Journal of Computer Science

Review

Deep Learning Approach for Human Action Recognition Using Gated Recurrent Unit Neural Networks and Motion Analysis

[1]Neziha Jaouedi, [2]Noureddine Boujnah and [1]Med Salim Bouhlel

[1]*SETIT, Sfax, Tunisia*
[2]*FSG, Gabes, Tunisia*

Article history
Received: 19-09-2018
Revised: 31-10-2018
Accepted: 26-07-2019

Corresponding Author:
Neziha Jaouedi
SETIT, Sfax, Tunisia
Email: neziha_jaouedi@yahoo.fr

Abstract: Human action recognition is a computer vision task. The evaluation of action recognition algorithms relies on the proper extraction and learning of the data. The success of the deep learning and especially learning layer by layer led to many imposing results in several contexts that include neural network. Here the Recurrent Neural Networks (RNN) with hidden unit has demonstrated advanced performance on tasks as varied as image captioning and handwriting recognition. Specifically Gated Recurrent Unit (GRU) is able to learn and take advantage of sequential and temporal data required for video recognition. Moreover video sequence can be better described on both visual and moving features. In this paper, we present our approach for human action recognition based on fusion and combination of sequential visual features and moving path. We evaluate our technique on the challenging UCF Sports Action, UCF101 and KTH dataset for human action recognition and obtain competitive results.

Keywords: Deep Learning, Recurrent Neural Networks, Gated Recurrent Unit, Video Classification, Motion Detection

Introduction

In the last few years there has been a growing interest of recognizing human actions in real-world environment finds applications in a variety of domains including intelligent human-computer interactions (Metaxas and Zhang, 2013; Triki *et al*., 2012) and video surveillance (Cristani *et al*., 2012; Choi *et al*., 2013), customer attributes and shopping behavior analysis. However, precise recognition of actions is a extremely challenging task due to cluttered backgrounds and viewpoint variations. Therefore, we can mention, that the most popular state-of-the-art methods for human action recognition (Yang *et al*., 2015; Geng and Song, 2015) use engineered motion and texture descriptors calculated around spatio-temporal interest points. In addition, most of these approaches follow the conventional paradigm of pattern recognition, which consists of two steps in which the first step evaluates complex handcrafted features from video frames and the second step learns classifiers based on the obtained features. In real-world scenarios, it is infrequently known which features are very important for the task at hand, since the choice of features is highly problem-dependent. Particularly for human action recognition, different action classes may appear different in terms of their appearances and motion patterns.

In this paper, we demonstrate the importance of feature to classify and recognize human action, thus we show how to use two different approaches to extract complex features on sequences:

- Gaussian mixture model and Kalman filter
- Deep-learning recurrent neural network

In a recursive computation, these features should help to analyze the next frame in a video sequence. In addition, to a semantically meaningful localized content description, such features should represent descriptions of human motions. In this challenging, we evaluate our method on the UCF Sports action, UCF101 and KTH human action dataset where we find variety of action in different background. Overall, our contributions are as follows:

- We present the dataset used to evaluate our experimental resultats

- We show the first approach based on motion detection and tracking using GMM and kalman filter
- We show, the second approach, that our recurrent models are able to conserve an abstract state over time, track and interpret motion
- We merge the two approaches to improve the performance of our human action recognition method on much larger dataset

The remainder of this paper is organized as follows. In second section, we discuss related work. We evaluate our model in third section and discuss the results in fourth section.

Related Work

Despite the diversity of recognition methods, they are all based on two major steps in which the first step extract the features of all frames in a video sequence and the second classify the features obtained for recognize the action. The most well-known approaches in the literature to motion detection are ViBe methods ("VIsual Background Extractor") (Zivkovic and van der Heijden, 2006), KDE (Kernel Density Estimation) (Wren *et al.*, 1997) and the temporal averaging filter (Langmann *et al.*, 2010). Practically all this methods are limited by noise sensitive (nonstationary medium, climate change) and change of motion on sequences.

Ideally, methods based on background subtraction have achieved great success in many applications including complex motion. Particularly, the method proposed by Stauffer and Grimson (1999) which is Gaussian Mixture Model (GMM).

Moreover, the motion tracking is important to extract features. Therefore, once the object has been detected then it can be tracked along its path. A tracking moving objects can track and display the movement of objects. Such tracking complete every movement of the objects that it tracks and can thus provide important data, such as the speed and acceleration of an object. In addition, in recent years there has been much work on the tracking of moving objects with in a scene. Systems developed for such tasks as people tracking (Pnevmatikakis and Polymenakos, 2006; Wren *et al.*, 1997) face tracking (Ekenel and Pnevmatikakis, 2006) and vehicule tracking (Beymer *et al.*, 1997; Sullivan, 1994) have come in many shapes or sizes (Magee, 2001).

