with a cascade of boosted classifiers. In case of long shot type, when the face region is very small to be detected the human detection is carried with aggregate channel features and by using the anatomical proportion of human body, the head region is identified.

2.2 Learning Based Method – Transfer Learning for Shot Classification

In recent years, the popularity of handcrafted feature-based learning method appears to be overtaken by the Convolutional Neural Network (CNN) in classification, detection and segmentation task. Existing famed deep neural networks contain millions of parameters and are computationally and memory intensive. With minimal data, transfer learning could be able to rebuild a machine learning model. The principal concept of transfer learning is to utilize a composite and effective pre-trained model, trained from a cumbersome data source to a petty amount of data.

Alexnet: AlexNet is a relatively simple layout framed by Krizhevsky et al. [15] and declared as the winning solution for ImageNet Large-Scale Visual Recognition Challenge (ILSCRC) in 2012 classifying 1.2 million images into 1000 different categories. It comprises of five convolutional, three max pooling and three fully connected layers. Traditionally, deep convolutional neural networks use saturating nonlinearities i.e., hyperbolic tangent which facilitate vanishing of gradients and are computationally expensive. A computationally efficient non-saturating nonlinearity function Rectified Linear unit (ReLU) that train several times faster than their equivalents with tanh units is introduced. To reduce overfitting i.e., data augmentation and dropout are introduced in AlexNet architecture. The last fully connected layer connects to 1000 classes as the ILSCRC challenge aims in classifying the test images into 1000 different image classes. The structure of Alexnet needs to be modified to classify three categories as close-up, mid and long shots.

The key innovations of AlexNet include the use of deeper architectures, rectified linear units (ReLU) activation functions, dropout regularization, and the concept of utilizing Graphics Processing Units (GPUs) to accelerate training. The main characteristics of the AlexNet architecture are as follows:

- Input Layer: The network takes input images of fixed size (typically 224x224 pixels).

- **Convolutional Layers:** The initial layers consist of multiple stacked convolutional layers with small kernel sizes (e.g., 3x3 and 5x5) to learn low-level features such as edges, textures, and basic patterns.
- **ReLU Activation:** AlexNet employed the ReLU activation function, which helps mitigate the vanishing gradient problem and speeds up training by allowing the network to learn more quickly and efficiently.
- **Max-Pooling Layers:** After each convolutional block, max-pooling layers are used to reduce the spatial dimensions of the feature maps while preserving the most important information.
- **Fully Connected Layers:** Towards the end of the network, there are several fully connected layers that combine high-level features to make predictions.
- **Softmax Output:** The final layer is a softmax layer that provides the predicted probabilities for different classes in the classification task.
- **Dropout Regularization:** AlexNet introduced the concept of dropout, a regularization technique where random neurons are temporarily "dropped out" during training to prevent overfitting and improve generalization.
- **GPU Acceleration:** AlexNet was one of the first deep learning models to leverage powerful GPUs for training, significantly reducing the training time compared to traditional CPUs.

VGG Net: The enhancement over AlexNet by changing large kernel-sized filters with multiple 3×3 kernel-sized filters is demonstrated by the architecture from Visual Geometry Group, Oxford (VGG Net) [16]. It proves that with a given receptive field, multiple stacked smaller size kernels are better than the one with a larger size kernel. This configuration utilizes 1×1 convolution filters and partial pooling for some of the intermediate convolution layers. A stack of convolutional layers is followed by three Fully Connected layers with first two having 4096 channels and the third contains 1000 channels aimed at performing 1000 - class ILSVRC classification. The architecture of VGG Net can be summarized as follows:

- **1. Input Layer:** The network takes input images of fixed size (e.g., 224x224 pixels).
- **2. Convolutional Blocks:** Each convolutional block typically consists of two or more stacked 3x3 convolutional layers, followed by a max-pooling layer with a 2x2 window and a stride of 2. These blocks are used to extract hierarchical features from the input image.
- **3. Fully Connected Layers:** After several convolutional blocks, the network is followed by a few fully connected layers to further process the extracted features and make predictions.
- **4. Output Layer:** The final layer is a softmax layer that provides the predicted probabilities for different classes in the classification task.

