Journal of Korea Multimedia Society Vol. 23, No. 10, October 2020(pp. 1236-1249)
https://doi.org/10.9717/kmms.2020.23.10.1236

LSTM Network with Tracking Association for Multi-Object Tracking

Xurshedjon Farhodov[†], Kwang-Seok Moon[††], Suk-Hwan Lee[†††], Ki-Ryong Kwon[††††]

ABSTRACT

In a most recent object tracking research work, applying Convolutional Neural Network and Recurrent Neural Network-based strategies become relevant for resolving the noticeable challenges in it, like, occlusion, motion, object, and camera viewpoint variations, changing several targets, lighting variations. In this paper, the LSTM Network-based Tracking association method has proposed where the technique capable of real-time multi-object tracking by creating one of the useful LSTM networks that associated with tracking, which supports the long term tracking along with solving challenges. The LSTM network is a different neural network defined in Keras as a sequence of layers, where the Sequential classes would be a container for these layers. This purposing network structure builds with the integration of tracking association on Keras neural-network library. The tracking process has been associated with the LSTM Network feature learning output and obtained outstanding real-time detection and tracking performance. In this work, the main focus was learning trackable objects locations, appearance, and motion details, then predicting the feature location of objects on boxes according to their initial position. The performance of the joint object tracking system has shown that the LSTM network is more powerful and capable of working on a real-time multi-object tracking process.

Key words: Deep Learning, Object Tracking, LSTM Network, RNN, Object Detection, Multi-Object Tracking, MOT, CNN, Keras, Dense Layer, Neural Network

1. INTRODUCTION

Most of multi object tracking research outcomes has been shown that working with multiple objects the same time is more challenging and demanding task, when applying neural network-based approach. There are number of indeterminable and uncertain circumstances which we may face some difficulties to handle with RNN or CNN based tracking techniques. However, most of state-of-the-art approaches has been developed with the deep learning network [1-4]. Even though multi-object tracking systems are more complex and computationally overloaded for the real tools in case of working on accuracy measures and highly leveled functionalities of tracking systems. A number of attempts to figure out multi-object tracking most common challenges, such as, similar appearance, frequent occlusion, motion of several objects the same time, edge problems realized that adaption of deep learning approaches can face an additional charges that for analyzing unnecessary information from the video sequence and discrete set of detections. Moreover, here comes one more

※ Corresponding Author : Ki-Ryong Kwon, Address: (48513) 45 Yongso-ro, Nam-gu, Busan, Pukyong National University, Korea, TEL : +82-51-629-6257, FAX : +82-51-629-6230, E-mail : kiryongkwon@gmail.com
Receipt date : May. 11, 2020, Revision date : Sep. 2, 2020
Approval date : Sep. 22, 2020
[†] Dept. of IT Convergence and Application Engineering, Pukyong National University
(E-mail : life9940502@gmail.com)
[††] Dept. of Electronics Engineering, Pukyong National University
(E-mail : ksmoon@pknu.ac.kr)
[†††] Dept. of Computer Engineering, Donga University
(E-mail : skylee@dau.ac.kr)
[††††] Dept. of IT Convergence and Application Engineering, Pukyong National University
※ This work was supported by a Research Grant of Pukyong National University (2019).

strategy of work is tracking by detection [5], whereas due to various challenges such as view point change, scales, density of object distribution and occlusion, they have proposed a model for detection as a channel interdependencies by using "Squeeze-and-Excitation" (SE) blocks that adaptively recalibrates channel-wise feature responses which works with customized Deep SORT network for object detection along with deep association network for their tracking algorithm.

Our recent research in multi-object tracking has been achieved remarkable effectiveness along with high accuracy. Main steps of our suggesting strategy is given Fig. 1 below. The main goal in multi-object tracking is tracking or identifying multiple and different type of objects in video sequence. Currently, most of resent multi object tracking techniques that relayed on tracking by detection approach. However, those methods are not much efficient while their performances are fully depends on state-of-the-art detectors accuracy and proficiency of accomplishment when they detect in a different environment. These methods are mostly localization based on bounding box recovery, but that kind of methods can be helpful for finding an object on a flat image surface. In multi-object tracking process runs with a large number of parameters, where the deep detection models requires huge amounts of inputs that can make the state-of-the-art approaches impasse or losing essential feature elements of object while tracking objects across in video frames. Our proposing model which we are going to describe, is contains two main part that a LSTM Network and tracking association which gives us a fully tracking results. The network we have created is formed with LSTM cell based layers that to explore the training set of image sequences. Main idea of proposed method is to focus on the objects appearance and its location, by estimating the initial location of objects in starting point and saving it into network cells along with classified features of the sequence frames. Moreover, the whole process of learning and tracking steps are separately divided, initial process is to train our network with training dataset in offline mode and in a second step tracking multi-objects with the integration of tracking association. Tracking process goes online across a sequential video frames tracked objects with identical numbers. Our LSTM network is applicable to get entire sequences of data to analyze and finding tracking object features and remembering values of bounding boxes over arbitrary time intervals into cells than will be controlled by three gates that the flow of coming information into and out of the cell.

