1

Journal of Computer Science Research

Volume 1 | Issue 1 | April 2019 | ISSN 2630-5151 (Online)

BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Editor-in-Chief

Editorial Board Members

Volume 1 Issue 1 · April 2019 · ISSN 2630-5151

Journal of Computer Science Research

BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Volume 1 | Issue 1 | April 2019 | Page 1 - 35
Journal of Computer Science Research

Contents

ARTICLE

REVIEW

BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
http://ojs.bilpublishing.com/index.php/jcsr

REVIEW

High-Resolution Traffic Flow Prediction Model Based on Deep Learning

Zhihong Yao[1,2,3]* Yibing Wang[4]

1. School of Transportation and Logistics, Southwest Jiaotong University

2. National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Southwest Jiaotong University, Chengdu, Sichuan, 610031, China

3. TOPS Laboratory, Department of Civil and Environmental Engineering, University of Wisconsin-Madison, 1415 Engineering Drive, Madison, WI 53706, USA

4. Department of Architecture Engineering, Yantai Vocational College, Yantai, Shandong, 264670, China

ARTICLE INFO

Article history:
Received: 8 December 2018
Accepted: 27 December 2018
Published: 31 December 2018

Keywords:
Traffic flow prediction
Deep learning
Time resolution
Platoon dispersion
Signal timing optimization
Real time

ABSTRACT

The time resolution of the existing traffic flow prediction model is too big to be applied to adaptive signal timing optimization. Based on the view of the platoon dispersion model, the relationship between vehicle arrival at the downstream intersection and vehicle departure from the upstream intersection was analyzed. Then, a high-resolution traffic flow prediction model based on deep learning was developed. The departure flow rate from the upstream and the arrival flow rate at the downstream intersection was taking as the input and output in the proposed model, respectively. Finally, the parameters of the proposed model were trained by the field data, and the proposed model was implemented to forecast the arrival flow rate of the downstream intersection. Results show that the proposed model can better capture the fluctuant traffic flow and reduced MAE, MRE, and RMSE by 9.53%, 39.92%, and 3.56%, respectively, compared with traditional models and algorithms, such as Robertson's model and artificial neural network. Therefore, the proposed model can be applied for real-time adaptive signal timing optimization.

1. Introduction

As one of the most important components of adaptive control systems (e.g., TRANSYT[1] and SCOOT[2], the platoon dispersion model is also the basis of traffic flow prediction, simulation, and signal timing optimization. The first platoon dispersion model based on the hypothesis that the vehicle's velocity follows normal distribution was proposed by Pacey[3] in 1965. From the view of traffic flow, Robertson[1] proposed a platoon dispersion model supposing that travel times obey a

**Corresponding Author:*
Zhihong Yao,
School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, 610031*, China,*
E-mail: Arry.zhyao@gmail.com

 DOI: https://doi.org/10.30564/jcsr.v1i1.381

shifted geometric distribution. Since the shifted geometric distribution has the merit of simpleness and convenient calculation, Robertson's platoon dispersion model has been incorporated in a large amount of software or systems, including TRANSYT-7F[1], SCOOT[2], SATURN[4], and TRAFLO[5]. Most of the later studies[6,7] are based on the assumption that the travel speed or travel time follows a certain statistical distribution. These models are accurate or not depend on the assumption. However, traffic flow may operate in unstable states which caused by the signal intersection. Moreover, in real-time transportation systems, access to information constraints will lead to the distribution of restricted traffic data different from that of theoretical traffic data. Under these complex situations, the platoon dispersion models may be unable to capture the dispersion of traffic flow, and may become inapplicable for real-time transportation systems. To some extent, this shortcoming limits the potential application of some real-time traffic signal control systems included with the platoon dispersion models. With the development of big data technology, many methods[8-15] were proposed in big data environment. However, the forecasting time resolution of these models is too big, such as 5, 10, 30, and even 60 minutes. Therefore, these models have good predicted effect, but the big time resolution is not enough to be applied in adaptive control systems.

In summary, a proposed traffic flow prediction model must be able to accurately capture the change of traffic flow and must satisfy the optimal time resolution of the signal timing optimization.