VGG Net's simple and uniform architecture made it easy to understand, implement, and train. However, the drawback of VGG Net is its high computational cost and large number of parameters, especially in the deeper versions like VGG-19, which makes it more resource-intensive and slower to train compared to other architectures.

GoogLeNet: GoogLeNet [17], the winner of the ImageNet Large Scale Visual Recognition Competition 2014 is a pre-trained convolutional neural network that is 22 layers deep. Unlike other network architectures, GoogLeNet contains 1×1 Convolution and global average pooling. GoogLeNet has a pile of a 9 inception blocks and global average pooling to create its estimates. The inception module allows the network to learn and extract a wide range of features in parallel, capturing both local and global contextual information. This approach enables the network to be more expressive and computationally efficient compared to traditional networks with a large number of parameters. Maximum pooling between inception blocks reduced the dimensionality. To further reduce the number of parameters and computational cost, GoogLeNet also introduced the concept of "bottleneck" layers. These bottleneck layers use 1×1 convolutions to reduce the number of input channels before applying larger convolutional filters, effectively creating a bottleneck in the network and reducing computation. The inception layer covers a larger region, also concentrates on fine details on the images by convolving in parallel with varying filter sizes. The concept is in concurrence with of Gabor filters with varying scales and orientation that could handle better multiple objects scales. The advantage is that the inception layer is learnable.

ResNet: Residual Neural Network (ResNet) [18] is an innovative architecture with skip connections and batch normalization. Skip connection enables to have deeper network without vanishing gradient problem. ResNet is 20 times deeper than AlexNet and 8 times deeper than VGG. He et al. (2015) proved that ResNet has less error on classification task than 34 layers plain Network. Furthermore, ResNet reports 28 % improvement on COCO the image recognition benchmark dataset. The primary innovation of the ResNet architecture is the use of residual blocks, which enable the network to tackle the problem of vanishing gradients in very deep networks. As CNNs get deeper, training becomes challenging because of the vanishing gradient problem, where gradients diminish as they backpropagate through numerous layers, hindering the learning process.

To address this issue, ResNet introduces "skip connections" or "identity shortcuts." Instead of strictly cascading layers one after another, the network introduces shortcuts that allow the output of one layer to be directly added to the output of another layer further down the network. This process creates residual connections, where the network can learn residual mappings—meaning it learns to model the difference between the input and the desired output, rather than learning the entire transformation from scratch. The basic building block of ResNet is the residual block, which typically consists of two or more convolutional layers with batch normalization and ReLU activations, along with a shortcut connection. These residual blocks allow gradients to flow directly through the shortcut connections, making it easier to optimize very deep networks. ResNet architectures can be adapted to different depths by stacking multiple residual blocks together.