Nowadays, the importance of multi object tracking is more urgent task than ever, there are some fields that requires controlling and tracking identities of several type of objects. Generally, in such methods of tracking requires to use common and complex datasets in order to analyze problem in-depth and multiple cases. We have used two open source datasets for training and testing our observation in detail, these are KITTI [6] and MOT2016 [7]. These datasets are quite common and eligible to test our approach, allows to observe

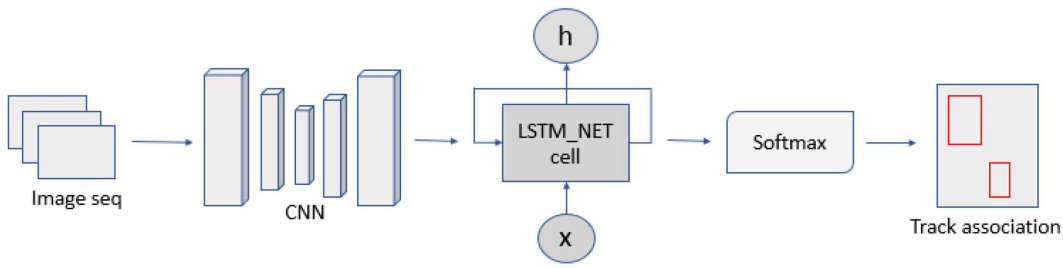

Fig. 1. Main structure of the proposed tracking structure.

our strategy with common real-time tracking performance. The results of our experiment bared that in multi object tracking identifying every objects in frames and assigning them as identical item with their appearance is more complex to get accurate outputs. However, the performance of our tracking system were more accurate and high performance results with visually qualitive outputs.

In the following section we are going to explain our proposing method in detail with mathematical formulation and along with some basics of our approach, related works as well. Moreover, after this step we are going to explain experimental results with all datasets along with further discussion.

2. LSTM NETWORK WITH TRACKING ASSOCIATION

In multi-object tracking there are number of classical problems to overcome across trackable frames. The main goal of our work is to create a simple online multi-object tracking platform that would be able to solve some of basic problems, like appearance, motion, edge detection and so on. Additionally, we are going to identify classes of trackable objects by taking the last step of the prediction value which follows through with at the end of the sequence. Tracking process is associated with prediction part of the tracking system that works correspondingly by providing predicted object features, object classes as well. Our full proposed method are illustrated above step by step in Fig. 1.

2.1 LSTM basics and related work

Currently, tracking-by-detection method has been succeed as one of the most successful technique in field of Computer Vision by the result of contemporary research strategies of object detection [8], methods and its integration with other tracking algorithms [9]. Additionally, to overcome that kind of problems in recent multi-object tracking approaches started improving their technique with data association [10], to provide tracking system with detected object information in terms of continuously tracking. As we mentioned before, essential part in multi-object tracking is memory unit, where we are going to detect and track several objects that similar to each other. In visual multi object tracking identifying each object as one an identical object is important, and similarity between multi objects has considerable role in case of identify them separately [11]. Exploring the entire image as a whole grid instead of computing exactly object located regions of interest will decrease the accuracy of detection and may lead overwhelming of object features and misclassification.

In the last few years, one more technique has fascinated several field researchers who working on with sequential data modeling with different output. It's called LSTM that was proposed in 1997 by Sepp Hoch Reiter and Jurgen Schmid Huber [12]. The main idea in LSTM has been introduced by solving Constant Error Carousel (CEC) units, that deals with the vanishing gradient problem. Basically, LSTM one of the type of an artificial Recurrent Neural Network's class architecture that contains basic elements of RNN and widely used in the field of deep learning. Because of its capability we can apply it not only process of single data points (such as images), alternatively we can apply this architecture to entire sequential data, such as speech or video processing. For example, there are several LSTM based works has been purposed in a different field of sequential data processing, such as handwriting recognition, speech recognition, and anomaly detection in network traffic or IDS(Intrusion Detection Systems) [13]. Most of regular used LSTM architecture is composed of a cell, an input gate, an output gate and a forget gate, where the entire process of regulation flow of information into and out of the cell will be conducted by over arbitrary time intervals remembering learned values of the sequential data.

There are some well-known and popular LSTM networks that well-suited for classifying, processing and making predictions based on time series data. Because of the indefinite sequential data time series or unknown duration between important events in a time series, LSTM network capable of taking vital part of the sequence data and dealing with the vanishing gradient problem that can be encountered when training traditional RNNs. Common architecture of the LSTM illustrated below in Fig. 2 [12], and detailed description of the architecture given following units.