In fact, apart from the classic platoon dispersion models, some intelligent methods (e.g., neural network[16,17], support vector machine[18], Kalman filter[19,20], etc.) are also applied to predict traffic flow. However, these methods only consider the time series characters of the traffic flow at downstream intersection, and do not consider the correlation between the arrival and departure rate of the downstream and upstream intersection.

From the perspective of platoon dispersion model, there is a certain relationship between the arrival and departure flow of the downstream and upstream intersection. Considering the deep learning can describe any complex stochastic and non-linear systems, a high-resolution traffic prediction model based on deep learning is proposed. Then, the real-time data of the upstream intersection is used to forecast the arrival flow of the downstream intersection based on the proposed model. In addition, the prediction time resolution of the proposed model can be determined by the actual demand. In this study, we choose 5 seconds as the prediction time resolution, which fully meets the minimum requirements of the adaptive control algorithms or systems.

This paper is organized as follows. Section 2 reviews the classical Robertson's platoon dispersion model. A high-resolution traffic flow prediction model based on deep learning is proposed in Section 3. In Section 4, the predicted results of the proposed model and traditional models and algorithms are compared. Section 5 closes the paper with conclusions and further research.

2. The Platoon Dispersion Model

2.1 Notation of Platoon Dispersion

Owing to the existence of the urban signal intersections, the continuous traffic flow is forced to split into separate platoons. Meanwhile, due to the differences of drivers' driving behavior and safety awareness, these factors result in the discrepancy of vehicle's speed. Consequently, the platoon becomes longer as vehicles travel further downstream, and this phenomenon is commonly called "platoon dispersion". This phenomenon is illustrated in Figure 1. When the signal light is red, vehicles have to queue up at the stop line. While the signal light turns green, vehicles through the intersection with saturation flow, but it is not the saturation flow rate when vehicles reach the downstream section. Therefore, the flow rate is decreasing with time as the platoon reaches downstream intersection. In addition, the peak of the platoon will become smoother and smoother when the distance between the upstream and downstream section is getting longer and longer.

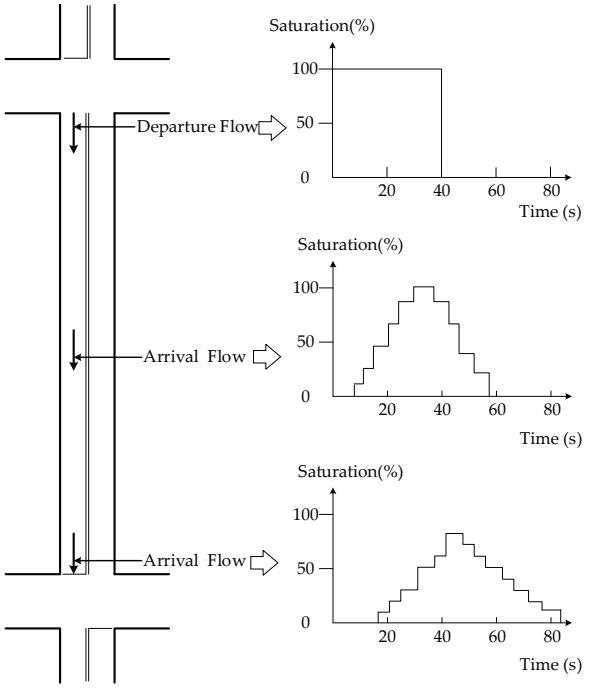


Figure 1. Diagram of platoon dispersion

DOI: https://doi.org/10.30564/jcsr.v1i1.381

2.2 The Robertson's Platoon Dispersion Model

The platoon dispersion model aims to describe the relationship between the departure flow of the upstream intersection and the arrival flow of the downstream intersection, and realize real-time prediction of arrival flow at downstream sections or intersections. The classical platoon dispersion has been given by Robertson[1], who used observed data to derive an iterative method to capture the behavior of platoon. The traffic flow prediction model was used for optimization of traffic signals to obtain the minimum vehicle delay. For each time interval, the arrival flow rate of the downstream stop-line is calculated by Eq.(1).