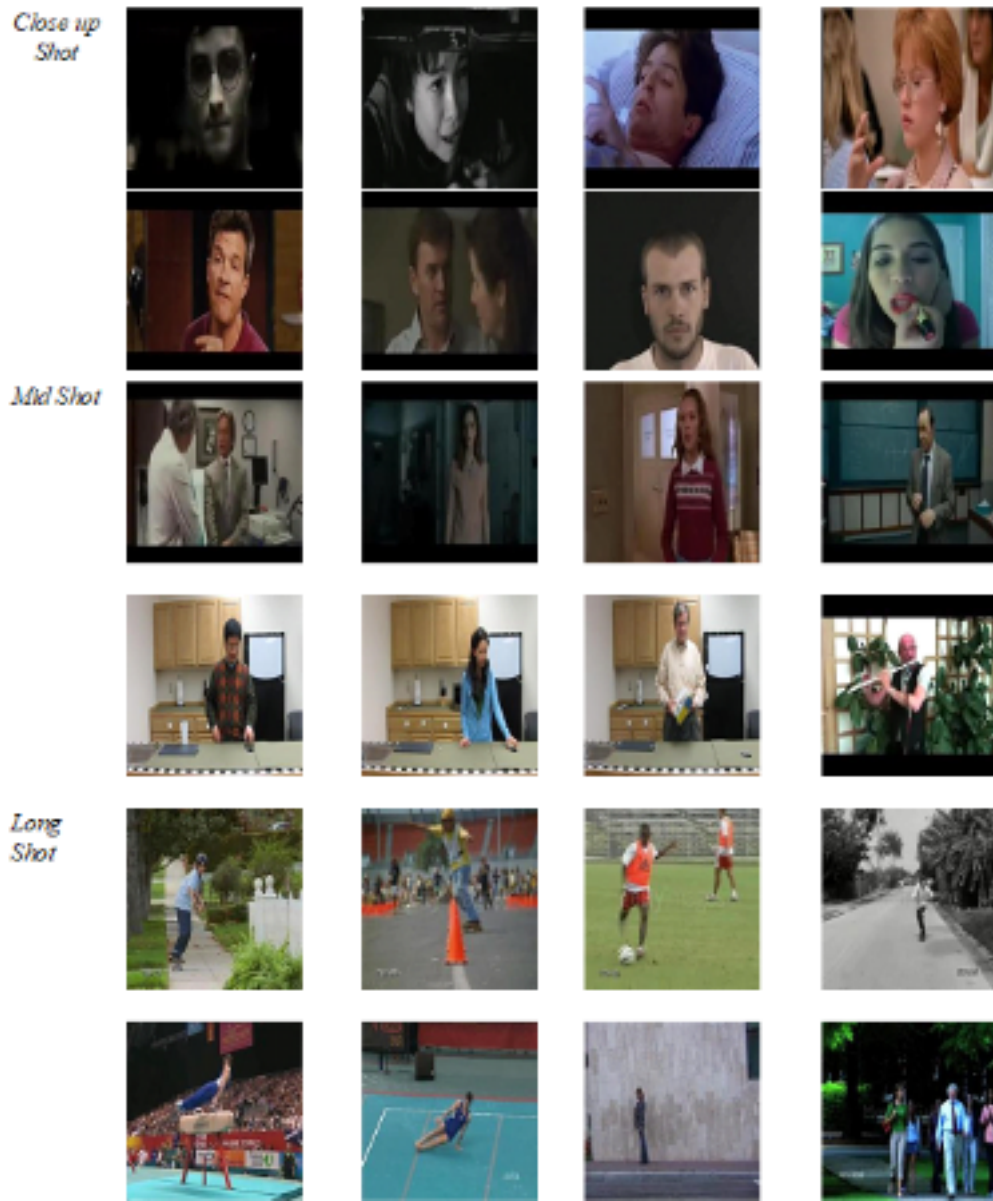


Figure 2: Sample Shots from the Database used for performance evaluation

Methodology of Shot Classification using Transfer learning: The methodology for automated shot type classification using transfer learning is briefed here. In all the pre-trained models used for experimentation, the last three layers are replaced to learn the characteristics of intended image categories. The step-by-step procedure is outlined below:

- Resize the input images in the dataset to be consistent with the size of the input layer of

the pre-trained network model.

- Review network architecture and replace the last three layers of the pre-trained network with a set of fully connected layer, softmax layer, and a classification output layer to categorize the images into three output classes.
- Train the network on the data for the task of shot scale classification.
- Test the accuracy of the new network on the testing dataset and report the performance metrics like Accuracy, Sensitivity, Specificity, Precision, False Positive Rate, and F1score.





3 Results and Discussion – Shot Classification

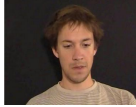


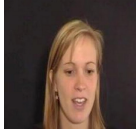
























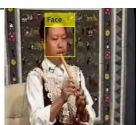







In this section, the strength of the proposed methodology for shot classification evaluated on the assorted database is discussed. The shots used for evaluation contains composite scenarios with challenges such as varying illumination condition, shadows, multiple subjects, complex background etc. For evaluation, sequences from public datasets, PETS 2006 [19], PETS 2016 [20], AFEW dataset [21], University of Rochester Activities of Daily Living Dataset [22], UCF101 - Action Recognition Data Set [23], Surrey Audio-Visual Expressed Emotion (SAVEE) [24] Database and HMDB: a large human motion database [25] are used. The datasets consist of 38,429 samples of close-up shot, 35,747 samples of midshot and 46,314 samples of long shot with a total of 1,20,490 samples. Some of the samples used for experimentation are shown in Figure 2.









