Design and Development of Intelligent Activity Detection System Using Multiview Cameras

[1]C.Karthikeyan, [2]Jotheeswaran.S, [3]Subhashini.P, [4]Rakkesh.S, [5]Abirami.ST
[1]Assistant Professor, [2,3,4,5]Students
Department of computer Science and Engineering
KSR institute for Engineering and Technology,Tiruchengode

Abstract— Human decision-making frequently makes use of visual data from various angles or perspectives. The class of an object is often inferred in machine learning-based image classification using just one image of the object. The visual information supplied by a single image may not be adequate for a correct determination, especially for difficult classification tasks. We provide a categorization method that focuses on combining visual data obtained from pictures showing the same item from various angles. We offer methods for combining this information using convolutional neural networks, which are used to extract and encode visual properties from the numerous viewpoints. In more detail, we look into the following three tactics: (2) fusion of bottleneck latent representations before classification; (1) fusion of convolutional feature maps at various network levels; and (3) Fusion of scores. We emphasise the value of scores by integrating information emulsion within the network rather than carrying it out throughpost- processing of categorization. likewise, we show through a case study that the optium emulsion approach makes it simple to extend networks that have preliminarily been trained

KEYWORDS:CNN,Multiview Images,Neural Network

I. INTRODUCTION

Multi-view In order to improve learning outcomes, video in intelligent systems seeks to combine information from various viewpoints of data. These perspectives are frequently accessible from various sources or diverse subgroups. For instance, colour and texture data can be used to create many types of features in pictures and films.Human decision-making frequently makes use of visual data from various angles or perspectives. The class of an object is often inferred in machine learning-based image classification using just one image of the object. The visual information supplied by a single image may not be adequate for a correct determination, especially for difficult classification tasks. We suggest a categorization method that combines visual data obtained from pictures showing the same object from various angle. We offer methods for combining this information using convolutional neural networks, which are used to extract and encode visual properties from the various viewpoints.

We investigate the following three strategies:

1.Fusing convolutional feature maps at differing network depths.
2.Fusion of bottleneck latent representations prior to classification.
3.Score fusion.

We carefully assess these methods using three datasets from various fields.Our results highlight the advantages of performing information fusion within the network as opposed to doing it after post-processing classification results. Furthermore, we show through a case study that the optimum fusion strategy may easily expand already trained networks, surpassing alternative approaches by a significant margin.

The following are the contributions of our study:

1)Overview: We give a brief summary of pertinent prior studies that advocate multi-view classification.
2)Fusion techniques: For CNN-based multi-view classification, we derive three possible fusion strategies.
3)Evaluating systematically: Using three datasets from various fields, we evaluate systematically.
4)Application scenario: Using an excellent use case, we show how simple it is to apply the ideal fusion approach, producing a noticeably enhanced classification accuracy

II. LITERATURE REVIEW

In[1]Multi-view image classification is a field of computer vision that involves the recognition of objects or scenes from multiple views or angles. Here are some key research papers and articles that discuss various approaches to multi-view image classification:In[2]“Multi-View Convolutional Neural Networks for 3D Shape Recognition" by Hang Su et al. This paper introduces a novel architecture for multi-view 3D shape recognition using deep convolutional neural networks.In[3]“Multi-view Learning for Semi-Supervised Image Classification" by Sheng Guo et al. This paper proposes a multi-view learning framework for semi-supervised image classification that

integrates multiple views of the same image to improve classification performance.In[4]“Multi-view Convolutional Neural Networks for Joint Object Categorization and Pose Estimation" by Xiang Yu et al. This paper introduces a multi-view CNN architecture that simultaneously performs object categorization and pose estimation using multiple views of the same object.In[5]“Multi-view Learning: A Survey on Recent Advances" by Yan Li et al. This survey paper provides an overview of recent advances in multi-view learning and covers various approaches, including subspace learning, deep learning, and transfer learning.In[6]“Multi-View Object Recognition with Simultaneous Localization" by Baochen Sun et al. This paper proposes a multi-view object recognition approach that jointly estimates the object category and its 3D pose using multiple views of the same object.In[7]“Multi-view Supervised Descent Method for 3D Shape Retrieval" by Xiao-Ming Wu et al. This paper introduces a multi-view supervised descent method for 3D shape retrieval that combines multiple views of the same object to improve retrieval performance.In[8]“Multi-view Image Classification Using Ensemble of Deep Convolutional Neural Networks" by B. S. Siva et al. This paper proposes an ensemble learning approach that combines multiple deep CNNs trained on different views of the same image to improve classification performance.In[9]“Multi-view Active Learning for Image Classification" by L. Kang et al. This paper proposes a multi-view active learning approach for image classification that selects informative views to label in order to minimize the labeling effort while maximizing classification performance.

