Mansoura Journal of Computers and Information Sciences

Video Analysis For Human Action Recognition Using Deep Convolutional Neural Networks

Nehal N.Mostafa
Faculty of computers and information systems , C.S dep.
Mansoura University, Egypt
nihalnabil1990@ gmail.com

Mohammed F. Alrahmawy
Faculty of computers and information systems , C.S dep.
Mansoura University, Egypt
mrahmawy@mans.edu.eg

Omaima Nomair
Faculty of computers and information systems , C.S dep.
Mansoura University, Egypt
omnomir@yahoo.com

ABSTRACT

In the last few years, human action recognition potential applications have been studied in many fields such as robotics, human computer interaction, and video surveillance systems and it has been evaluated as an active research area. This paper presents a recognition system using deep learning to recognize and identify human actions from video input.

The proposed system has been fine-tuned by partial training and dropout of the classification layer of Alexnet and replacing it by another one that use SVM. The performance of the network is boosted by using key frames that were extracted via applying Kalman filter during dataset augmentation. The proposed system resulted in oromising performance compared to the state of the art approaches. The classification accuracy reached 92.35%.

Keywords

Video analysis, Human action recognition, Deep learning, machine learning, video analysis, Image segmentation, feature extraction, Kalman filter, human action, Convolutional Neural Network.

1. INTRODUCTION

Human action recognition (HAR). refers to the automatic discovery of person behavior and the way he or she interacts with the surrounding environment. In the last few years, HAR has become an important research area due to the need for it as a prospective system in many branches like; automation for human behavior characterization, human computer interaction, and video surveillance systems. HAR requires many vitality recognition systems. HAR is the process of labeling action labels on image sequences. The process of human activity recognition from nonmoving images or video sequences has many challenges like appearance, lighting, viewpoint, changes in scale, partial occlusion, and background clutter. Another significantproblem in human action recognition is to first identify human within the video sequence [1], [2]. Deep learning has been considered as a new branch in machine learning since 2006.Deep learning is affecting everything from medicinal services to transportation to assembling, and that is only the beginning. Organizations are swinging to deep learning to solve of difficult issues, similar to machine interpretation, object recognition and, speech recognition.

Deep learning alludes to a somewhat broad category of machine learning architectures and techniques, with the sign of using multi non-linear layers information. Computers have techniques for image features recognition; however, the outcomes were not usually acceptable. Computer vision has been a primary recipient of deep learning. Nowadays, deep learning is used in computer vision to solve problems on most image recognition missions [3]. Three significant classes broadly categorize how the techniques and architectures are intended in deep learning [4]:

1. **Generative or unsupervised learning deep networks**,
2. **Supervised learning deep networks**,
3. **Hybrid deep networks**.

Recently, the techniques of deep learning have evolved and have influenced a wide range of work on signal and data implementation in their modern and standard form within the broad areas that include artificial intelligence and machine learning basic concepts. Deep learning is a subfield within the machine learning which applies learning algorithms to a multi-level representation. It models complex relationships within data, thereby defining high-level attributes and concepts based on what is inferior to them. The hierarchical structure of the deep learning and almost all other models rely on learning method without the supervision representations [5].

In this paper, we introduce a recognition system to recognize and identify human action from video input. The proposed system is built using deep convolutional neural networks. Alexnet architecture powers the introduced CNN architecture. The proposed method is integrated with Kalman filter to mark the original frames in input video. The proposed system has been fine-tuned by partial training and dropout classification layer which is replaced by another one.

The paper is structured as following: section 2 represents different correlated techniques and procedures that are used in the proposed architecture. The proposed system is presented in third section by specifying its basic theory and details. Section 4 represents the proposed system performance, and analyzes the results. In the last section (5), we conclude our work and overview the possibilities for the future work for practitioners and researchers.

2. REVIEW WORK

Many researchers have introduced different methods and models to recognize human action recognition. Their work has different perspectives with pros and cons. We present here some recent proposals and estimate some of their performance against our model.