3.1 Quantitative Method - Shot Classification based on Percentage of Face

The viola-jones face detection framework is the primary face detection structure to give competitive face detection. The area of the face region in the image is calculated and the changes in the face scale with respect to the shot type are explicitly visualized. Some sample face pixel ratio for different shot types are shown in Table 1.

Table 1: Percentage of Face Pixels for Shot Classification

| Shot Type | Input Image | Face Detected Image | Face Region | Percentage of Face Pixel Ratio |
|---------------|---|---|--|--------------------------------|
| Close-up shot |  |  |  | 8.5430 |
| |  |  |  | 8.9701 |

| | | | | |
|-----------|---|---|---|---------|
| |  |  |  | 16.3002 |
| |  |  |  | 20.7350 |
| |  |  |  | 22.3952 |
| |  |  |  | 40.7545 |
| Mid Shot |  |  |  | 1.5625 |
| |  |  |  | 3.3802 |
| |  |  |  | 1.9805 |
| |  |  |  | 3.5156 |
| |  |  |  | 3.8584 |
| |  |  |  | 5.5983 |
| Long Shot |  |  |  | 0.1956 |
| |  |  |  | 0.3061 |

| | | | |
|---|---|--|--------|
|  |  |  | 0.2932 |
|  |  |  | 1.5625 |
| |  |  | 0.2631 |

From Table 1, in case of long shot type if multiple people are present in the frame the individual face pixel ratio decides that the individual is in long shot, mid shot or close up shot type.

The distribution of face pixel ratio for different shot type is shown in Figure 3. From the Figure 3, it is perceived that there is a clear distinction in the values of long shot with other two shot types and some overlapping is observed between mid and close-up shot types.

The distance measure of face pixel ratio between shot types is displayed in Table 2. From Table 2 it is perceived that, the distance between long shot and close-up shot is maximum.

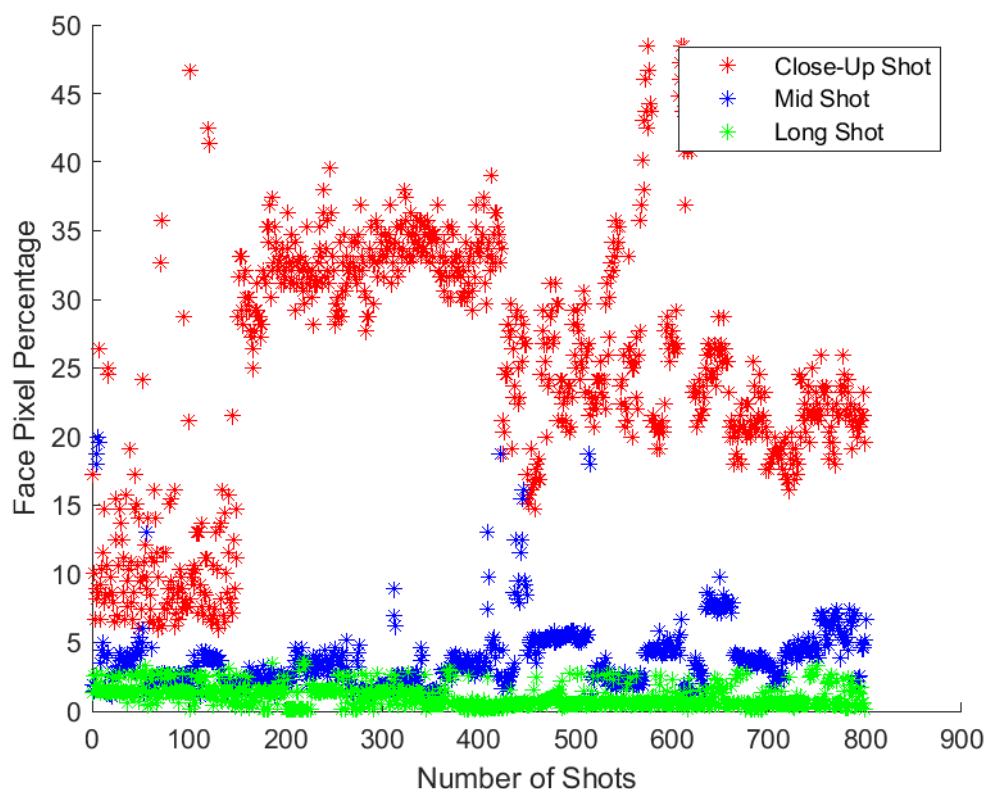
Table 2: Distance of Face Pixel Ratio for different Shot types