$$q_d(t) = \sum_{i=1}^{t-T_{\min}} q_u(i) \cdot F \cdot (1-F)^{t-T_{\min}-i}, \quad (1)$$

$$q_d(t) = F \cdot q_u\left(t - T_{\min}\right) + (1-F) \cdot q_d(t-1), \quad (2)$$

where $q_d(t)$ represents the arrival flow rate at time interval t of the downstream intersection. $q_u(i)$ represents the departure flow rate at the time interval i of the upstream intersection. T_{min} represents the minimum travel time for the road segment, the value is equal to 0.8 times the mean travelling-time. F represents a smoothing factor.

In Eqs.(1-2), the departure flow of the upstream intersection can be obtained by loop detectors. Therefore, the arrival flow rate of the downstream intersection can be calculated by Eqs.(1-2) with estimated parameters. The minimum travel time T_{min} can be calculated according to the historical data, and the smoothing factor can be calculated by the Eq.(3).

$$F = \frac{1}{1 + \alpha\beta T}, \quad (3)$$

where α is the platoon dispersion coefficient, which has been found to be 0.5 in central business district (CBD), β is the travel time factor, generally taken the value of 0.8. Readers could refer to TRANSYT-7F user's guide[21] for more details. T is the travel time for the road segment.

In Robertson's platoon dispersion model, the basic assumption is that the travel time follows a shifted geometric distribution. However, the shifted geometric distribution has a long tail and hence the Robertson's model predicts a greater dispersion of the platoon than the actual situation. Later, the actual data fitting proves that the travel time or speed follows various probability distributions, such as normal distribution, lognormal distribution, mixture Gaussian and truncated distribution of these distributions[6,7,21-27], etc. When the traffic condition changes, the travel time or speed distribution will also change greatly. However, the parameters of the Robertson's model are static which cannot reflect the real-time traffic flow characteristics.

3. Deep Learning-based High-resolution Traffic Flow Prediction Model

3.1 Deep Learning

Deep learning describes a high dimensional function via a sequence of semi-affine non-linear transformations[28]. The deep learning architecture[29]. is organized as a graph shown in Figure 2.

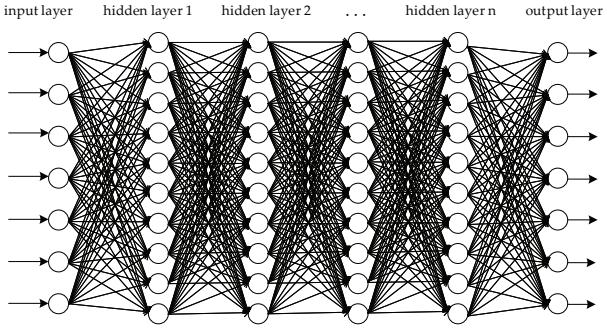


Figure 2. Basic structure of deep learning

A deep learning predictor, denoted by $\hat{y}(x)$,, takes an input vector $x = (x_1, \ldots x_p)$ and outputs via different layers of abstraction that employ hierarchical predictors by composing L non-linear semi-affine transformations. Specifically, a deep learning architecture is as follows. Let $f_1, \ldots, f_n$ be given univariate activation link functions, e.g. sigmoid $1/(1+e^{-x})$,$\cosh x$,$\tanh x$, Heaviside gate functions ($I(x > 0)$), or rectified linear units ($\max\{x,0\}$) or indicator functions (I (x∈R)) for trees. The composite map is defined by

$$\hat{y}(x) = G(x) = \left(f_{w_n,b_n} \circ \cdots \circ f_{w_1,b_1}\right)(x), \quad (4)$$

where $f_{w,b}$ is a semi-activation rule defined by

$$f_{w_l,b_l}(x) = f\left(\sum_{j=1}^{N_l} w_{lj} x_j + b_l\right) = f(w_l^T x_l + b_l), \quad (5)$$

where N_l represents the number of units at layer l. The weights $w_l \in R^{N_l \times N_{l-1}}$ and offset b∈R needs to be learned from training data.