Weilong Yang, et al [12] proposed a patch based matching algorithm for action recognition. They matched input clips and known actions template clips with a set of motion patches. They proved that the proposed matching scheme is operative for difficult case action recognition only for one action clip training. They also confirmed that accuracy and computational efficiency could be improved by learning a transferable weighting on these patches. These generic weights applied straight to new video sequences without furthermore learning. The experiments based on KTH dataset, and Weizmann dataset. The experiments result in 86.67% accuracy.

Lin Sun, et al [5] combined deep learning with slow feature analysis methods to learn hierarchical video data representations. They used a two-layered Slow Feature Analysis learning structure with three-dimension convolution and max pooling processes to measure the method to huge inputs and capture structural features and abstract from video. The experiments based on UCF Sports, KTH, and Hollywood2 datasets. The experiments result in 93.1% accuracy for the KTH dataset.

Jun Lei, et al [13] utilized Convolutional Neural Network (CNN) model to automatically learn directly features of high level action from raw inputs. The Latent Dynamic Conditional Random Field (LDCRF) model is used to model the extrinsic and intrinsic dynamics of actions. CNN embedded in bottom layer of LDCRF, which converts the structure of LDCRF from shallow to deep. This proposed system incorporates action feature learning and continuous action recognition procedures in a unified way. Their training model is in end-to-end fashion. The parameters of CNN and LDCRF optimized by gradient descent algorithm. The experiments based on KTH dataset, and HumanEva dataset. The experiments result in accuracy of 91.41% for KTH 2D action dataset and 93.2% for HumanEva dataset.

T.Subetha, et al [2] summarized Human Activity Recognition techniques, issues and challenges. They explored Human Activity Recognition systems variations like; Human-Human Interactions and Human Object Interactions. They analyzed experimental evaluation of many papers efficiently with various performance metrics like Precision, Recall and Accuracy.

Michalis Vrigkas, et al [1] provided a survey of recently developments in human activity recognition area. They also suggest a human activity categorization methodologies and explained the benefits and restrictions. Due to the data usage or not from dissimilar modalities, they divided the methods of human activity to two categories. To reproduce what type of activities they are concerned and how they model human activities, they explored every category into subcategories. They provided a complete analysis of the available datasets of human activity classification and study the dataset demands for perfect human activity recognition.

3. Problem formulation

Human Action Recognition (HAR) is a computer vision process. Since, most computer vision process includes preprocessing, a region of interest detection, segmentation, feature extraction and label assignment; HAR has applied the method of all listed subprocesses. To recognize an action that is represented as a label, input scene must be pre-processed and analyzed to detect if there is a person in the currentview. Due to human existence; which represents the region of interest, segmentation is required. In turn, feature extraction is applied as final phase before classification. In subsequent sections, we have discussed in details every process formore information.

4. THE PROPOSED METHDEOLOGY

4.1 Kalman Filter

Kalman filter offers a linear optimal filtering problem solution. This solution could be applied to non-stationary and stationary situations. It appraises the state that is calculated by earlier estimate and the new data inputs, thus earlier estimate only needs storage. The Kalman filter is more effective over calculating the forecast straitly of the entire last observed at each filtering process step data. Figure 1 represents the dynamical system conception of the discrete-time. The state vector, state, represented as x_k, is stated as nominal data set. It is adequate to define the natural dynamical system behavior individually; k specifies discrete time [6].

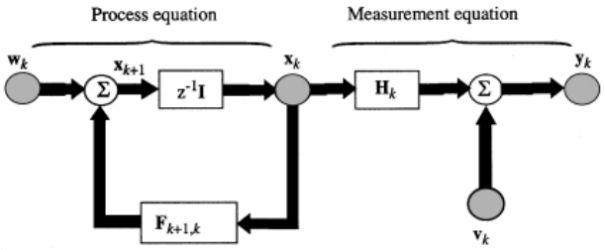

Fig1 : Discrete System State: Kalman Filter[6]