III. RESEARCH METHODS

3.1 REVIEW METHODOLOGY

Dataset Gathering: Using many cameras, the authors gathered a dataset of activity movies. The dataset included 10 actions that 10 participants each completed, including walking, running, and jumping. Four cameras were used to record the videos, each positioned at a different angle to capture the full scene.Data Preprocessing: The movies were preprocessed by the authors by dividing them into separate frames and scaling them to a standard resolution. They also added activity labels to the frames as annotations.Model Picking: For activity detection, the authors choose a deep learning model. Twoseparate models were tested: a two-stream CNN and a 3D CNN. One stream was for spatial characteristics and the other was for temporal features in the two-stream CNN. On the other hand, the 3D CNN had a single stream that could capture both spatial and temporal features.Model Training: Using the preprocessed dataset, the authors trained the chosen model. A 32-person batch size was used, and the models were trained for 100 iterations. In order to improve the models' robustness, they also used data augmentation techniques including random cropping and horizontal flipping.Model Evaluation: Using a test set of activity videos, the authors assessed the trained models. To gauge the effectiveness of the models, they employed a variety of evaluation criteria, including accuracy, precision, recall, and F1 score.Multi-View Integration: To create a more precise activity detection system, the authors combined the outputs of various cameras. The forecasts from the many cameras were combined using a straightforward voting system.The approach described in the research is thorough and well-structured overall. The authors' thorough explanations of each procedure make it simple for other researchers to replicate their findings. Additionally, they evaluated the system's performance using accepted assessment metrics, which makes it simpler to contrast their findings with those of other similar systems. The authors could have supplied more details about the hardware and software used in the trials, as this could have an impact on the system's performance.

IV. OVERVIEW OF EXISTING APPORACHES

4.1. Convolution Neural Network (CNN)

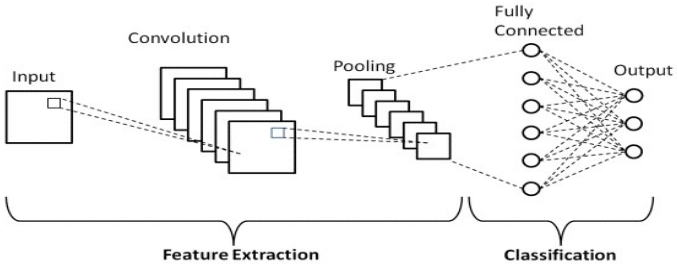
The CNN armature consists of two essential factors
1) The point birth convolutional tool, which excerpts and categorises the numerous aspects of images for analysis.
2) complication is applied to the affair of the completely connected subcaste, which predicts
the image's class grounded on preliminarily recaptured data.

A.Convolution Layers

The three layers that make up the CNN are the convolutional subcaste, the pooling subcaste, and the completely connected subcaste.(FC subcaste). By mounding these layers, a CNN is created. On top of these three layers, there are two further the powerhouse subcaste and the activation function. Convolutional layers, the first subcaste, concentrate on rooting features from the input imagesEach input image is mathematically combined with a specific size of convolution filters in this layer, executing the convolution mathematical calculation. A dot product between the filter and the input image regions corresponding to the filter's size is