Kalman filtering (Liu *et al.*, 2007; Jang *et al.*, 2002) is very popular in the research field of navigation and aviation because of its magnificent accurate estimation characteristic. Since then, the Kalman filter is an estimator for what is called the "linear quadratic problem", which focuses on estimating the instantaneous "state" of a linear dynamic system perturbed by white noise. In addition, the Kalman tracker adapts the learning parameters of the adaptive background module. Moreover Kalman filter was succefully applied by Rameshbabu *et al.* (2012) to track human body while walking and detect her face. As well as Patel and Thakore (2013) uses kalman filter to track any single moving objects in indoor as well as in outdoor environments on surveillance video of CAVIAR and PETS datasets.

In last few years, deep learning has the most successful method used to learn a hierarchy of features (Geng and Song, 2015; Xu *et al.*, 2010) in many applications including video, image, speech and signal processing. Therefore, a feed forward neural network models can achieve state-of the-art accuracy in object classification, sometimes exceeding human-level performance. Models are trained by using a large set of labeled data and neural network architectures that contain many layers. Baccouche *et al.* (2011) introduces a multiple 3d Deep Learning of Spatio-Temporal Features for Human Action recognition which obtains a high performance on KTH dataset. In addition, the recent work of Deldjoo *et al.* (2017) uses deep learning for movies recommendation with successful results. However, sharing parameters across time is insufficient for capturing all of the correlations between input samples. Additionally, local connectivity limits the output to a function of a small number of neighboring input samples.

More recently, Recurrent Neural Networks (RNNs) have demonstrated great success in sequence labeling and prediction tasks such as handwriting recognition and language modeling. However, various types of hidden units for Recurrent Neural Networks (RNN) have been used to solve range of problems with impressive results in several applications involving sequential or temporal data.

The Long Short-Term Memory unit (LSTM) proposed by Hochreiter and Schmidhuber (1997) have been very successful with recurrent neural networks in diversity tasks such as speech recognition (Graves and Jaitly, 2014; Chorowski *et al.*, 2015), image and video captioning (Karpathy and Fei-Fei, 2014; Vinyals *et al.*, 2015; Venugopalan *et al.*, 2015) and handwriting recognition (Cristani *et al.*, 2012). Thus various architectures of LSTM networks have been created to optimize a variety of applications. Bidirectional LSTM (BLSTM) networks have been proposed by Alex Graves *et al.* (2099; Eyben *et al.*, 2009) for Frame wise phoneme classification and it is used (Wollmer *et al.*, 2011) to model a multi-stream framework for continuous conversational speech recognition.

LSTM Projected (LSTMP) have been proposed by Hasim Sak for Large Scale Acoustic Modeling (Sak *et al.*, 2014).

Cho *et al.* (2014) have been proposed, a simplify model of LSTM unit, the Gated Recurrent Unit (GRU). GRU is a lighter version of an LSTM where the complexity in the structure is reduced by decreasing the gates in the architecture. Units (GRU) have empirically demonstrated their ability to model long-term temporal dependency in various task on computer vision. Recently in 2017 Abien Fred (Agarap, 2017) have used A Neural Network Architecture Combining Gated Recurrent Unit (GRU) and Support Vector Machine (SVM) for Intrusion Detection in Network Traffic Data.

Our paper is included in the same context of Abien Fred, though we have used GRU for recurrent neural networks to extract sequential data and moving detection to extract human path using GMM and kalman filter.

Experiments

Human Action Dataset

We perform experiments on the UCF Sports Action, UCF101 and KTH datasets to demonstrate important characteristics of our model for action recognition.

UCF Sports Action: UCF Sports Action: UCF Sports is one of the earliest actions recognition dataset. It consists of a set of actions collected from various sports.

UCF101: UCF101 is the largest and the one of the most challenging action dataset in terms of actions and scale. It contains a variety of action with a large variations in camera motion and cluttered background.

KTH: The KTH dataset is the most popularly used public human actions dataset. It contains 6 types of actions (walking, jogging, running, boxing, hand-waving and handclapping).

Some of actions which used on experiments of our model are presented in Fig. 1.

Features Extraction

The goal of features extraction is to extract visual features vector and human moving features (Jaouedi *et al*., 2016a). We have considered vectors extracted from kalman filter and vectors extract from pre-trained deep-learning using Gated Recurrent Unit Neural Networks.

Human Motion Path

Background subtraction is a common first step in the field of video processing and motion tracking Fig. 2, however, it is used to reduce the effective image size in subsequent processing steps by segmenting the mostly static background from the moving or changing foreground. Background subtraction by GMM is a popular method to detect moving object in the video scenes. In addition, Gaussian is a more successful model to moving detection and moving path extraction. The probability density function using GMM model is given by:

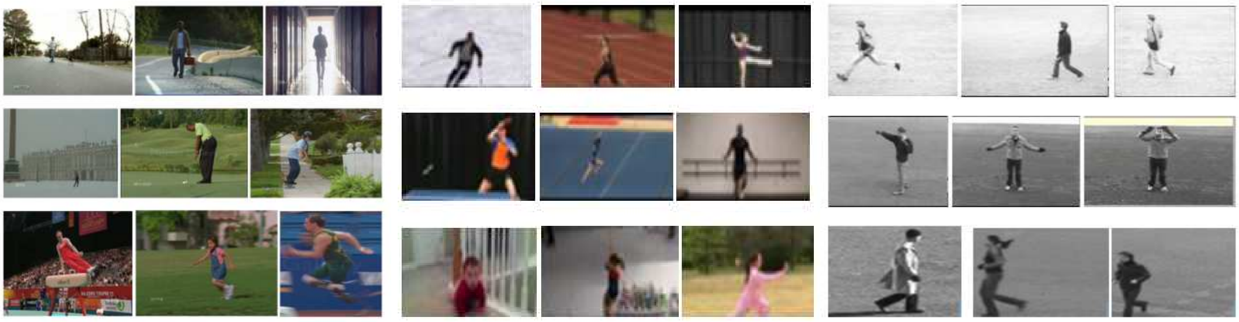
$$p\left(X_t\right)=\sum_{i=1}^{J}\omega_{i,t}\eta\left(X_t,\mu_{i,t},\sum{}_{i,t}\right) \quad (1)$$

where, $\omega_{i,t}$ is the weight of the *i*th Gaussian in the mixture at time t and $\sum_{i=1}^{J}\omega_{i,t}=1$. j represents the number of distributions and also the number of features in each image. $\eta(X_t, \mu_{i,t}, \Sigma_{i,t})$ is the density function of multivariate Gaussian probability with mean $\mu_{i,t}$ and covariance matrix $\Sigma_{i,t}$.

The probability density function of the multivariate Gaussian process is:

$$\eta\left(X_t,\mu_{i,t},\Sigma_{i,t}\right)=\frac{1}{(2\pi)^{\frac{n}{2}}\left|\Sigma\right|^{\frac{1}{2}}}e^{-\frac{1}{2}(X_t-\mu)^T\,\Sigma^{-1}(X_t-\mu)} \quad (2)$$

With n is the dimension of the pixel model.



UCF Sports Action UCF101 KTH

Fig. 1: Example frames of used datasets

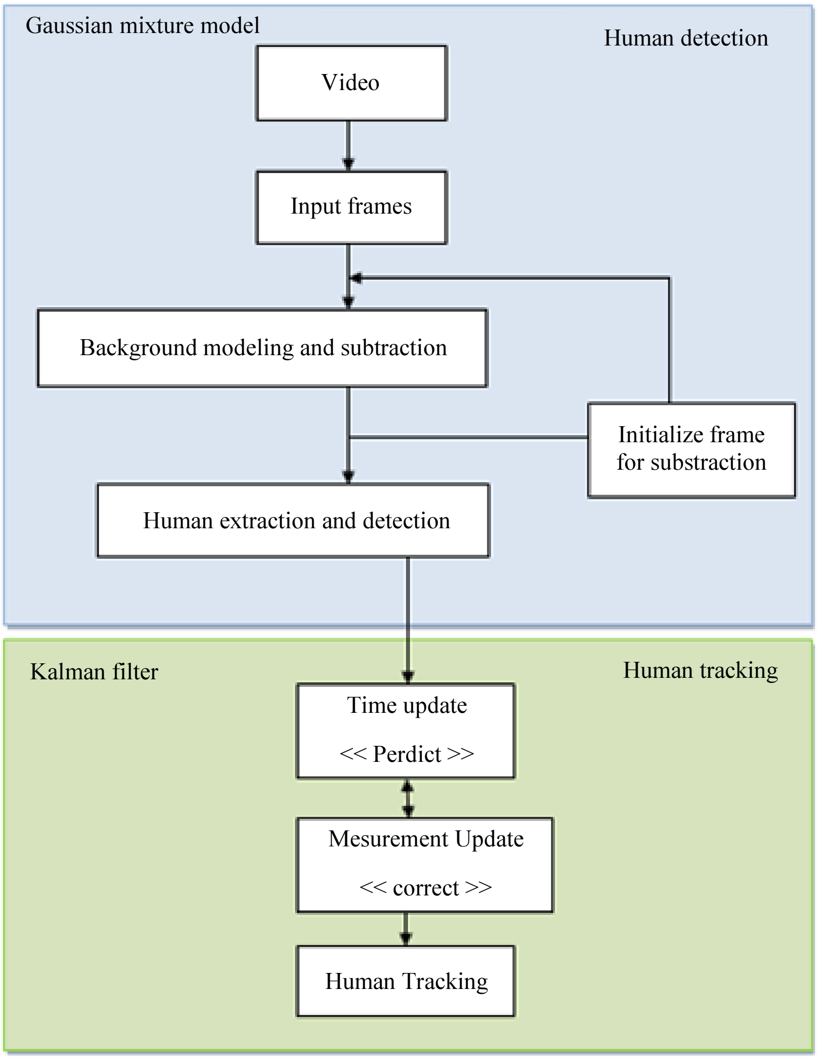


Fig. 2: Human motion path steps

Kalman Filter is the next step. This method is much related to GMM. Therefore, it extracts the location of a target feature in an image sequence over time. Kalman filter progresses cyclically in two phases: Prediction and correction. The prediction phase is to produce an estimate of the current state using the previous state. Our goal is to get a more accurate estimate. The state and observation equations are given by the following system:

$$\begin{cases} X_t = AX_{t-1} + \xi_t \\ Y_t = HX_t + \mu_t \end{cases} \tag{3}$$

where, $X_t = [x_t, y_t, v_t, w_t]^{\mathrm{T}}$, (x_t, y_t) is the position and (v_t, w_t) is the velocity vector. Y is the observation vector at time t. $[y_t, y_t]^T$ is the position vector. A is the state transition matrix. H is the measurement matrix, ξ_t stochastic process, modeling the state error $\xi_t \sim N(0,Q)$. μ_t stochastic process, modeling the observation $\mu_t \sim N$. The estimated value of X_t is modeled as:

$$\hat{X}_t = \hat{X}_t^- + G_t\left(Y_t - H\hat{X}_t^-\right) \tag{4}$$

Where:

$$\hat{X}_t^- = A\hat{X}_{t-1} \tag{5}$$

G_t is the filter gain at each iteration and it is updated as:

$$G_t = \hat{P}_t^- H^T\left(H\hat{P}_t^- H^T + R\right)^{-1} \tag{6}$$

Where:

$$\hat{P}_t = E\left[\left(X_t - \hat{X}_t^-\right)\left(X_t - \hat{X}_t^-\right)^T\right] = A\hat{P}_{t-1}A^T + Q \tag{7}$$

$$\hat{P}_t = E\left[\left(X_t - \hat{X}_t\right)\left(X_t - \hat{X}_t\right)^T\right] = \left(I - G_t H_t\right)\hat{P}_t^- \tag{8}$$

Deep Gated Recurrent Unit Networks Features

Gated Recurrent Unit (GRU) Architecture

A Gated Recurrent Unit (GRU) simplifies the gated recurrent neural network on two gates, a reset gate *r* and an update gate *z*. The reset gate determines how to combine the new input with the previous memory and the update gate control what to keep from the previous memory. The GRU structure can be seen in Fig. 3.

The Gated Recurrent Unit for Recurrent Neural Networks can be described with the hidden state and the candidate activation presented by the following equations:

$$h_t = \left(1 - z_t\right) * h_{t-1} + z_t * \tilde{h}_t \tag{9}$$

$$\tilde{h}_t = \tanh\left(W_h X_t + \left(r_t * U_h h_{t-1}\right) + b_h\right) \tag{10}$$

With the two gates presented as:

$$z_t = \sigma\left(W_z X_t + U_z h_{t-1} + b_z\right) \tag{11}$$

$$r_t = \sigma\left(W_r X_t + U_r h_{t-1} + b_r\right) \tag{12}$$

where, the variables are the external input vector X_t, the previous hidden state h_{t-1}, the parameters for two matrices *W* and *U* and the vector bias *b*. The total number of parameters for the candidate activation and the two gates are respectively, W_h, U_h, b_h, W_z, U_z, b_z, W_r, U_r and b_r. These parameters are all updated at each training step. The Activation functions used by Gated Recurrent Unit are the hyperbolic tangent function *tanh* and the logistic sigmoid function σ:

$$\sigma(x)\frac{1}{1+\exp(-x)} \tag{13}$$

GRU Recurrent Neural Networks Architecture

An alternative way to extract visual features in each time from video, we have used Recurrent Neural Networks with Gate Recurrent Unit. The proposed model presented in Fig.4 and may be summarized as follows:

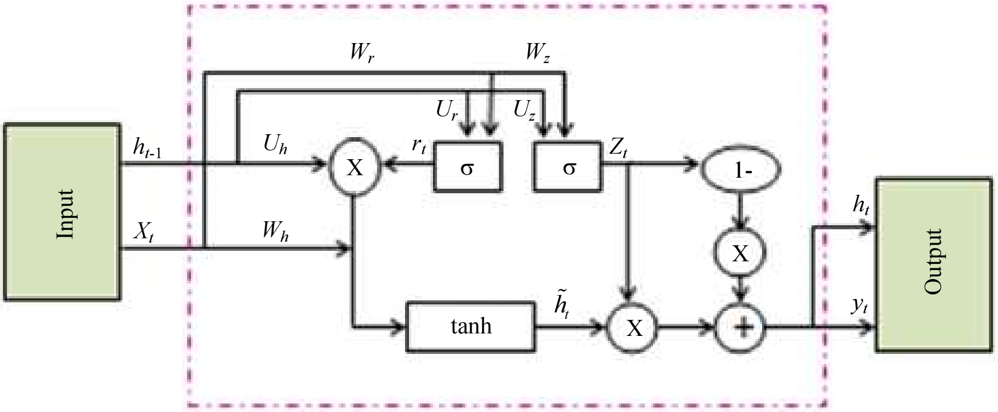
- Input layers present vector $X(X_1, X_2,.., X_n)$ of video dataset
- One hidden layer trained to classifier each input, which each Gate Recurrent Unit is connected to other at the next time step
- Output of our model is visual features vector Y $(Y_1, Y_2,.., Y_n)$ of input video

Features Fusion

In this step, from the same video, we employed the fusion of moving features obtained by kalman filter and visual features obtained by pre-trained of Gated Recurrent Unit for Recurrent Neural Networks. Therefore each video are represent by a fixed-length descriptor.