The state x_k is unidentified. To guess it, we used experiential data set, represented using vector y_k which is computed using process equation; Eq. (1) and measurement equation, Eq. (3).

$$x_{k+1} = F_{k+1,k}\, x_k + w_k \qquad \text{(Eq. 1)}$$

Where the transition matrix $F_{k+1,k}$ is taking the state x_k from time k to k+1. The noise of process w_k is expected to be improver, Gaussian, and white, with covariance matrix and zero mean defined by

$$E[w_n w_k^T] = \begin{cases} Q_k & for\ n = k \\ 0 & for\ n \neq k \end{cases} \qquad \text{(Eq.2)}$$

where T represents matrix transposition. The dimension of the state space is denoted by measurement equation represented as

$$y_k = H_k x_k + V_k \qquad \text{(Eq.3)}$$

where$_k$ is noticeable at time k and H_k represents measurement matrix. The noise of measurement v_k is expected to be improver, Gaussian, and white, with covariance matrix and zero mean defined by

$$E[V_n V_k^T] = \begin{cases} R_k & for\ n = k \\ 0 & for\ n \neq k \end{cases} \quad \text{(Eq.4)}$$

furthermore, the noise of measurement V_k is not correlated with the noise of process w_k. N represents measurement space dimension [6].

4.2Transfer learning

It is the system ability to recognize and apply to novel tasks, skills and knowledge learned in previous task. Fig. 2 represents the changes between transfer and traditional learning [7], [8].

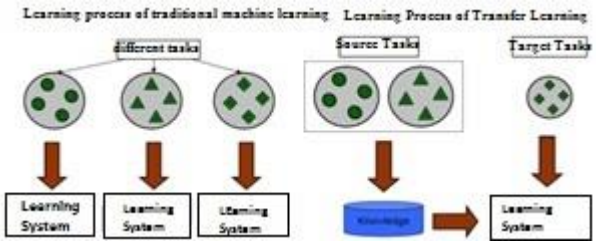

Fig2: Traditional vs. Deep learning[7]

Initially, we describe the notation the domain as D composed of two components: a marginal probability distribution P (X) and feature space X, where $X = \{x_1, \dots, x_n\} \in X$. In turn, if two domains are dissimilar, then they may have dissimilar feature spaces or dissimilar marginal probability distributions. Source domain data is represented as $D_S = \{(x_{S1}, y_{S1}), (x_{Sn}, y_{Sn})\}$, where $x_{Si} \in X_S$ is the instance of data and $y_{Si} \in Y_S$ is the resultant label of class. In turn, for classification example, D_S can be a set of term vectors together with their correlated true or false class labels. Likewise, target domain data is represented as $D_T = \{(x_{T1}, y_{T1}) \dots (x_{Tn}, y_{Tn})\}$, where the input x_{Ti} is in X_T and $y_{Ti} \in Y_T$ is the equivalent result [7].

4.3. Convolutional Neural Networks (CNN)

Linear processes and nonlinear processes set are convoluted in a weighted manner denoting convolutional neural network (CNN). CNN is encompassed of many convolutional layers followed by other completely associated layers like those in neural network. CNNs are trained simply and have lower parameters than other correlated networks using similar hidden units. convolutional neural network architecture is structured to have the vantages of any 2D input like image inputs or speech signal. CNNs are known for its substantial performance in applications as the visual tasks and natural language processing [4], [9]. These layers are learned in a joint manner [10].

A CNN composed of some subsampling layers optionally followed by totally connected layers and convolutional layers. The convolutional layer input is rx n x m image where n and m is the width and height of the image respectively, and r represents number of channels, e.g r equals to 3 in RGB image. The convolutional layer having k filters of size q x n x n where n is smaller than the image dimension and q is lower than channels number r or the same and each kernel it may differ. The size of filters provides increase to the correlated structure, each convolved with the image to produce k feature maps of size m−n+1. Each plan is then subsampled with mean or max pooling over p x p connecting expanses where p varieties from 2 and 5. Either before or after the subsampling layer an improver bias and sigmoidal nonlinearity are applied to every feature map. Figure 3show full CNN layer illustration is composed of subsampling and convolutional sublayers. Dots having different color denote different filter maps and dots having the same color have tied weights [9], [11].