2.2 Overview of the proposed multi-object tracking originality

Basically, in Computer Vision there are a number of image processing studies that related to controlling systems to handle mess situation of multiple objects movement while doing visually monitoring. The proposing method is differs from other learning approach with feature learning strategies of datasets, that as an input it takes labeled image sequences for learning objects inside every frames of the video sequence. At the same time the original images sequences will be combined to have a clear object features and locations as well in results. As we know labeled images is the process managed by human-powered task of annotating an image with labels, such as classes, objects locations on image, and so on. These labels are predetermined by dataset creators (KITTI-tracking and MOT16) are chosen to give it related computer vision model information about what is shown in the image. Our proposal is to applying labeled image datasets into our network by integrating image sequence to get more accurate outcome model for feature tracking process. In our network model labeled images are applied as an input, which we are going to learn object locations and features by taking coordinates of the located objects configuration and regression, as well as image dimensions configurations into *dense layer* of the network that often follows LSTM layers and is used for outputting a prediction is called Dense, which used instead of fully-connected layer. Actually, it's the layer where each neurons of the layer N is connected to all of the neurons N+1 from the next layer. It implements the operation $output = X * W + b$ where X is input to the layer, and W and b are actual weights and bias of the layer. Next step will be combination of all coordinates, dimension of configuration and regression in differently concatenation layer in case of combining overall results into one cell memory unit. There are several approaches has been introduced by applying LSTM network for multi object tracking process, such as MOT with neural gating using bilinear LSTM[14] based on intuitions drawn from recursive least squares, bilinear LSTM stores building blocks of a linear predictor in its memory, which is then coupled with the input in a multiplicative manner, instead of the additive coupling in conventional LSTM approaches. Moreover, a multi-object tracking in videos based on LSTM and Deep Reinforcement learning[15] has been introduced by applying object detection and Markov decision process for sequence decisions to utilizing into deep learning reinforcement. Our model differs from this given recent work with learning strategy, used data structures, tracking proposals as well. In addition, conventional method of LSTM network can be applied different type of tasks and applications to achieve a various results, which is the network cannot be useful convention-

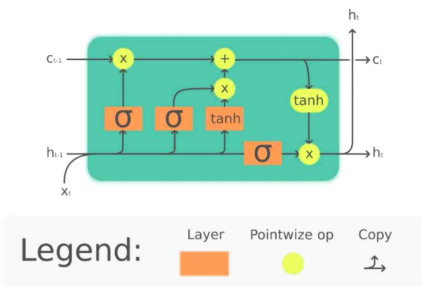

Fig. 2. The Long Short-Term Memory (LSTM) cell can process data sequentially and keep its hidden state through time.

ally, without integration of any tracking algorithm. (fully revised)

2.3 LSTM network baseline

The core concept of the LSTM network looks similar in a controlling flow chart as a recurrent neural network that the data passing process goes on information as it propagates forward. Theoretically the cell states can carry or keeping temporarily relevant details across all the processing of the sequences. From the name of method, even detailed particulars from initial time steps can make its way to later time steps, reducing the effects of short-term memory.

As we mentioned before, the LSTM cells has gates, while the cell state goes on its way, features gets inserted or erased it to the cell state via gates. The LSTM network gates perform different task while training, also they decides which information is allowed on the cell state, additionally they can learn which features are necessary to keep or forget it during training process. Network operations has its tasks to do in every steps, inside of the LSTM cell there are operations: sigmoid, tanh, pointwise multiplication, pointwise addition, and vector concatenation. These operations controls learned feature information of the network while predicting and tracking process as well.

The mathematical equations of the LSTM for the forward pass of cell state unit with a forget gate are given below, which describes the calculation of the operations mathematically [1]:

$$f_t = \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \quad (1)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (2)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (3)$$

$$\hat{c}_t = \sigma_g(W_c x_t + U_c h_{t-1} + b_c) \quad (4)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \hat{c}_t \quad (5)$$

$$h_t = o_t \circ \sigma_h(c_t) \quad (6)$$

where the initial values are $c_0 = 0$ and $h_0 = 0$ and the operator $\circ$ denotes the Hadamard product (element-wise product). The subscript t indexes the time steps.

The description of the variables:

- $x_t \in \mathbb{R}^d$: input vector to the LSTM unit
- $f_t \in \mathbb{R}^h$: forget gate's activation vector
- $i_t \in \mathbb{R}^h$: input/update gate's activation vector
- $o_t \in \mathbb{R}^h$: output gate's activation vector
- $h_t \in \mathbb{R}^h$: hidden state vector also known as output vector of the LSTM unit
- $\hat{c}_t \in \mathbb{R}^h$: cell input activation vector
- $c_t \in \mathbb{R}^h$: cell state vector
- $W \in \mathbb{R}^{h \times d}, U \in \mathbb{R}^{h \times h}$ and $b \in \mathbb{R}^h$: weight matrices and bias vector parameters which need to be learned during training

where the superscripts $c_0 = 0$ and $h_0 = 0$ refer to the number of input features and number of hidden units, respectively.