Human Action Recognition

To recognize a human action in sequence of video using Recurrent Neural Networks with Gated Recurrent Unit, we must trained model and classify action.



Grated recurrent unit memory block

Fig. 3: Gated recurrent unit architecture

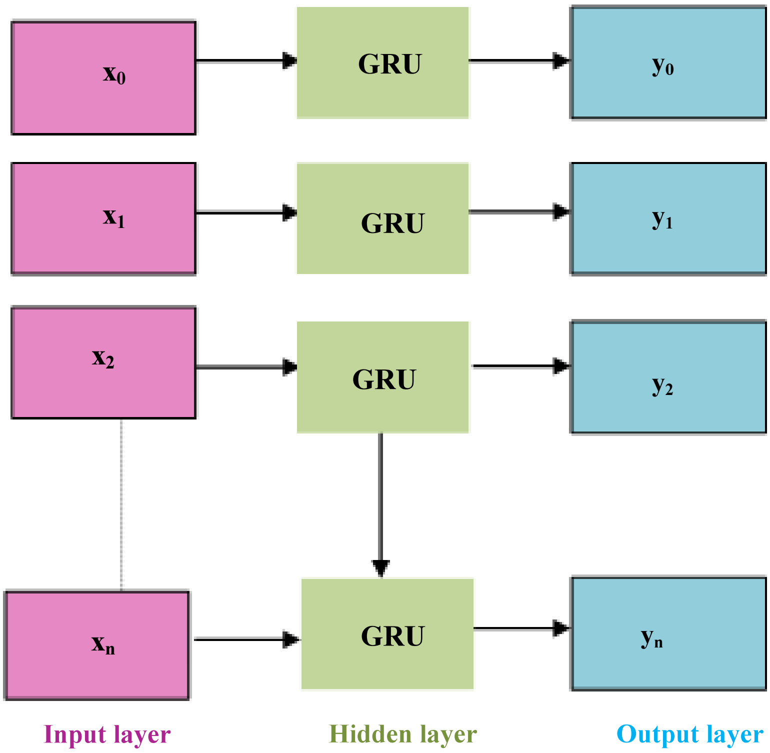


Fig. 4: Gated Recurrent Unit for Recurrent Neural Networks (GRU RNN) architecture

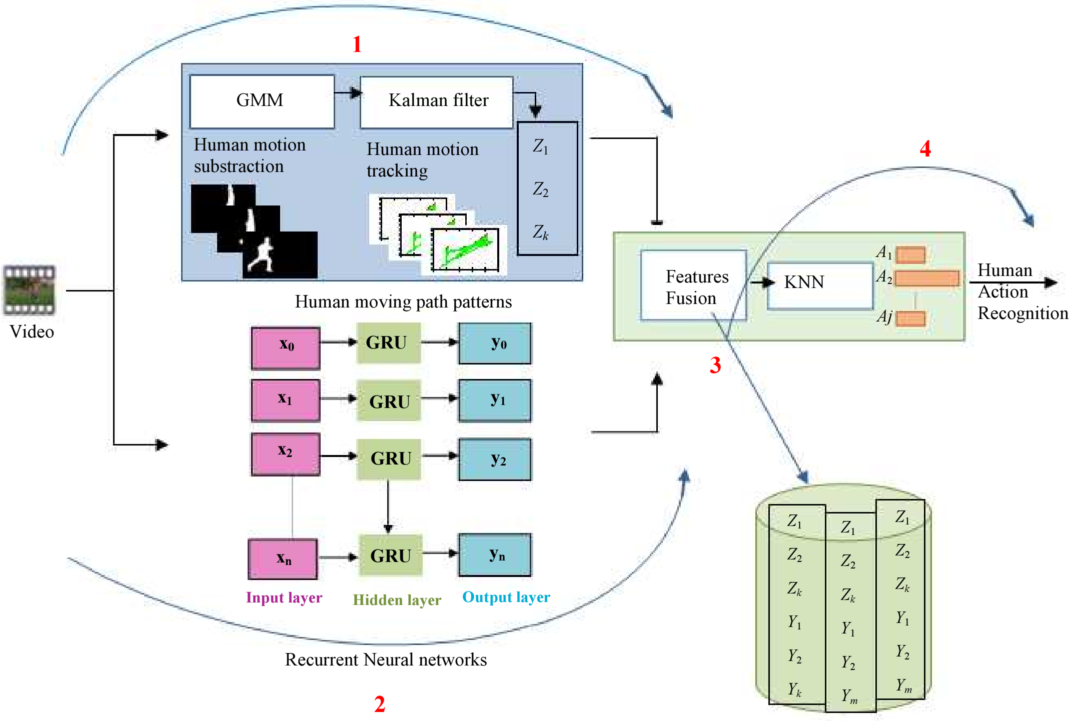


Fig. 5: Flowchart of the methodology used to recognize a human action

Training Phase

In the learning phase, we have three steps presented on Fig. 5:

Step1: The practice of Human moving path patterns, for each video, we detect human moving on each frame, we use the GMM method to color the human moving on white moving and the background on black color. This step reduces more time of video processing, in the next, we track human moving using kalman filter to obtain vector descriptor Z

Step2: Pre-training using Recurrent Neural Networks, in parallel way of human moving features detection, we extract Gated recurrent networks features. Inside, the proposed GRU-RNN model was trained as follows:

- Input the dataset features $X(X_1, X_2,.., X_n)$ to the GRU model
- Initialize the learning parameters weights and biases with arbitrary values (they will be adjusted through training)
- The cell states of GRU are computed based on the input features X_i and its learning parameters values
- Adjusts the weights and biases based on computed loss minimization between input and output GRU RNN
- Output the dataset features $Y(Y_1, Y_2,.., Y_n)$

Step3: Features fusion is the last step of learning phases; in this step we collect the moving features and visual features to the one length vector. Then we save the mixture features vector of each video on features database.

Classification Phase

In the practice of machine learning, the use of classification algorithms is more and more evaluated. The classification phase has the step 1, step 2 and step 4 of Fig. 5.

Step4: We project a collect of moving and visual feature of test video (step 1 and step 2) on feature dataset saved in step 3 and to classifier human action we have used K-NN classification method (Jaouedi *et al*., 2016b).

The principle of K-NN is to make a comparison between the features vector of test video and all of vectors on database features. The outputs of our approach model are the distribution of a probability of each class which represent a human action. Finally, we decide to which class, which has a max of probability, the test video belongs.

Discuses Results

To verify this method, we used 300 videos of 15 human actions on three datasets. To recognize a human action we used the distribution Three-quarters of videos to learning and one-quarter to test. The measurement results of classification rates used 5 actions of UCF Sports Action dataset (K = 5) are summarized in Table 1. Indeed the classification rate achieved 93, 4% for skate boarding-front. Then the results of human action classify using 5 actions of UCF101dataset (K = 5) are presented in Table 2, where we found the high classification rate 91% for skiing action. Finally the results of 5 actions of KTH (K = 5) dataset are resumed in Table 3 where the walking action have a high classification rate 98, 8%.

Overall, our experiments prove the effectiveness of human action recognition based on fusion of moving features and visual features automatically extracted from trailers of sequences.

Comparison with State of the Art

Table 4 shows the performance comparison with other approaches on the same KTH dataset. Indeed our approach based on fusion of visual features and moving features achieved classification rate 96, 70%, Overall, our method gives comparable results with the best related work on KTH dataset like CNN (Convolutional Neural Networks) method (Geng and Song, 2015) 92,49%, 3D CNN (Rameshbabu *et al*., 2012) 90,2% and 3D CNN + LSTM (Long Short-Term Memory) (Patel and Thakore, 2013). However, fusion of features is much better identifying human action.

Table 1: Classification results for UCF sports action

Human actions	Golf-swing-front	Kicking-side	Run-Side	Skate boarding-front	Walk-front	Average
Classification rate	90%	82%	87%	93,4%	92,6%	89%

Table 2: Classification results for UCF101

Human actions	Jumping jack	Swing	Baby crawling	Skiing	Javelin throw	Average
Classification rate	91%	80,1%	80%	89,5%	75%	83,10%

Table 3: Classification results for KTH

Human actions	Boxing	Running	Walking	Hand clapping	Hand-waving	Average
Classification rate	96%	97,4%	98,8%	95,8%	95,6%	96,70%

Table 4: Performance comparison of our approach

Methods	Classification rate
Our approach	96,70%
CNN	92,49%
3D CNN	90,2%
3D CNN + LSTM	94,39%

Conclusion and Future Work

This work presents a novel approach in the domain of human action recognition. The technique is based on the analysis of video content and extraction of two types of features. Moving feature based on moving detection and moving tracking using GMM and kalman methods. Then visual features based on all visual characteristic of each frame on video sequence using model of deep learning net works. The results of this paper provide a strong importance of collection and fusion features to obtain high classification rate and recognize human action.

For future work, the goal is to evaluate our approach to obtain the same classification rate on other datasets. Moreover, we are interested to investigate the possible improvement of human action based on the features provided by the different models of deep-learning networks.

Acknowledgment

This work was supported and financing by the Ministry of Higher Education and Scientific Research of Tunisia.