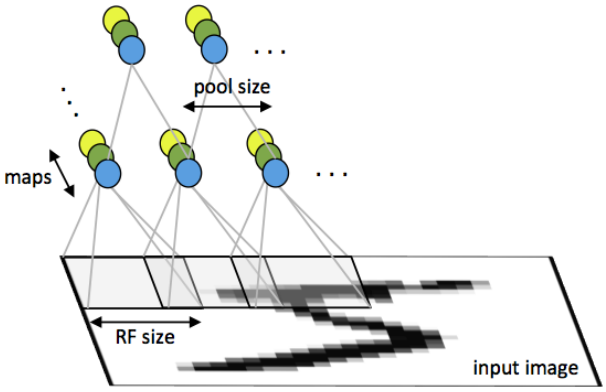

Fig 3: Full layer illustration in CNN

4.4 Human Action Recognition System

The proposed system is designed and developed over two stages; the first, includes preprocessing the acquired video and detect the region of interest ROI for recognizing scenes including actions using Kalman filter and marking them as keyframes in video streaming. While the second stage is the process of using Convolutional Neural Network (CNN) to segment ROI, extract visual features, update learning layers and apply classification based on Support Vector Machine (SVM). This paper proposes a hybrid Kalman-CNN model combining Kalman and CNN for continuous action recognition. We made complete usage of the strengths of them; Kalman is strong ability of detecting active formulation states and CNN's powerful ability of feature learning instead of the usual way of using the CNN, is by training CNN and