Activation functions:

- σ_g: sigmoid function.
- σ_c: hyperbolic tangent function.
- σ_h: hyperbolic tangent function or as the peephole LSTM paper suggests $\sigma_h(x) = x$

A. Sigmoid activation operation. The sigmoid function has a great role for learning details and controlling by cell state unit gates. Generally, a sigmoid function performs the same as the tanh activation, but a sigmoid activation takes between 0 and 1 values for squishing, instead of values between -1 and 1. That form may help updating or forgetting process of the cell state unit's gates. Because any taken value number multiplies 0 the outcome also becomes 0, that means causing values gives results to disappearing or forgotten. If its opposite, when any number multiplied by 1 is the same value stays the same or kept. Hence, the network decides which data is will be taken or forgotten while learning.

B. Forget gate operation. Firstly, the forget gate is essential unit of LSTM network that de-

cides what information should be thrown away or kept. Initially, the information from the previous hidden state h_{t-1} and from the current input x_1 is passed through the sigmoid function σ_g. As an output the values comes out between 0 and 1, if that values closer to 0 is going to forget f_t (1), else 1 means to keep it as a features.

C. Input gate operation. For updating the cell state input gate will be used, previously, the hidden state and current input together passes through a sigmoid activation σ_g, and the same input goes to tanh function at the same time that pointwise would be multiplied as well as the equation (2) will decide which input values will be updated by passing values between 0 and 1, if value is 0 means not important, or 1 means important. When the input passes through tanh function, it will be squished values between -1 and 1, in order to regulate the network.

D. Cell state operation. Now all passed forget and input gate information goes to calculation of the cell state, initially, the cell state gets pointwise c_{t-1} multiplied by the forget activation f_t vector (4), that stage has a possibility of dropping values from the cell state if it gets multiplied by values neat to 0. After taking output from input gate will be pointwise addition which updates the cell state to new values that the neural network finds relevant to keep for prediction (5), that gives us our new cell state.

E. Output gates operation. In final stage, the output gate determines which next hidden state h_t should be (6), that, the hidden state contains information on pervious inputs, also used for predictions. Firstly, the pervious hidden state and the current input goes through into a sigmoid activation (3), secondly, a newly modified cell state will be passed from a tanh function. Than two output value would be multiplied to decide what information the hidden state should carry, that output is the hidden state. The new cell state and the new hidden is then carried over to next time step learning process.

The rest time steps learning and prediction process follows as described above, where the forget gate decides that what learned information is relevant to keep from initial time steps. Moreover, for updating information the input gate will decide to add necessary learned features from current step. Lastly, the output gate determines the next hidden state for next time steps. The LSTM's control flow are a few tensor operations and loop section, where we can apply the hidden states for predictions, and combination of these all mechanisms forms the LSTM network, from that combination LSTM will choose an information which is relevant for learning and prediction.

2.4 Proposed LSTM Network based training

Currently, using RNN based methods becoming more common in a several research fields of the Computer Vision, such as image/video/audio processing, pattern recognition, biological vision, artificial intelligence, augmented reality, 2D sensors, photography and so on. The extensive use of RNN is due to their working ability of memory unit's direct control states, such as gate state, gate memory and forget gates that works in a certain period of time and without interruptions. In this section we are going to describe our LSTM based network model for tracking purposes. In this work the main idea is to create a network that to learn information directly from featured datasets which contains labeled images, ground truth info, sequence info, detection coordination of the trackable objects details for training procedure. The approach of training operation goes offline for learning features, estimating objects movement orientation to get proper output for integrating with tracking association. LSTM network has its own capability to exploit object characteristics to overcome classification problems, and other typical challenges, such as occlusion, objects similarity, multiple interconnected objects, and so on. Basically, proposing architecture

of the LSTM network build on multiple network layers that to utilize learning model to be more productive to exploit features of the dataset images for getting better performance with associated tracking method. To testify our approach we took datasets that commonly used with different methodology. They are KITTI tracking [6] and MOT16 [7], these datasets are totally different taken inside and outside environment content videos that includes single and multiple object scenes of human and vehicles.

A. Proposed model. The proposing LSTM based network model shown in Fig. 3 below that has been trained with two various type of datasets individually, in order to use trained output results with tracking association parallelly. Our offline trained output results gives us initial estimation of prediction and to identify the object class for tracking procedure. The training process starts by configuring dataset details with training options that could integrate the whole process correctly. KITTI-tracking training data set contains 21 video sequences that labeled into two class images. All images has different scenery of vision and viewpoints.

The LSTM network model takes the values of the sequence images to learn directly from labeled images for deeply excavate the features of the images by taking only the region of the interest for deeper exploration of the appearance similarity from the sequential images. At the same time, the network capable of to estimate the position of the trackable objects with the help of the input parameters of the dataset, such as labeled images, region of the interest (position of the multiple objects in the frame), and so on. Fig. 3 above illustrates the purposed LSTM network structure that shows the detailed steps of the learning model. Here we can see the network is taking the configuration of the object located coordinates and values of the regression to learn the features, predicting trackable object classes as well. Moreover, the network can save a predicted object location by the rules of the LSTM cell state units. The necessary steps of the predicting unit and memorizing follows in every frame of the video sequence by learning features parallelly. After receiving a trained model output in offline mode, we are going to use it as an initial feeding example to start the tracking process, also will be used for evaluation of our model.