Data dimension reduction of a high dimensional map G is performed via the composition of univariate semi-affine functions. Let z^l denote the l-th layer hidden features, with $x = z^0$. The final output is the response y, can be numeric or categorical. The explicit structure of a deep prediction rule is than

 DOI: https://doi.org/10.30564/jcsr.v1i1.381

$$
\begin{aligned}
z^1 &= f(w_0^T x + b_0) \\
z^2 &= f(w_1^T z^1 + b_1) \\
&\vdots \\
z^n &= f(w_{n-1}^T z^{n-1} + b_{n-1}) \\
y(x) &= w_n^T z^n + b_n
\end{aligned} \tag{6}
$$

In many cases there is an underlying probabilistic model, denoted by $p(y|\hat{y}(x))$. This leads to a training problem given by optimization problem

$$\min_{w,b} \frac{1}{T}\sum_{i=1}^{T} -\log p\left(y_i \middle| \hat{y}_{w,b}(x_i)\right), \tag{7}$$

where $p(y|\hat{y}(x))$ is the probability density function given by specification $y_i = G(x_i) + \epsilon_i$. Efficient algorithms[31] exist to solve those problems, even for high dimensional cases.

In summary, when the structure of the deep neural network, the training data samples and the error threshold are given, the deep neural network can be optimized through training by the optimization algorithm. Therefore, the actual process of the application of deep neural network is divided into four steps: designing the deep neural network structure, obtaining the data samples, training the network, and using the trained network to predict the value based on the new input. So, how to build a traffic flow prediction model based on deep learning will be discussed in next section.

3.2 High-Resolution Traffic Flow Prediction Model

In Figure 1, there is a length of Δx of road segment which between the upstream and downstream intersection. The minimum and maximum travel time of the road segment are calculated according to the speed limit of road segment or historical data. T_{min} and T_{max} are the minimum and maximum travel time, respectively.

The departure flow of the upstream intersection can be acquired in real time by detectors which are set at exit lane of the upstream intersection. According to the idea of the platoon dispersion model: the number of vehicles arriving at the downstream section which come from the vehicles of time interval ($[t - T_{\max}, t - T_{\min}]$) at upstream intersection. So, this relationship can be expressed by the follow formula.

$$q_d(t) = G\left(q_u(t - T_{\max}), q_u(t - T_{\max} + 1), \cdots, q_u(t - T_{\min})\right), \tag{8}$$

where $G(\cdot)$ is a mapping relation.

According to the Eq.(8), there is a correlation between the arrival flow rate of the downstream intersection at the time interval t, and which maybe come from the upstream intersection for each time interval of time period $[t - T_{\max}, t - T_{\min}]$. This relationship cannot be expressed in a general form of function. In the classical platoon dispersion model, there is a basic assumption that vehicle's speed or travel time follows a certain probability distribution. Then, the relationship formula between the downstream and the upstream flow rate is derived. Therefore, these models need to select an appropriate probability distribution to describe the characteristics of traffic flow, and the probability distribution can only characterize the dynamic traffic flow to a certain extent. Considering the deep neural network can be used to describe any linearly separable complex stochastic systems, and does not require any underlying assumptions. The deep neural network is used to describe the mapping relation between the arrival traffic flow of the downstream intersection and the departure traffic flow of the upstream intersection. Then, a high-resolution traffic prediction model based on the deep learning is developed.

Firstly, we need to determine the structure of the deep neural network from the Eq.(8). The number of input neuron is $T_{\max}-T_{\min}+1$,, input variables are $x=(q_u(t-T\max), q_u(t-T_{\max}+1), \cdots, q_u(t-T\min))$; The number of output neuron is 1, output variable is $y=q_d(t)$. The number of hidden layers and hidden layer neurons can be selected according to comparative analysis. Among them, the number of input nodes of deep neural network with different time resolution is also different, which is equal to the length of time interval $[t - T_{\max}, t - T_{\min}]$ divided by the selected time resolution.