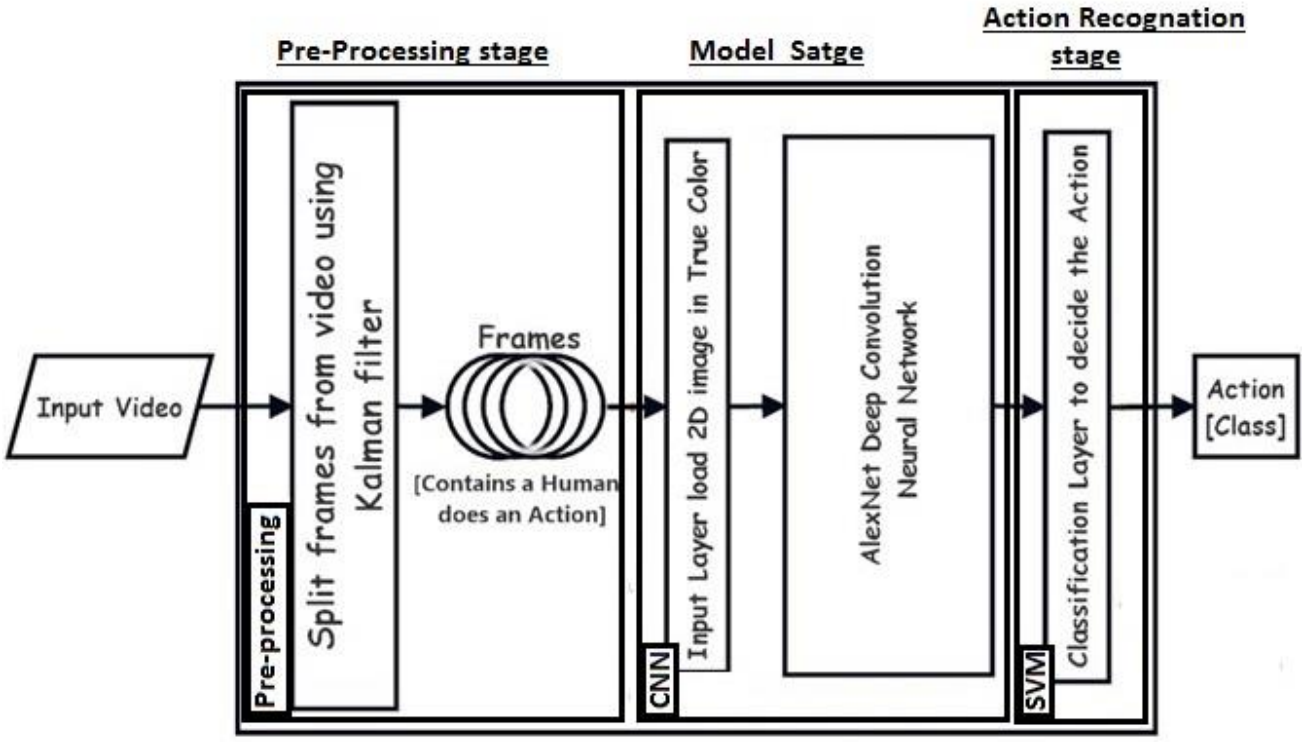

Fig 4 : Hybrid Kalman-CNN model

then uses the features learned by CNN to train a classifier for action recognition. we integrate the CNN and Kalman seamless in a unified framework. It integrates the feature learning and recognition procedures in an incorporated way. Figure 3 shows the abstract architecture of our *hybrid Kalman-CNN model*. The continuous action video is segmented into small clips, and each clip is inputted to the Kalman. The output scene from Kalman is the input nodes of CNN. The CNN infers the features representing the action in the video. We have adapted a predesigned CNN architecture to automatically segment and learn efficient and robust action features from keyframe that were marked by Kalman from raw video data. The CNN is trained guided by dynamic information that is available by Kalman, making the learningprocess more appropriate for dynamicmodeling.

Stage (1): Pre-processing

Since The Kalman filter formulates the unknown state problem of conjointly the measurement equations and process, the state can be expressed using the entire observed data; y_1; y_2; . . .; y_k, to calculate for every k+1which estimates the state; x_i. The Kalman filter estimates a linear dynamical system measurement, defined by equation (1) and equation (3). In HAR system, the problem is nonlinear, so we outspread the Kalman filtering usage over a linearization process. The outcome filter denoted as the extended Kalman filter (EKF) [14]. The main concept of the extended Kalman filter is the linearization of the state space model of equations(Eq.5) (Eq.6)

$$y_k = h(k, x_k) + V_k \quad \text{(Eq.5)}$$

$$x_{k+1} = f(k, x_k) + w_k \quad \text{(Eq. 6)}$$

for every round the greatest fresh state estimate is being considered. The typical Kalman filter equations are used, once a linear model is obtained, to mark current frame of the input video even if it contains a human doing action or not.

Stage (2): CNN using AlexNet (model)

Different works have been developed for extracting features for action representation. Deep models [3] have accomplished major achievements in different computer vision applications. CNN [10] [15] is validated as an effecient model for high-level features directly learninig from real data. One of the most common known CNN architecture is AlexNet; CNN Model. AlexNet is a deep CNN model, developed by Krizhevsky et al. [9], to model the 2012 ImageNet for the Large Scale Visual Recognition Challenge (ILSVRC-2012).

AlexNet was trained using more than 1.2 million images belonging to 1000 classes. AlexNet consists of

- Five convolutional layers which play an essential role in the operation of CNN. Every convolutional layer has activation function; a nonlinear ReLU layer is stacked after each layer.
- Three pooling layers that reduce the dimensionality of the representation, the number of parameters and the computational of the model complexity.
- Two normalization layers are stacked after the first and the second convolutional layers.

Three fully connected layers that were cited at the top of the model preceded by softmax layer.