B. Building an LSTM network. In order to build our network we started by integrating feature extraction part of the network. Hence, we are going to create a simple feature extractor in case of providing the network with bounding box information to learn the dataset properly, also producing the

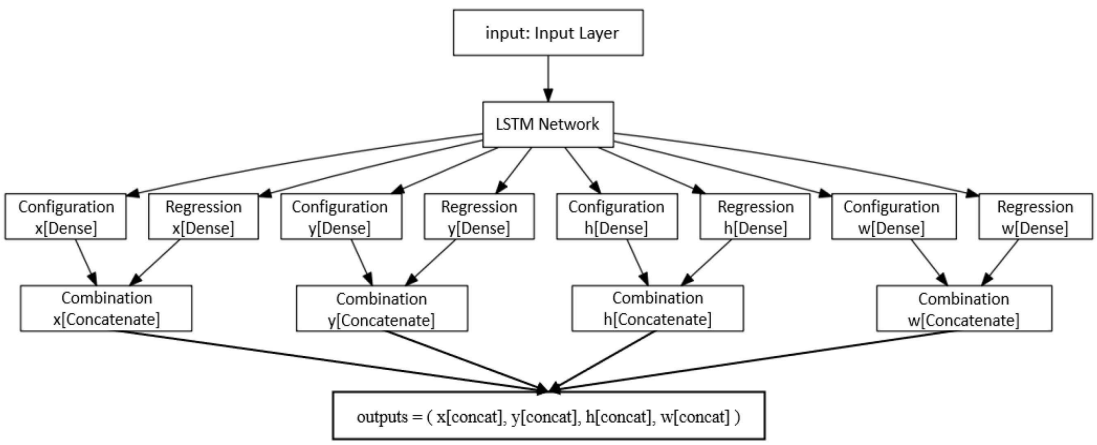

Fig. 3. Proposing LSTM Network model structure.

network with object position map that contains location of the object, region of interest and so on. The extracting features of the images and bounding boxes helps to learn the searching area of the interests and makes it smaller, in addition, it maintains the network from overbalance feature details. The network created on Keras model that includes libraries and all useful modules, such as layers, configurations and so on. The prediction section of the training model gets information from trained model features, bounding box learning section and ext. The loss function of training mode indicates the result of a bad prediction where a number indicating how bad the model's prediction was on a single example. In our model we used Huber loss function that changes the distance point from a quadratic to linear, also applied SoftMax cross entropy to identify a number of label classes to calculate losses. Optimization problems considered as a analyzing action to solve a problem for some sort of data and reduce the set of model selection and validation while learning process. In this work the adaptive moment estimation (Adam) optimizer has been used to minimize the cost function in training neural network. Other metrics such as, prediction, labels, and accuracy measures has been calculated by taking input details from learned features and labels information.

C. Training configuration. The evaluation of training process relays on the calculation of prediction metrics, losses, labels, accuracy of prediction and labels, precision and ext. Training process goes offline mode, where all required options and input datasets will be set to necessary part of the assistant. We have trained the model 100 epoch times to get trained model output, graph output, and evaluation files as well. We tested our model by applying with two common dataset individually, where a training process hold on individually and received outputs in a different locations. The learning rate has been set as a 0.001 manually in case of exploring the object features deeply, when the learning rate is small the training will take longer time than as usual. Training and estimation process goes parallelly into one action while training and shows the results of evaluation every 1000 iteration, which includes training evaluation precision, accuracy, loss, and other outcomes. In addition, the results of training outcome files will be stored into provided location. The estimated results of the training process shows the model productivity, learning ability, and mutual compatibility of model.

2.5 LSTM tracking association

The Multi-Object tracking systems are a little bit more complicated mechanisms rather than observing a single object in the video sequence. The LSTM network model has been associated with tracking process by integrating tracking facilities to the network with the help of prediction ability of LSTM cell state units, that by extracting features and providing with object located bounding boxes to learn the object labels, as well as by comparing it with network predicted bounding boxes, the tracking association has been provided. Tracking multiple targets the same time is quite challenging task, in order to identify them one by one and solving other detection based problems parallelly. In case of tracking a single target track the state space that location of the bounding box and their height and width has been applied. As a starting point for taking a shape of the objects and classes by taking covariance matrix of the initial state distribution is more important step for giving them an unique track identity. Additionally, identifying every object as one identical among multiple objects in a one frame, when we going to learn and remember a number of time step tracks by network is more time demanding task. Multi-target tracking differs visibly from a single target tracking process, which we should take into consideration that we are going to apply our trained LSTM model. We have an LSTM graph file that created

while training our model, which we are going to use it for predicting the next target by running graph runner unit. There are a list of active tracks at the current time step with number of object classes in one time step sequence to be tracked. In order to obtain the next prediction from each track, it returns an array of the shape with number of tracks and number of classes as well. The following Fig. 4 illustrates the structure of the tracking association with the LSTM model:

Updating the performance of the tracks goes by taking an array of detections and predictions at the time step of the target shapes, additionally this full updating process will be repeated. The next step is drawing the bounding boxes for each tracks on given images, if the track state matches the bounding box will be drown with tracking ID, otherwise the process will be returned to the back stage.