After the deep neural network structure is determined, the network needs to be trained through the historical data. If we want to predict the traffic flow rate at the th time interval, the historical data before the th time interval can be used to train the network. Considering the network training needs a certain period of time, so we train or optimize the network every once in a while, such as 5 minutes, to ensure that the deep neural network training can be completed in the interval. Finally, the trained network can be used to forecast the traffic flow rate of the downstream intersection based on the departure flow of the upstream intersection which can be acquired in real time by detectors. Moreover, the historical arrival flow rate of the downstream intersection can also be obtained by detectors, which can be used to train the deep neural network.

3.3 High-Resolution Traffic Flow Prediction Algorithm

In this study, a high-resolution traffic flow prediction algorithm can be divided into the following 5 steps, which is shown in Figure 3.

Step 1. The minimum and travel time $T_{\min}$, $T_{\max}$, network

  DOI: https://doi.org/10.30564/jcsr.v1i1.381

training update time Δt and the current time are determined according to the actual situation;

Step 2. The deep neural network structure is determined according to the relevant parameters in Step 1;

Step 3. The deep neural network is trained and optimized by using the historical data of the current time t;

Step 4. The traffic flow of the downstream intersection is predicted based on the trained network and the real-time data obtained by the detectors at the upstream section, then $t=t+1$;

Step 5. If the current time meets the network training update time, then skip to Step 3; otherwise, go to Step 4.

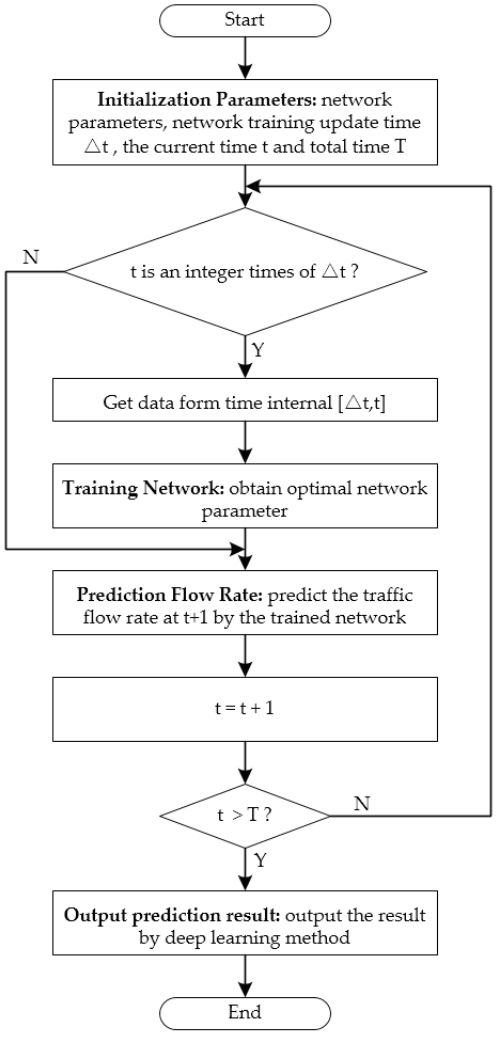


Figure 3. Flowchart of the proposed algorithm

Through the above 5 steps, the traffic flow of the downstream intersection can be predicted in real time.

4. Case Study

In this section, the predicted performance of Robertson's model, artificial neural network, and the proposed model will be discussed based on the survey data.

4.1 Data Collection

In order to prove the proposed model have better performance, Wushan Road in Guangzhou is selected for field investigation, as shown in Figure 4. There are 14 bus lines via this segment, and in general the traffic condition is unsaturated. Specifically, the survey time interval is 7:30 am – 11:20 am and the traffic flow is volatile, forming a distribution with a typical morning peak. The travel times can be obtained by comparing vehicle license plates in the upstream and downstream section (the distance between two places is 650 m). After data preprocessing, we get 1,621 pieces of effective data as shown in Table 1. Afterward, the estimated value of all parameters can be calculated by a statistical method. Then, the flow rate of the upstream and downstream section are obtained in Figure 4.