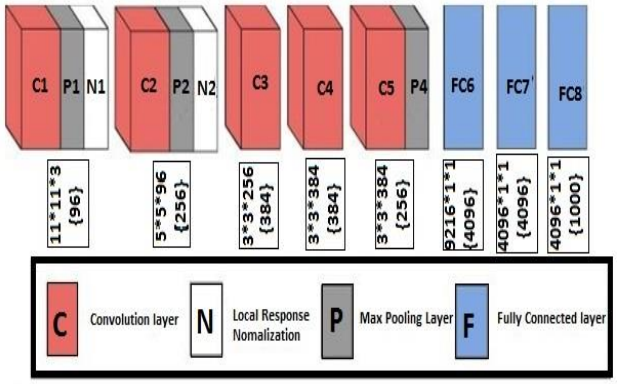

Fig 5 : CNN Architecture of AlexNet[8]

The proposed system is powered by AlexNet gaining the power of transfer learning. The proposed altered the classification layer to classify between 7 actions rather than 1000 image category. During the process of introducing the designed model, instead of learning a new model from scratch, the weights of AlexNet is used as an initialization to be fine-tuned using the dataset corresponding to our problem instead of the training process which requires a massive amount of labeled data as well as high processing power. In turn, a little learning rate; 5: 10 %, was used to update the weights of the convolutional layers. Stochastic Gradient Descent (SGD) algorithm is used to update the network weights using the human action dataset as the weights of these layers will not change dramatically. As for the fully-connected layers, the weightsare randomly initialized since they are considered data specific layers, unlike the convolution layers.

Stage (3): Action recognition stage

In the proposed system, we have transferred learned feature to build support vector machine. The proposed framework trains SVM classifier using extracted features set from the CNN. The derived feature sets are divided into training set and testing set. The division data have 80:20 rate. The division is performed in the random manner.

5. IMPLEMENTATION AND EVALUATION

The environmental implementations include software & hardware details. Hardware details are, PC with processor of core i7 and DDRAM of 8 GB. Software details are, 64-bit MATLAB 2016. To evaluate the proposed system, we have used dataset of KTH video that was provided by Schuldt et al. [16]. It contains six types of actions (boxing, running, jogging, walking, hand-clapping, and handwaving); which are shown in fig. 5. The size of images is 160 × 120 pix. The resolution of images is 25 frames per sec. There are substantial differences in viewpoint and duration. All actions were captured using similar backgrounds, but with hard shades. All sequences are stored using AVI format and are available online. Unpressed version is up on request. There are 600 (25x6x4) video files for every group of 25 subjects, four scenarios, and6 actions. Each file contains about four subsequences used as a sequence in the proposed system experiments.

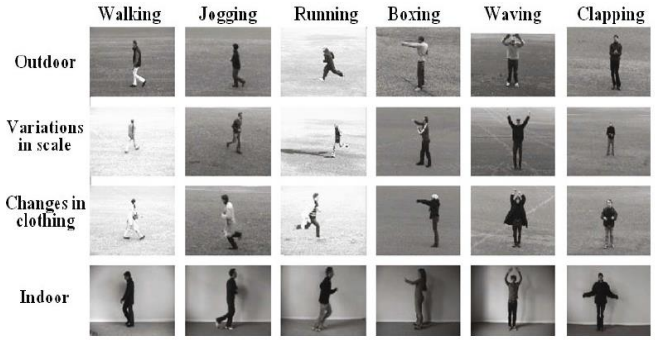

Fig 6 : KTH dataset sample(s)

To evaluate the proposed system, we have controducte many different types of experiments. There is about four scenarios seek optimization of accuracy level during the training phase. The proposed system has achieved the best score in different scenarios. Every scenario has input datasets; which is described in table 1; the dataset is split into 80% for training and 20% for testing randomly. In table 1, different data augmentation scenarios are described. Besides, the achieved level of accuracy is recorded to distinguish between different developed models.

Table 1. Enhancement of the proposed model

Scenario	Dataset Augmentation Description
A	Uses 60 percent of total number of video per action from the dataset and use all extracted frames from the video as input.
B	Uses 60 percent of total number of video per action from the dataset and use all frames from video plus increasing the number of frames by duplicating frames and rotating frames.
C	Use all videos in KTH dataset, use all extracted frames from videos
D	Use all videos in KTH dataset, use Kalman filters to extract frames including human.