3. EXPERIMENT RESULTS AND DISCUSSION

We have been tested our network model with two different opensource datasets in case of exploring the proposed network capability, adaptability and working stability with different data. The training process is conducted in offline mode, by applying KITTI-tracking and MOT16's training set around 13K images which means: in the KITTI-tracking training set has 20 video sequences, around 8K images; in the MOT16 training set 7 video sequences around 5.3K images within has been trained. These datasets have in different position multi and single target located video sequences that helps to test the network learning and prediction facilities by training and testing it with test set of the datasets. In this experiment our method showed better results comparing to other MOT researches, however, we have some minority drawbacks like time consuming and unstable accuracy issues while training, training losses, and not much high accuracy results from other methods. This evaluation below can show only results of multi object tracking performance of our proposals and other techniques. As we mentioned in pervious sections, there are not much LSTM based multi object tracking papers available, even in existing ones also used different type of opensource datasets, structures as well. Therefore, we have faced some shortcoming comparison difficulties to evaluate our model with other conventional methods. However, we tried to be accurate and discrete in every comparison measurements evaluation.(added)

A. Training evaluation. Evaluation of the training is like a systematic process that to analyze a training program and initiatives whether if it's effective and efficient, and can work with different class datasets. We have evaluated our training

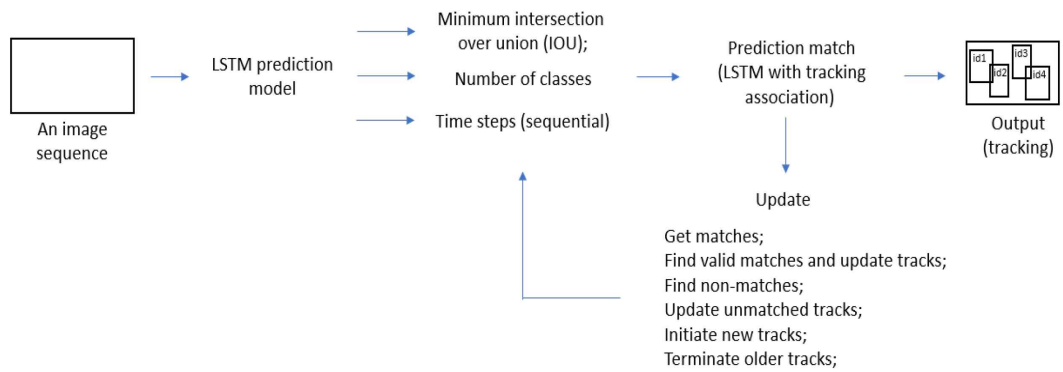

Fig. 4. Tracking association integration with the LSTM model.

Table 1. Training evaluation

Number of Iteration	Train Accuracy(%)	Loss(%)	Precision Evaluation(%)	Mean IoU Evaluation(%)	Accuracy Evaluation(%)
1(1257)	41.1	23.8	51.2	60.4	38.9
2(29318)	53.4	10.3	67.7	67.6	59.12
3(78200)	71.1	2.1	85.3	71.6	88.29
4(133478)	67.8	1.09	76.8	70.3	60.18
5(198321)	82.6	0.05	90.4	82.9	78.37
6(242412)	76.3	1	96.7	86.1	70.23

process by calculating the network capability, that, includes training accuracy, loss, precision accuracy, mean of interconnected over union (Mean IoU), and accuracy of evaluation. A given Table 1 below shows the training evaluation results:

A given results on this table above shows a randomly taken iteration steps outcome while training the network. The last column is the final iteration number with training results. In this table the train accuracy column is the calculation of training quality which means how our proposed LSTM network performed well did not overfit. Loss column of the table shows that how much percentage data has not been learned or skipped by our network model while training process. A precision evaluation unit of the table indicates the accuracy of the LSTM network prediction. Mean Interconnected over Union section part shows that how the network performed while learning the interconnected objects in every image. The last column of the training evaluation of the table describes the overall performance of the training process.