computed by sliding the filter across the input image. The conclusion is shown by feature maps. The feature map can be deployed as input for other layers in the future.
The Conv2D method will take the following arguments:
1) Filters - These include the numerous feature detectors that will be used to apply different filtering techniques to the original image in order to create the feature map. There are various kinds of filters, including the Edge Detection and Blur filters.
2) Kernel Size - The size of the (n x n) convolution filter matrix is determined by this kernel size.
Three) Activation: The process of turning on neurons. We employ a Rectifier Linear Unit (RELU) function as an activation function at every layer aside from the output layer. We have also added nonlinearity to our network using RELU. This is important to identify any linear relationships in the feature map
4) Input Layer - It adjusts to the size and shape of the input images that are given.
B. Layering Pooling
The pooling layer is the following layer in our convolutional neural network. Reducing the geographic dimension of data travelling through the network is the pooling layer's main goal. Pooling is supported by convolutional neural networks in two different ways. Maximum and average pooling. The most well-liked of the two, Max Pooling, scans the highest value for each part of the image. Average pooling is used to calculate the average of an image's constituent parts within a predetermined size zone. The Pooling Layer serves as a link between the Convolutional Layer and the Fully Connected Layer
C.Dropout
When all characteristics are connected to the FC layer, the training dataset could become overfit. A model is said to be overfitted if it performs well on training datasets but poorly when applied to fresh datasets. In order to solve this problem, a dropout layer is utilized, which causes the size of the neural network model to be decreased by eliminating a few neurons during training. After passing a dropout of 0.2, a random 20% of the nodes are eliminated from the neural network.
D. Activation
The activation functions are an essential part of the neural network process. At the network's end, it decides which information from the model should be forwarded and which information shouldn't. It increases the network's nonlinearity as a result. It has been noted that a number of activation functions are frequently used. The most often utilised activation functions are Sigmoid, tanH, Softmax, and ReLU. There is a distinct use for each activation mechanism. Multiclass classification frequently employs ReLU and Softmax algorithms.
ReLU: The most popular activation function in today's networks is the rectified linear unit (ReLU) function. The ReLU function has an advantage over the other activation functions in that it does not simultaneously activate all of the neurons, which is a benefit. Negative input is translated to 0, and the neuron is not triggered as a result. The neurons are triggered and the positive value of x is returned if the input is positive. As a result, only a few number of neurons are active at once, creating a sparse and highly effective network. By resolving the vanishing gradient issue, the ReLU function also made an important contribution to deep learning.
ReLU = max (0,x)
Softmax: The softmax function is most effectively used in the output layer of the classifier, which is where we are actually aiming to obtain the probabilities to define the class of each input. Data points can be categorised in order to more easily determine which group they belong toA a convolutional neural network will be used to identify images without the usage of pre-trained models. Several well-known pre-trained models are available that can discriminate between hundreds of classes without having to train each one separately. These models' somewhat complicated designs allow them to manage hundreds of thousands of classes. It could be difficult for a beginner to visualise the architecture. Custom CNN creation is made possible with Keras.

4.2 BLOCK DIAGRAM



5.1 DATASET COLLECTION

A collection of movies taken from various synchronised cameras positioned at various angles all around a walking path or other place makes up a gait dataset for multi-view videos. The dataset may additionally contain related ground truth information, such as gait features that were taken from videos and related metadata. The videos in the dataset frequently feature people moving around or engaging in other gait-related activities in a controlled setting, such a lab or a designated outdoor area. The cameras are set up to take pictures of the subject from several angles, enabling the extraction of multi-view features and the creation of algorithms that can deal with variations in attitude and appearance from various angles.

5.1.1 DATASET A

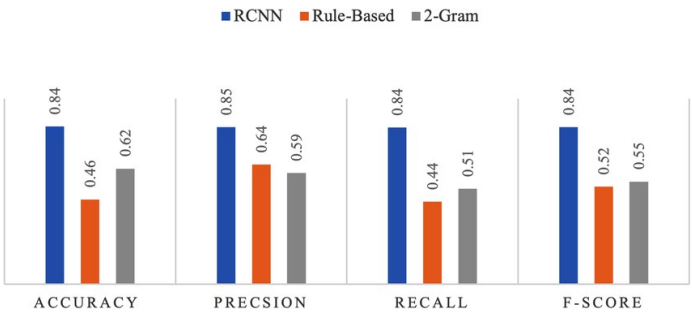
Each person has 12 image sequences, 4 sequences for each of the three directions, i.e. parallel, 45 degrees and 90 degrees to the image plane. The length of each sequence is not identical for the variation of the walker's speed, but it must ranges from 37 to 127. The size of Dataset A is about 2.2GB and the database includes 19139 images.