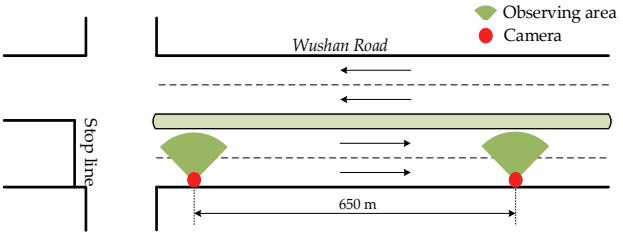
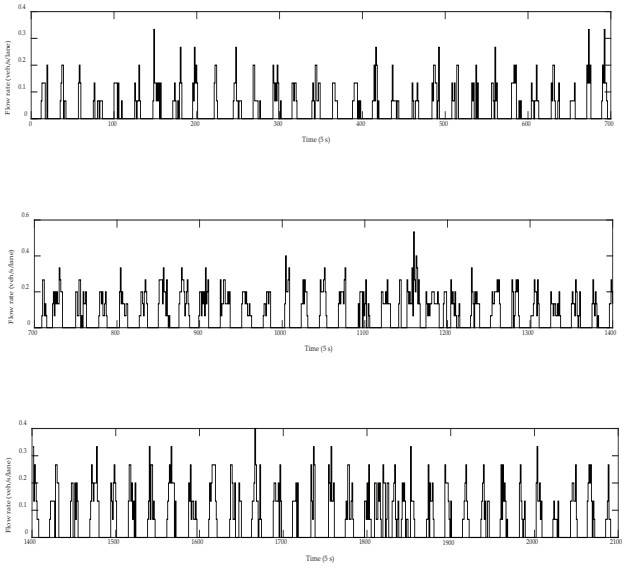
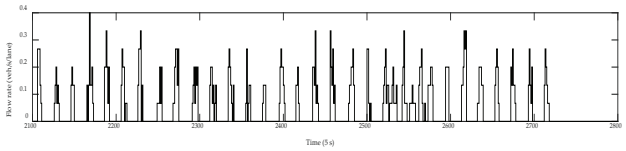


Figure 4. Diagram of the survey road segment.

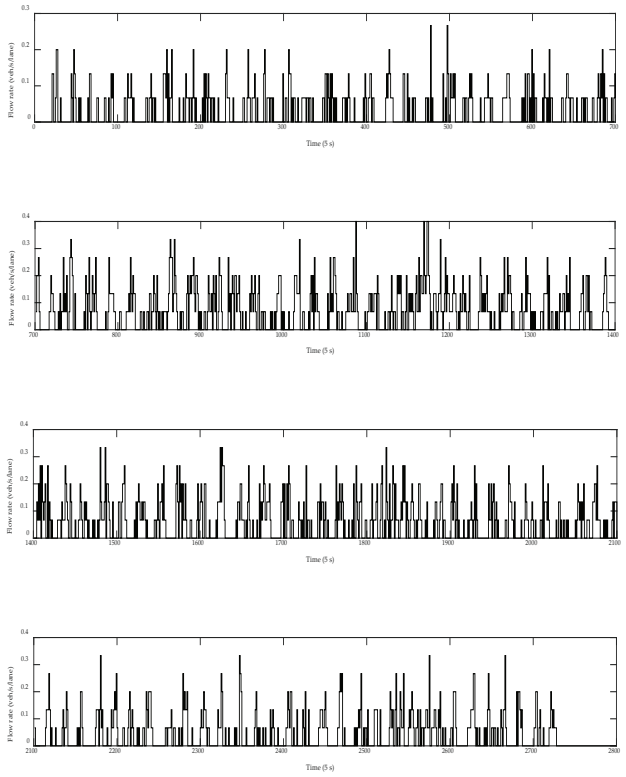
Table 1. The statistical parameters

Statistical Parameter	Value
The number of vehicles (vehicle)	1621
The minimum travel time (second)	30
The maximum travel time (second)	130
The average travel time (second)	41.6





(a) The departure flow rate of the upstream section.



(b) The arrival flow rate of the downstream section.

Figure 5. The flow rate of departs and arrivals during time intervals of 5 s

As demonstrated in Figure 5, the fluctuation of traffic flow at the upstream and downstream section is large, and the regularity is not very significant. The forecast effect of these three models will be discussed based on the survey data as following.