B. Tracking evaluation. The following unit of the experiment results describes a tracking association performance with the LSTM network model. In case of creating multi object tracking model we have integrated a tracking association with LSTM network model. The entire system can work an online mode. LSTM network based multi object tracking model has been tested with different type of open source datasets, in order to evaluate the model performance in a different situation. Following Table 2 shows multi object tracking evaluation on KITTI-tracking dataset:

In this Table 2 above shown the results of LSTM network based multi object tracking on KITTI tracking dataset. We took the video sequence ID's randomly here from KITTI tracking dataset, that every video sequence has different number of frames as well as in a different scenario. Moreover, in this table given a result of multi object tracking accuracy (MOTA), multi object tracking precision (MOTP), mostly tracked (MT) objects ratio and mostly loss (ML), and so on. Number of matched

Table 2. LSTM network based multi object tracking evaluation on KITTI tracking dataset

Video Sequence ID	Number of frames	MOTA (%) ↑	MOTP (%) ↑	MT ↑	ML ↓	IDs	Number of matched ↑	Number of objects
0000	153	76.4	79.1	15	20	2	623	708
0004	313	81.2	84.03	41	98	192	900	1108
0008	389	84.8	86.77	28	119	192	1157	1365
0013	339	91.8	92	68	23	101	1353	1474
0016	208	97.4	96	28	12	72	3042	3122
0019	1058	96.8	97.12	106	85	251	8538	8820

column describes the same targets has been tracked on the frame, which means one single target or similar target has been tracked several times.

C. Comparison results. In order to explore the performance of the proposed tracking method we have compared the tracking results with other method's outcomes. In this comparison result given below tables 3 and 4 describes two different open-source dataset testing results with different MOT methodologies, not only with LSTM based MOT models, because of a few resources in LSTM based MOT study. Nonetheless, tried to put RNN based works to compare and evaluate our model properly in multi object tracking procedure.(added) The given Table 3 below indicates a comparison results of multi object tracking on KITTI tracking dataset:

A given Table 3 above indicates a comparison result of tracking methods on KITTI dataset, with in new column feature - tracking mode, from this given result we can say that our proposed LSTM network based multi object tracking model has achieved best result among given recent multi object tracking methods, in most basic performance quality features (MOTA, MOTP, MT).

The following itinerary information below describes a comparison results of multi object tracking methods on MOT16 dataset:

The table 4 shows the results of multi object tracking comparing with most recent works on MOT16 dataset. The numerical results on this table shows the average percentage of performance on MOT16 dataset with different multi object tracking approaches. Percentage of accuracy rate on this table in the first place with 69.12, but in precision column our method shows second results. Although, LSTMNET_TA model performance of MOT accuracy and precision was higher than RNN_LSTM based MOT. Moreover, our model presented good results in different measurements. (added)

The main goal of this work is to implement a new multi object tracking method that should be capable to track multiple targets at the same time with IDs. Our multi target tracking method showed

Table 3. Comparison of our method performance on KITTI-tracking dataset with recent works

Methods	Tracking mode	MOTA ↑ (%)	MOTP ↑ (%)	MT ↑ (%)	ML ↓ (%)	IDs
MDP	online	76.59	82.10	52.15	13.38	130
LP-SSVM	online	77.63	77.80	62.61	8.76	62
NOMT	online	78.15	79.46	57.23	13.23	31
MCMOT-CPD	online	78.90	82.13	52.31	11.69	228
JCSTD	online	80.57	81.81	56.77	7.38	61
LSTMNET_TA(ours)	online	88.06	89.22	59.12	10.42	135

Table 4. Comparison of our method performance on MOT16 dataset with recent works

Methods	Tracking mode	MOTA ↑ (%)	MOTP ↑ (%)	MT ↑ (%)	ML ↓ (%)	IDs
RNN LSTM	online	19.0	71.0	5.5	45.6	1490
MDP	online	30.3	71.3	13.0	38.4	680
JPDA	online	23.8	68.2	5.0	58.1	365
Response	online	62.0	73.6	37.7	20.7	909
DPT_DPT	online	61.3	79.1	32.1	18.6	739
LSTMNET_TA(ours)	online	69.12	75.17	45.26	17.18	810

Fig. 5. A qualitative results of the proposed LSTM network based on MOT with two class of objects, tested on KITTI-tracking dataset, along with in these frames every target has their identical ID' s.

that the LSTM network based tracking association can work and adopt easily with new targets. A visual qualitative results of KITTI-tracking dataset tested results of proposed multi object tracking method given below in Fig. 5.

4. Conclusion and feature work discussion

In conclusion, the method in this research showed that LSTM network based target tracking has more capability and adaptation rather than other approaches. This method has more advantages and distinguishes from other type of the neural network, such as memory cells, controlling gates, working ability with in time series video sequences, adopting easily with any kind of situation, and so on. In addition, the network has a prediction skills and memorizing it into cells in a time series period. In this study we have presented our network model that extracts features and learns as an object. As we told in pervious steps of our description above, LSMT is one of the RNN based structure that can be directed to any kind of purposes. So, there are not much works available on MOT field with LSTM model. Even so, there is not much outcome to compare with, we tried make a specific comparison with RNN_LSTM based MOT, where our model showed better results. (added) The entire process can work in online mode, integration of tracking association supports tracking multi targets easily by taking detected and predicted bounding boxes from LSTM network unit. Comparison studies demonstrated that our presented network model along with tracking association multi object tracker performance is much better than other neural network based techniques. Moreover, our system is not complicated to set in a real project application.