| Parameter | Value |
|--------------------------------|--------|
| dist(Long Shot, Mid Shot) | 78.28 |
| dist(Long Shot, Close up Shot) | 582.41 |
| dist(Mid Shot, Close up Shot) | 539.51 |

$$T1 = \frac{\text{dist}(\text{Mid Shot, Close up Shot})}{\text{dist}(\text{Long Shot, Close up Shot})} = 0.93 \quad (2)$$

$$T2 = \frac{\text{dist}(\text{Long Shot, Close up Shot})}{\text{dist}(\text{Long Shot, Mid Shot})} = 7.44 \quad (3)$$

Face Pixel Ratio is calculated for the test data samples and classified based on the threshold value. If the Face Pixel Ratio is less than T1 the shot is declared as long shot, if the value lies between T1 and T2 the shot is identified as mid shot and if it is greater than T2 it is classified as close-up shot.



3.2 Deep Learning based Method – Transfer Learning

In this study, four renowned pre-trained architectures such as AlexNet, GoogleNet, VGG-16 and ResNet-18 are deployed. The selective visual features are extracted by fine-tuning of transfer learning to increase the efficiency of a CNN network by replacing the last three layers of the pre-trained network.

This scenario works by transferring the weights of the pre-trained network from source dataset (ImageNet) to the target dataset. The common practice is to truncate the softmax layer of the pre-trained network and replace it with the proposed softmax layer that is relevant to the research problem considered (1000 classes of ImageNet are replaced with the three classes). In the second scenario, pre-trained network layers are frozen and treated as fixed features. This scenario works by deriving the weights of the pre-trained model from source dataset (ImageNet), and the desired features vector can be used from fully connected layers or from convolutional layers for training a linear classifier (SVM) on the data of target task.

In this case, Stochastic Gradient Descent Momentum (SGDM) is used with 0.9 momentum in each architecture. The value of batch size is set to 10 with initial learn rate of 0.0001 and maximum epochs of 30. All fine-tuned networks are applied to the validation data to get the classification accuracy. The accuracy of each network on each dataset for each strategy are listed in Table 3.

The close-up shot contains only the face regions where the smooth regions of face and fine details like eyes lips and nose dominates the frame, whereas mid shot displays the upper body

Table 3: Classification Accuracy for Shot Classification

| Methods Used | Accuracy | Sensitivity | Specificity | Precision | False Positive Rate | F1_score |
|---|----------|-------------|-------------|-----------|---------------------|----------|
| Alexnet | 0.9107 | 0.9105 | 0.9555 | 0.9107 | 0.0445 | 0.9104 |
| Features from Alexnet and SVM | 0.8997 | 0.8996 | 0.9499 | 0.8997 | 0.0501 | 0.8996 |
| Features from Alexnet and Ensemble Classifier | 0.8973 | 0.8978 | 0.949 | 0.8973 | 0.051 | 0.897 |
| Features from Alexnet and KNN Classifier | 0.8459 | 0.8699 | 0.9315 | 0.8459 | 0.0685 | 0.84 |
| GoogleNet | 0.9461 | 0.9459 | 0.9732 | 0.9461 | 0.0268 | 0.9459 |
| Features from GoogleNet and SVM Classifier | 0.9203 | 0.9208 | 0.9607 | 0.9203 | 0.0393 | 0.9198 |
| Features from GoogleNet and Ensemble Classifier | 0.9325 | 0.932 | 0.9665 | 0.9325 | 0.0335 | 0.9321 |
| Features from GoogleNet and KNN Classifier | 0.9144 | 0.9243 | 0.9604 | 0.9144 | 0.0396 | 0.9131 |
| ResNet | 0.9109 | 0.9106 | 0.9556 | 0.9109 | 0.0444 | 0.9106 |
| Features from ResNet and SVM Classifier | 0.9045 | 0.9061 | 0.9531 | 0.9045 | 0.0469 | 0.904 |
| Features from ResNet and Ensemble Classifier | 0.8941 | 0.8963 | 0.948 | 0.8941 | 0.052 | 0.8925 |
| Features from ResNet and KNN Classifier | 0.8877 | 0.9017 | 0.9484 | 0.8877 | 0.0516 | 0.8855 |
| VGGNet | 0.9291 | 0.9311 | 0.9656 | 0.9291 | 0.0344 | 0.9284 |
| Features from VGGNet and SVM Classifier | 0.9271 | 0.9311 | 0.9656 | 0.9291 | 0.0344 | 0.9284 |
| Features from VGGNet and Ensemble Classifier | 0.9437 | 0.9441 | 0.9721 | 0.9437 | 0.0279 | 0.9436 |
| Features from VGGNet and KNN Classifier | 0.8811 | 0.9032 | 0.9486 | 0.8811 | 0.0514 | 0.8759 |