Table 2 describes the accuracy of the proposed system over different phases. The accuracy was measured twice; during the training phase and the testing phase. Every resultshow that the system has higher rates of accuracy for pre-input samples. But, the proposed methods have lower rates during the testing phase. Model D is the best achievement in the experimental results. This model has been evaluated against the six actions which are available in the dataset. Figure 7 compares the accuracy of different models during the training verse the testing. Table 3 represents the convolutional matrix of the model D. The convolutional matrix shows high true positive recognition rate for all six actions.

Table 4 represents the comparison rates of the precision for a different model. The models have rates of precision per action. Model D has the highest rates per action against the other models; A, B and C. In models A, B and C, actions jogging, running and walking have large deviation in rates. The most stable rates for jogging, running, and walking exist for model D.

Table 2. Training vs. testing Accuracy of Different models

Model	Accuracy	Testing Accuracy
A	98.24	88
B	98.23	89.72
C	98.23	90.27
D	**98.21**	**92.35**

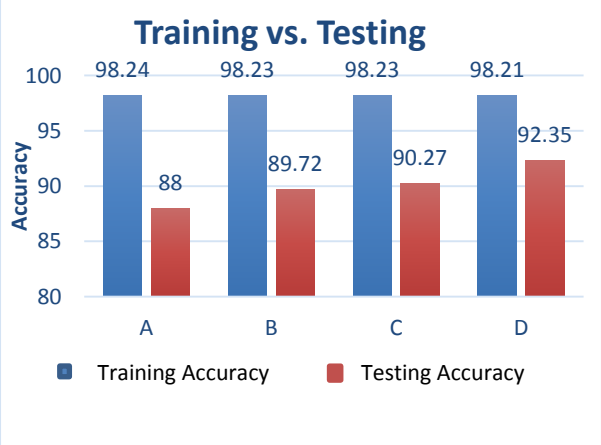

Fig 7 : Accuracy rates during training and testing

Table 3. Convolutional Matrix of the proposed System "Case D"

	Boxing	Handclapping	Hand waving	Jogging	Running	Walking
Boxing	**1013**	0	1	3	1	1
Handclapping	0	**991**	32	0	0	0
Hand waving	3	26	**983**	0	0	0
Jogging	1	2	1	**943**	170	104
Running	0	0	0	53	**838**	15
Walking	2	0	2	20	10	**899**

Table 4 Precision: Comparison evaluation of Different models

Model	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
A	0.98	0.97	0.97	0.79	0.75	0.99
B	0.93	0.96	0.98	0.85	0.93	0.8
C	1	0.98	0.94	0.87	0.9	0.78
D	0.99	0.97	0.97	0.77	0.92	0.96

Table 5 shows the model D is more sensitivity over the six actions. Although, the model D has 80% rates for running and walking against at least 93% for the rest actions. The model D achieves overall higher sensitivity rate. In table 6, a comparison of specificity rates for all models is represented. The model D gained the comparison and got the highest rates of models A, B, and C.

Table 5 Sensitivity: Comparison with Different models

Model	Boxing	Handclapping	Hand waving	Jogging	Running	Walking
A	1	0.98	0.97	0.77	0.92	0.77
B	1	0.98	0.96	0.76	0.77	0.97
C	0.96	0.94	0.98	0.76	0.83	0.98
D	0.99	0.97	0.96	0.93	0.82	0.88

Table 6 Specificity: Comparison evaluation of Different models

Model	Boxing	Handclapping	Hand waving	Jogging	Running	Walking
A	1	0.99	0.99	0.96	0.94	1
B	0.99	0.99	1	0.97	0.99	0.95
C	1	1	0.99	0.98	0.98	0.95
D	1	0.99	0.99	0.95	0.99	0.99

Table 7 compares the proposed model D against different models that were introduced by researchers using the same dataset; KTH.

Table 7. Comparative study versus the proposed

Figure 8 visualizes the comparison of the proposed model D verse the recent models that were proposed by scientists. The proposed model is more accurate than the proposed model by Jun Lei,et al.[13] in 2016 with ratio up to 1%.