This presented multi target tracking system can be applied any kind of required filed of tracking purposes, because of its adaptability and also can be managed easily. In the future work, we are going to improve the network configuration and compatibility with other additional properties, additionally, accuracy and performance as well.

REFERENCE

[1] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, and P.H.S. Torr, "Fully-convolu-

tional Siamese Networks for Object Tracking," *Proceeding of the European Conference on Computer Vision*, pp. 850-865, 2016.

[2] A. Milan, S.H. Rezatofighi, A. Dick, I. Reid, and K. Schindler "Online Multi-Target Tracking Using Recurrent Neural Networks," *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence(AAAI-17)*, pp. 4225-4232, 2017.

[3] S. Sun, N. Akhtar, H.S. Song, A. Mian, and M. Shah, "Deep Affinity Network for Multiple Object Tracking," *Journal of Latex Class Files*, Vol. 13, No. 9, pp. 1-15, 2017.

[4] Kh. Farkhodov, K. Oh-Heum, M. Kwang-Seok, K. Oh-Jun, L. Suk-Hwan, and K. Ki-Ryong "A New CSR-DCF Tracking Algorithm based on Faster RCNN Detection Model and CSRT Tracker for Drone Date", *Journal of Korea Multimedia Society,* Vol. 22, Iss. 12, pp. 1415-1429, 2019.

[5] A. Jadhav, P. Mukherjee, V. Kaushik, and B. Lall, "Aerial Multi-object Tracking by Detection Using Deep Association Networks," *Proceeding of 2020 National Conference on Communications (NCC)*, pp. 1-6, 2020.

[6] A. Milan, L. Leal-Taix´e, I.D. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark for Multi-object Tracking," *arXiv preprint arXiv: 1603.00831*, 2016.

[7] A. Geiger, P. Lenz, and R. Urtasun, "Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite," *Proceeding of Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 3354-3361, 2012.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 6, pp. 1137-1149, 2017.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-time Object Detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788, 2016.

[10] A.A. Butt and R.T. Collins, "Multi-target Tracking by Lagrangian Relaxation to Min-cost Network Flow," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1846-1853, 2013.

[11] S.H. Bae and K.J. Yoon, "Confidence-based Data Association and Discriminative Deep Appearance Learning for Robust Online Multi-object Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 3, pp. 595-610, 2018.

[12] Wikipedia, https://en.wikipedia.org/wiki/Long_short-term_memory

[13] B.J. Radford, M. Leonardo, Apolonio, A.J. Trias, and J.A. Simpson, "Network Traffic Anomaly Detection Using Recurrent Neural Networks", *arXiv Preprint arXiv:1803*.10769 v1, 2018.

[14] C.H. Kim, F. Li, and M.R. James, "Multi-object Tracking with Neural Gating Using Bilinear LSTM" *Proceeding of European Conference on Computer Vision*, pp. 200-215, 2018.

[15] J.M. Xin, D. Chao, P.Z. Geng, W.L. Fang, and S. Xing, "Multi-object Tracking in Videos Based on LSTM and Deep Reinforcement Learning" *Complexity,* Vol. 2018, pp. 1-12, 2018.

Farkhodov Khurshedjon

He received the B.S. degree Computer Engineering Tashkent University of Information Technologies, Uzbekistan in 2017. He is currently a Master student in the department of IT Convertgence and Application Engineering in Pukyong National University. His research interests include Digital Image Processing and Machine Learning.

Kwang-Seok Moon

received the B.S., M.S., and Ph. D degrees in Electronics Engineering in Kyungpook National University, Korea in 1979, 1981, and 1989 respectively. He is currently a professor in department of Electronic Engineering at Pukyong National University. His research interests include digital image processing, video watermarking, and multimedia communication.

Suk-Hwan Lee

He received a B.S., an M.S., and a Ph.D. degree in Electrical Engineering from Kyungpook National University, Korea in 1999, 2001, and 2004 respectively. He is currently working as a Professor in the Department of Computer Engineering at Dong-A University. His research interests include multimedia security, digital image processing, and computer graphics.

Ki-Ryong Kwon

He received the B.S., M.S., and Ph.D. degrees in electronics engineering from Kyungpook National University in 1986, 1990, and 1994 respectively. He worked at Hyundai Motor Company from 1986-1988 and at the Pusan University of Foreign Language from 1996-2006. He is currently a professor in the Department of IT Convergence and Application Engineering at the Pukyong National University. He has researched the University of Minnesota in the USA in 2000-2002 with Post-Doc. and Colorado State University in 2011-2012 with visiting professors. He was the President of Korea Multimedia Society in 2015-2016. His research interests are in the area of digital image processing, multimedia security and watermarking, bioinformatics, weather radar information processing, and machine learning.