and the dress colour along with facial features. However, it's not intense as in close-up shot and some background region covers the frame. From experimentation, it is observed that confusion occurs between close-up and mid shots. Since the long shot displays the subject in midst of the environment, the features from the surrounding and the entire body of the subject dominates. Thus, there is no misperception in distinguishing the longshot from other shot types.

Table 4: Classification Accuracy of Test Dataset using Percentage of Face Pixel Analysis

| Dataset | Action Categories | Number of Frames | Closeup | Mid Shot | Long Shot | Recognition Rate (%) |
|-----------|-------------------|------------------|---------|----------|-----------|----------------------|
| eNTERFACE | Angry | 906 | 906 | 0 | 0 | 100 |
| | Disgust | 773 | 773 | 0 | 0 | 100 |
| | Fear | 748 | 748 | 0 | 0 | 100 |
| | Happy | 700 | 700 | 0 | 0 | 100 |
| | Sad | 836 | 836 | 0 | 0 | 100 |
| | Surprise | 755 | 755 | 0 | 0 | 100 |
| SAVEE | Angry | 564 | 564 | 0 | 0 | 100 |
| | Disgust | 598 | 598 | 0 | 0 | 100 |
| | Fear | 571 | 571 | 0 | 0 | 100 |
| | Happy | 579 | 579 | 0 | 0 | 100 |
| | Neutral | 1102 | 1102 | 0 | 0 | 100 |
| | Surprise | 586 | 586 | 0 | 0 | 100 |
| AFEW | Angry | 321 | 208 | 88 | 25 | 92.21 |
| | Disgust | 211 | 123 | 63 | 25 | 88.15 |
| | Fear | 203 | 108 | 28 | 67 | 67 |
| | Happy | 367 | 247 | 70 | 50 | 86.38 |
| | Neutral | 389 | 311 | 56 | 22 | 94.34 |
| | Surprise | 187 | 137 | 21 | 29 | 84.49 |
| KTH | Boxing | 1893 | 0 | 74 | 1819 | 96.09 |
| | Handclapping | 1556 | 0 | 86 | 1470 | 94.47 |
| | Handwaving | 2273 | 0 | 73 | 2200 | 96.79 |
| | Jogging | 1324 | 0 | 46 | 1278 | 96.53 |
| | Running | 1307 | 0 | 74 | 1233 | 94.34 |
| | Walking | 1098 | 0 | 44 | 1054 | 95.99 |
| Weizmann | bend | 15 | 0 | 5 | 10 | 66.67 |
| | jack | 23 | 0 | 0 | 23 | 100 |
| | jump | 17 | 0 | 0 | 17 | 100 |
| | pjump | 22 | 0 | 0 | 22 | 100 |
| | run | 14 | 0 | 0 | 14 | 100 |
| | side | 17 | 0 | 0 | 17 | 100 |
| | skip | 22 | 0 | 0 | 22 | 100 |
| | walk | 24 | 0 | 0 | 24 | 100 |
| | Wave1 | 32 | 0 | 0 | 32 | 100 |
| | Wave2 | 31 | 0 | 0 | 31 | 100 |