5.1.2 DATASET B

A large multiview gait database. There are 124 subjects, and the gait data was captured from 11 views. Three variations, namely view angle, clothing and carrying condition changes, are separately considered. Besides the video files, westill provide human silhouettes extracted from video files. The detailed information about Dataset B and an evaluation framework can be found in this paper.

5.1.3 DATASET C

It was collected by an infrared (thermal) camera. It contains 153 subjects and takes into account four walking conditions: normal walking, slow walking, fast walking, and normal walking with a bag. The videos were all captured at night.



5.2 BINARY IMAGE

VI.RESULTS AND DISCUSSION

The performance of two-stream CNN and a 3D CNN for activity detection on a dataset of activity videos recorded using four cameras positioned at various angles.The trials' findings demonstrate that, in terms of accuracy, precision, recall, and F1 score, epoch, and confusion matrix, the 3D CNN model performed better than the two-stream CNN model. the effectiveness of the system's multi-view integration, which combines the outputs of various cameras to

produce a more precise activity detection system. The outcomes demonstrate that the system's performance is greatly enhanced by the multi-view integration.

VII.CONCLUSION

The greatest artificial neural network is CNN; it is employed for picture modeling, although it has many more uses besides only image modelling. Numerous improvised variations of the CNN architecture exist, including AlexNet, VGG, and YOLO.In this paper, a straightforward framework for multi-view video synopsis is proposed. Five stages make up the framework. The first three steps make use of currently used techniques, and the latter two stages include new techniques. Performance comparison with existing systems has been done in terms of activity recognition accuracy, synopsis length reduction percentage, and quality assessment ratio. CNNs will discover the best filters for identifying particular objects and patterns. A CNN, however, learns more than one filter. In fact, it even learns multiple filters in each layer! Every filter learns a specific pattern, or feature.

REFERENCES

[1]. Shuai Zheng, Junge Zhang, Kaiqi Huang, Ran He and Tieniu Tan. Robust View Transformation Model for Gait Recognition. Proceedings of the IEEE International Conference on Image Processing, 2011

[2]. Shuai Zheng, Kaiqi Huang, and Tieniu Tan. Evaluation framework on translation-invariant representation for cumulative foot pressure image. Proceedings of the IEEE International Conference on Image Processing, 2011.

[3]. Shuai Zheng, Kaiqi Huang and Tieniu Tan. Translation Invariant Representation for Cumulative foot pressure Image, The second CJK Joint Workshop on Pattern Recognition(CJKPR), 2010.

[4]. Liming Shi, Shuai Zheng. Representation for cumulative foot pressure images, China Criminal Police Journal 2010

[5]. Shiqi Yu, Tieniu Tan, Kaiqi Huang, et.al.A Study on Gait-Based Gender Classification. IEEE Trans. on Image Processing. pp:1905-1910,V18(8),2009.

[6]. Shiqi Yu, Daoliang Tan, and Tieniu Tan. A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition. In Proc. of the 18'th International Conference on Pattern Recognition (ICPR06). Hong Kong, China. August 2006.

[7]. Daoliang Tan, Kaiqi Huang, Shiqi Yu, and Tieniu Tan. Efficient Night Gait Recognition Based on Template Matching. In Proc. of the 18'th International Conference on Pattern Recognition (ICPR06). Hong Kong, China. August 2006.

[8]. Shiqi Yu, Daoliang Tan, and Tieniu Tan. In Proc. of the 7'th Asian Conference on Computer Vision (ACCV06). Hyderabad, India. Jan. 2006.

[9]. Yuan Wang, Shiqi Yu, Yunhong Wang and Tieniu Tan. In Proc. of the International Conference on Biometrics 2006, pages 605-611. Hong Kong, China. Jan. 2006.