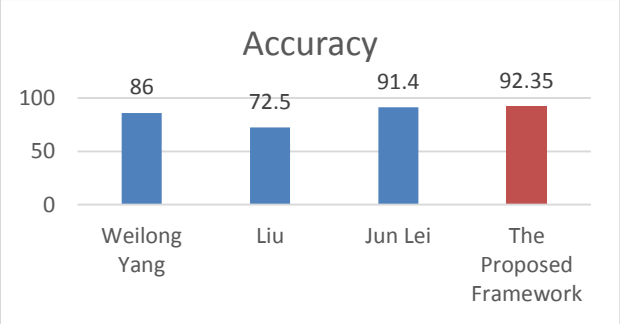

Fig 8 : Comparison of different accuracy levels

6. Conclusions and Future Work

This paper describes and evaluates a transfer learning based AlexNet for human action recognition. The architecture of the previously learned network was adapted to be compatible with the new classification problem. Human Action video dataset with six classes was used to investigate the learning performance. Dataset augmentation was usedto boost the performance of the network through detecting keyframes using Kalman filter, which has contributed to increasing the classification accuracy to 92.35%. Moreover, the proposed system admits to build SVM model utilizing learned feature from architecture of tuned AlexNet. Furthermore, it has also been proved that fine-tuning the fully connected layers of AlexNet is recommended than refining the complete network especially for relatively small datasets, which will also decrease the learning time.

7. REFERENCES

[1]M. Vrigkas, C. Nikou, and I. A. Kakadiaris, “A Review of Human Activity Recognition Methods,” Frontiers in Robotics and AI, vol. 2, p. 28, 2015.

[2]T. Subetha and S. Chitrakala, “A Survey on human activity recognition from videos,” in Information Communication and Embedded Systems (ICICES), 2016 International Conference on, 2016, pp. 1–7.

[3]G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” Neural computation, vol. 18, no. 7, pp. 1527–1554, 2006.

[4]S. Srinivas, R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. Kruthiventi, and R. V. Babu, “A taxonomy of deep convolutional neural nets for computer vision,” arXiv preprint arXiv:1601.06615, 2016.

[5]L. Sun, K. Jia, T.-H. Chan, Y. Fang, G. Wang, and S. Yan, “DL-SFA: deeply-learned slow feature analysis for action recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2625–2632.

[6]S. S. Haykin and others, Kalman filtering and neural networks. Wiley Online Library, 2001.

[7]S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2010.

[8]M. S. Elmahdy, S. S. Abdeldayem, and I. A. Yassine, “Low quality dermal image classification using transfer learning,” in Biomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on, 2017, pp. 373–376.

[9]A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[10]Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[11]U. K. Thikshaja and A. Paul, “A Brief Review on Deep Learning and Types of Implementation for Deep Learning,” Deep Learning Innovations and Their Convergence With Big Data, p. 20, 2017.

[12]W. Yang, Y. Wang, and G. Mori, “Human action recognition from a single clip per action,” in Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, 2009, pp. 482–489.

[13]J. Lei, G. Li, S. Li, D. Tu, and Q. Guo, “Continuous action recognition based on hybrid CNN-LDCRF model,” in Image, Vision and Computing (ICIVC), International Conference on, 2016, pp. 63–69.

[14]S. J. Julier and J. K. Uhlmann, “A new extension of the Kalman filter to nonlinear systems,” in Int. symp. aerospace/defense sensing, simul. and controls, 1997, vol. 3, no. 26, pp. 182–193.

[15]Y. LeCun, K. Kavukcuoglu, and C. Farabet, “Convolutional networks and applications in vision,” in Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on, 2010, pp. 253–256.

[16]C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: a local SVM approach,” in Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, 2004, vol. 3, pp. 32–36.

[17]L. Liu, L. Shao, X. Li, and K. Lu, “Learning spatio-temporal representations for action recognition: A genetic programming approach,” IEEE transactions on cybernetics, vol. 46, no. 1, pp. 158–170, 2016.