Author's Contributions

Neziha Jaouedi: Written and participated in all experiments of the manuscript.

Noureddine Boujnah: Designed the research plan, Corrected and revisited the manuscript.

Med Salim Bouhlel: Designed the research plan and organized the study.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and there are no ethical issues involved.

References

Agarap, A.F.M., 2017. A neural network architecture combining Gated Recurrent Unit (GRU) and Support Vector Machine (SVM) for intrusion detection in network traffic data. avXiv:1709.03082v6.

Baccouche, M., F. Mamalet, C. Wolf, C. Garcia and A. Baskurt, 2011. Sequential deep learning for human action recognition. Proceedings of the 2nd International Conference on Human Behavior Unterstanding, (HBU' 11), Springer, Amsterdam, The Netherlands, pp: 29-39. DOI: 10.1007/978-3-642-25446-8_4

Beymer, D., P. McLauchlan, B. Coifman and J. Malik, 1997. A real-time computer vision system for measuring traffic parameters. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jun. 17-19, IEEE Xplore Press, San Juan, Puerto Rico, USA, pp: 495-501. DOI: 10.1109/CVPR.1997.609371.

Cho, K., B. van Merienboer, C. Gulcehre, D. Bahdanau and F. Bougares *et al*., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. Proceedings of the Empirical Methods in Natural Language Processing, Oct. 25-29, Association for Computational Linguistics, Doha, Qatar, pp: 1724-1735. DOI: 10.3115/v1/D14-1179

Choi, S., E. Kim and S. Oh, 2013. Human behavior prediction for smart homes using deep learning. Proceedings of the IEEE RO-MAN, Aug. 26-29, IEEE Xplore Press, Gyeongju, South Korea, pp: 173-179. DOI: 10.1109/ROMAN.2013.6628440

Chorowski, J.K., D. Bahdanau, D. Serdyuk, K. Cho and Y. Bengio, 2015. Attention-based models for speech recognition. Advances in Neural Information Processing Systems.

Cristani, M., R. Raghavendra, A.Del BueVittorio Murino, 2012. Human behavior analysis in video surveillance: A Social Signal Processing perspective. Neurocomputing, 100: 86-97. DOI: 10.1016/j.neucom.2011.12.038

Deldjoo, Y., M. Quadrana, M. Elahi and P. Cremonesi, 2017. Using mise –en –scene visual features based on MPEG7 and deep learning for movie recommendation. ArXiv:1704.06109.

Ekenel, H. and A. Pnevmatikakis, 2006. Video-based face recognition evaluation in the CHIL project. Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, Apr. 10-12, IEEE Xplore Press, Southampton, UK., pp: 90-90. DOI: 10.1109/FGR.2006.110

Eyben, F., M. Wollmer, B. Schuller and A. Graves, 2009. From speech to letters - using a novel neural network architecture for grapheme based ASR. Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Nov. 13-Dec. 17, IEEE Xplore Press, Merano, Italy, pp: 376-380. DOI: 10.1109/ASRU.2009.5373257

Geng, C. and J.X. Song, 2015. Human action recognition based on convolutional neural networks with a convolutional auto-encoder. Proceedings of the 5th International Conference on Computer Sciences and Automation Engineering, (SAE' 15), Atlantis Press. DOI: 10.2991/iccsae-15.2016.173

Graves, A. and N. Jaitly, 2014. Towards end-to-end speech recognition with recurrent neural networks. Proceedings of the 31st International Conference on International Conference on Machine Learning, Jun. 21-26, JMLR.org, Beijing, China, pp: II-1764-II-1772.

Graves, A., M. Liwicki, S. Fernández, R. Bertolami and H. Bunke *et al*., 2009. A novel connectionist system for unconstrained handwriting recognition. IEEE Trans. Patt. Anal. Mach. Intell., 31: 855-868. DOI: 10.1109/TPAMI.2008.137

Hochreiter, S. and J. Schmidhuber, 1997. Long short-term memory. Neural Comput., 9: 1735-1780. DOI: 10.1162/neco.1997.9.8.1735

Jang, D.S., S.W. Jang and H.I. Choi, 2002. 2D human body tracking with structural kalman filter. Patt. Recognit., 35: 2041-2049. DOI: 10.1016/S0031-3203(01)00201-1

Jaouedi, N., S. Zaghbani, N. Boujnah and M.S. Bouhlel, 2016a. Human motion detection and tracking. Proceedings of the 9th International Conference on Machine Vision, (CMV' 16), Nice, France. DOI: 10.1117/12.2268539

Jaouedi, N., N. Boujnah, O. Htiwich and M.S. Bouhlel, 2016b. Human action recognition to human behavior analysis. Proceedings of the 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, Dec. 18-20, IEEE Xplore Press, Hammamet, Tunisia, pp: 263-266. DOI: 10.1109/SETIT.2016.7939877

Karpathy, A. and L. Fei-Fei, 2014. Deep visual semantic alignments for generating image descriptions. arXiv preprint arXiv:1412.2306.