| | | | | | | |
|------------|------------------|------|---|----|------|-------|
| UCF Sports | Diving-Side | 15 | 0 | 0 | 15 | 100 |
| | Golf-Swing-Back | 15 | 0 | 3 | 12 | 80 |
| | Golf-Swing-Front | 24 | 0 | 12 | 12 | 50 |
| | Golf-Swing-Side | 11 | 0 | 4 | 7 | 63.64 |
| | Kicking-Front | 10 | 0 | 2 | 8 | 80 |
| | Kicking-Side | 10 | 0 | 1 | 9 | 90 |
| | Lifting | 31 | 0 | 4 | 27 | 87.1 |
| | Riding-Horse | 19 | 0 | 5 | 14 | 73.68 |
| | Run_Side | 24 | 0 | 3 | 21 | 87.5 |
| | Skating | 18 | 0 | 3 | 15 | 83.33 |
| | Swing bench | 29 | 0 | 4 | 25 | 86.21 |
| | Swing-Side Angle | 11 | 0 | 2 | 9 | 81.82 |
| | Walk_Front | 63 | 0 | 6 | 57 | 90.48 |
| PETS 2006 | | 2031 | 0 | 16 | 2015 | 99.21 |
| PETS 2016 | | 627 | 0 | 8 | 619 | 98.72 |
| AVSS | | 212 | 0 | 0 | 212 | 100 |
| ABODA | | 490 | 0 | 0 | 490 | 100 |

With this pre-learned model, the public video datasets available for human emotion recognition (eNTERFACE, SAVEE & AFEW), action recognition (KTH, Weizmann & UCF Sports) and surveillance (PETS 2006, PETS 2007, AVSS and ABODA) is experimented with the rate of 1 frame per second and the classification accuracy of 97.82% is reported for UCF sports dataset, 99.17% for AFEW, 99.08% for KTH and remaining datasets with 100% accuracy. Whereas, by using Face Pixel Ratio, 81.06% for UCF Sports, 86.05% for AFEW, 95.70% for KTH, 96.67% for Weizmann, 99.21% for PETS 2006, 98.72% for PETS 2016, AVSS and ABODA with 100% accuracy. Table 3 shows the detailed classification rate using Face Pixel Ratio. Table 4 shows the comparison of the classification accuracy by percentage of face pixel analysis and Transfer Learning Method.

Table 5: comparison of the classification accuracy by percentage of face pixel analysis and Transfer Learning Method

| Dataset | | eNTERFACE | SAVEE | AFEW | KTH | Weizmann | UCF Sports | PETS 2006 | PETS 2016 | AVSS | ABODA |
|------------------------------|--------------------------------|-----------|-------|-------|-------|----------|------------|-----------|-----------|------|-------|
| Average Recognition Rate (%) | Face Pixel Ratio | 100 | 100 | 86.05 | 95.7 | 96.67 | 81.06 | 99.21 | 98.72 | 100 | 100 |
| | Transfer Learning based Method | 100 | 100 | 99.17 | 99.08 | 100 | 97.82 | 100 | 100 | 100 | 100 |

4 Conclusion

Developing a reliable automatic suspicious behaviour detection system is very important to avoid human fatigue, when monitoring surveillance scenes over an extended period of time. This research aims at exploring the behavioural cues of the person depending upon the shot scale. Framework for estimating the shot scale by learning based approach by exploiting the inherent characteristics of shots is analysed. It is observed that, close up shots could reveal subject's emotions and in long shots, wide view of the surroundings around the subject is visualized. The face detection based method is the simplest one when the face region is detected. The deep learning based method inspite of the computational complexity learns the features well and could distinguish between the classes. Without annotation human supervisor intervention substantial clues are provided to the Human Behaviour Recognition System to proceed further with video analytics.