Langmann, B., S.E. Ghobadi, K. Hartmann and O. Loffeld, 2010. Multi-Modal Background Subtraction Using Gaussian. In: Mixture Models, Paparoditis, N., M. Pierrot-Deseilligny, C. Mallet and O. Tournaire, (Eds.), IAPRS, Saint-Mandé, France.

Liu, G., X. Tang, J. Huang, J. Liu and D. Sun, 2007. Hierarchical model-based human motion tracking via unscented kalman filter. Proceedings of the International Conference on Computer Vision, Oct. 14-21, IEEE Xplore Press, Rio de Janeiro, Brazil, pp: 1-8. DOI: 10.1109/ICCV.2007.4408941

Liwicki, M., A. Graves, H. Bunke and J. Schmidhuber, 2007. A novel approach to on-line handwriting recognition based on bidirectional long short term memory networks. Proceedings of the 9th International Conference on Document Analysis and Recognition, (DAR' 07), pp: 367-371.

Magee, D.R., 2001. Tracking multiple vehicles using foreground, background and motion models. Proceedings of the European Conference on Computer Vision, (CCV' 01).

Metaxas, D. and S. Zhang, 2013. A review of motion analysis methods for human Non verbal communication computing. Image Vis. Comput., 31: 421-433. DOI: 10.1016/j.imavis.2013.03.005

Patel, H.A. and D.G. Thakore, 2013. Moving object tracking using kalman filter. Int. J. Comput. Sci. Mobile Comput., 2: 326-332.

Pnevmatikakis, A. and L. Polymenakos, 2006. 2D person tracking using Kalman filtering and adaptive background learning in a feedback loop. Proceedings of the 1st International Evaluation Conference on Classification of Events, Activities and Relationships, Apr. 06-07, Springer, Southampton, UK, pp: 151-160. DOI: 10.1007/978-3-540-69568-4_11

Rameshbabu, K., J. Swarnadurga, G. Archana and K. Menaka, 2012. Target tracking system using kalman filter. Int. J. Adv. Eng. Res. Stud., 2: 90-94.

Sak, H., A. Senior and F. Beaufays, 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. 15th Annual Conference of the International Speech Communication Association, Sept. 14-18, Singapore, pp: 338-342.

Stauffer, C. and W.E.L. Grimson, 1999. Adaptive background mixture models for real-time tracking. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jun. 23-25, IEEE Xplore Press, Fort Collins, CO, USA, pp: 246-252. DOI: 10.1109/CVPR.1999.784637

Sullivan, G.D., 1994. Model-based Vision for Traffic Scenes Using the Ground-Plane Constraint. In: Real-time Computer Vision, Terzopoulos, D. and C. Brown (Eds.), Cambridge University Press, pp: 93-115.

Triki, N., M. Kallel and M.S. Bouhlel, 2012. Imaging and HMI: Fondations and complementarities. Proceedings of the 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications, Mar. 21-24, IEEE Xplore Press, Sousse, Tunisia, pp: 25-29. DOI: 10.1109/SETIT.2012.6481884

Venugopalan, S., M. Rohrbach, J. Donahue, R. Mooney and T. Darrell *et al*., 2015. Sequence to sequence video to text. arXiv preprint arXiv:1505.00487.

Vinyals, O., A. Toshev, S. Bengio and D. Erhan, 2015. Show and tell: A neural image caption generator. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 7-12, IEEE Xplore Press, Boston, MA, USA, pp: 3156-3164. DOI: 10.1109/CVPR.2015.7298935

Wollmer, M., F. Eyben, B. Schuller and G. Rigoll, 2011. A multi-stream ASR framework for BLSTM modeling of conversational speech. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 22-27, IEEE Xplore Press, Prague, Czech Republic, pp: 4860-4863. DOI: 10.1109/ICASSP.2011.5947444

Wren, C., A. Azarbayejani, T. Darrell and A. Pentland, 1997. Pfinder: Real-time tracking of the human body. IEEE Trans. Patt. Anal. Mach. Intell., 19: 780-785. DOI: 10.1109/34.598236

Xu, W., M. Yang and K. Yu, 2010. 3D convolutional neural networks for human action recognition. Proceedings of the 27th International Conference on Machine Learning, Jun. 21-24, Haifa, Israel, pp: 495-502. DOI: 10.1109/TPAMI.2012.59

Yang, J.B., M.N. Nguyen, P. San, X. Li and S. Krishnaswamy, 2015. Deep convolutional neural networks on multichannel time series for human activity recognition. Proceedings of the 24th International Conference on Artificial Intelligence, Jul. 25-31, AAAI Press, Buenos Aires, Argentina, pp: 3995-4001.

Zivkovic, Z. and F. van der Heijden, 2006. Efficient adaptive density estimation per image pixel for the task of background subtraction. Patt. Recognit. Lett., 27: 773-780. DOI: 10.1016/j.patrec.2005.11.005