References

- [1] Mercado, G. *The Filmmaker's Eye: Learning (and Breaking) the Rules of Cinematic Composition*. (Publisher Name,2010)
- [2] Chauhan, D., Patel, N. & Joshi, M. Automatic Summarization of Basketball Sport Video. *Proceedings On 2016 2nd International Conference On Next Generation Computing Technologies, NGCT 2016*. pp. 670-673 (2017,10)
- [3] Ekin, A., Tekalp, A. & Mehrotra, R. Automatic Soccer Video Analysis and Summarization. *IEEE Transactions On Image Processing*. **12**, 796-807 (2003)
- [4] Sigari, M., Soltanian-Zadeh, H., Kiani, V. & Pourreza, A. Counterattack Detection in Broadcast Soccer Videos Using Camera Motion Estimation. *Proceedings Of The International Symposium On Artificial Intelligence And Signal Processing, AISP 2015*. pp. 101-106 (2015)
- [5] Tong, X., Liu, Q., Duan, L., Lu, H., Xu, C. & Tian, Q. A Unified Framework for Semantic Shot Representation of Sports Video. *MIR 2005 - Proceedings Of The 7th ACM SIGMM International Workshop On Multimedia Information Retrieval, Co-Located With ACM Multimedia 2005*. **7**, 127-134 (2005)
- [6] Papachristou, K., Tefas, A. & Others Stereoscopic Video Shot Classification Based on Weighted Linear Discriminant Analysis. (2014)
- [7] Wang, H. & Cheong, L. Film Shot Classification. *Journal Name*. **19**, 1529-1542 (2009)
- [8] Chudasama, H. & Patel, N. A Unified Framework for Cricket Video Shot Classification using Low Level Features. *Journal Name*. **10** (2017)
- [9] Xu, M., Wang, J., Hasan, M., He, X., Xu, C., Lu, H. & Jin, J. Using Context Saliency for Movie Shot Classification. *Conference Name.*, 3714-3717 (2011)

- [10] Canini, L., Benini, S. & Leonardi, R. Classifying Cinematographic Shot Types. *Journal Name*. pp. 51-73 (2013)
- [11] Cherif, I. Shot Type Identification of Movie Content. *Conference Name*. (2014)
- [12] Wei, W., Lin, J., Liu, T. & Yang, Y. DeepNet Fusion to Classify Shots in Concert Video. *IEEE International Conference On Acoustics, Speech, And Signal Processing (ICASSP) 2017*. pp. 1383-1387 (2017)
- [13] Minhas, R., Javed, A., Irtaza, A., Mahmood, M. & Joo, Y. Shot Classification of Field Sports Videos Using AlexNet Convolutional Neural Network. *Applied Sciences (Switzerland)*. **9**, 1-22 (2019)
- [14] Savardi, M., Signoroni, A., Migliorati, P. & Benini, S. Shot Scale Analysis in Movies by Convolutional Neural Networks. *25th IEEE International Conference On Image Processing (ICIP)*. pp. 2620-2624 (2018)
- [15] Krizhevsky, A., Sutskever, I. & Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems 25*. pp. 1097-1105 (2012)
- [16] Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. (arXiv preprint arXiv:1409.1556,2014)
- [17] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. Going Deeper with Convolutions. (arXiv preprint arXiv:1409.4842,2014)
- [18] He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. (arXiv preprint arXiv:1512.03385,2015)
- [19] PETS 2006, <http://www.cvg.reading.ac.uk/PETS2006/data.html>.
- [20] Patino, L., Cane, T., Vallee, A. & Ferryman, J. PETS 2016: Dataset and Challenge. *IEEE Conference On Computer Vision And Pattern Recognition Workshops (CVPRW)*. (2016)
- [21] Dhall, A., Goecke, R., Lucey, S. & Gedeon, T. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE Multimedia*. **19**, 34-41 (2012)
- [22] Messing, R., Pal, C. & Kautz, H. Activity Recognition Using the Velocity Histories of Tracked Keypoints. *ICCV*. pp. 104-111 (2009)
- [23] Soomro, K., Zamir, A. & Shah, M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. (CoRR abs/1212.0402,2012)
- [24] Haq, S. & Jackson, P. Multimodal Emotion Recognition. *Machine Audition: Principles, Algorithms And Systems*. pp. 398-423 (2010)
- [25] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. & Serre, T. HMDB: A Large Video Database for Human Motion Recognition. *ICCV*. pp. 2556-2563 (2011)