4.2 Model Evaluation

Firstly, with 5 seconds as the time interval, the data are aggregated, and we obtain 2,720 data samples. Then, these data can be divided into two parts, the first part serves as the parameter calibration (the first 2,000 data samples), the second part serves as the model prediction effect validation (the last 720 data samples). Therefore, the parameters of Robertson's model can be calculated by the method in the literature[32] or the TRANSYT-7F manual[21]. The parameters of the artificial neural network and deep learning are calibrated by training based on the first part of data samples[10,12]. The key parameter values of these three models are shown in Table 2.

Table 2. Key parameter values for different models

Model	Parameter	Value
Robertson's model	T	41.6 s
	α	0.5
	β	0.8
Artificial Neural Network	Input neurons	10
	Hidden neurons	5
	Hidden layer number	1
	Output neurons	1
	Transfer function	S-function
	Training algorithm	Levenberg-Marquardt algorithm
	Training epoch	500
	Learning rate	0.05
Deep Learning	T_{min}	30 s
	T_{max}	130 s
	Input neurons	20
	Hidden neurons	[10,10,10,10,10]
	Hidden layer number	5
	Output neurons	1
	Transfer function	S-function
	Training algorithm	Levenberg-Marquardt algorithm
	Training epoch	1000
	Learning rate	0.05

Because the time interval is 5 seconds, and the difference between the maximum and the minimum travel time is 100 seconds, so the number of input neurons is 20. The number of hidden layers is 3 and hidden neurons are 10 calculated by the empirical formula[32] and the number of output neurons is 1. After the parameters of these three models are determined, we can predict the traffic flow rate of the downstream section by using Robertson's model, artificial neural network, and deep learning, respectively. The predicted results of these tgree models are shown in Figure 6. The performance of these three models can be assessed quantitatively by examining the prediction error statistics. Standard prediction measures include MAE, MRE, and RSME[33,34]. These measures for the predicted results shown in Figure 6 are given in Table 3.

DOI: https://doi.org/10.30564/jcsr.v1i1.381

Table 3. The Evaluation Index Value of Models

Measure	Robertson's Model	Artificial Neural Network	Deep Learning	Improvement
MAE	0.0576	0.0519	0.0494	14.24% / 4.82% / 9.53%
MRE	17.69%	23.51%	12.13%	31.43% / 48.41% / 39.92%
RMSE	0.0728	0.0732	0.0704	3.30% / 3.83%/ / 3.56%

Note: the value of improvement: improvement compare Robertson's model / improvement compare artificial neural network / average improvement compare these two models.

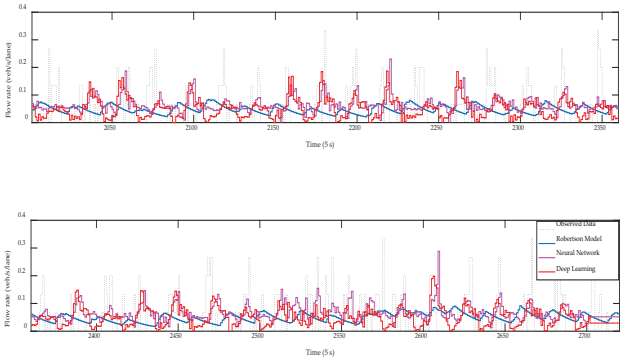


Figure 6. The actual and predicted arrival flow rate based on two models during time intervals of 5 seconds.

As shown in Figure 6, compared with the other two models, the deep learning prediction results can better capture the fluctuant characteristics of traffic flow. Because to a certain degree, the deep learning can get the relationship between the arrival and departure flow rate of the downstream section and the upstream intersection by training. On the contrary, Robertson's model is based on the strict assumption that travel times follow a shifted geometric distribution, and cannot accurately characterize the flow relationship between the upstream and downstream intersection. In addition, the artificial neural network only considers the time series characters of the traffic flow, and does not consider the correlation between the arrival and departure rate of the downstream and upstream intersection. The analysis shows that the performance of Robertson's model and artificial neural network do not work well when the traffic flow fluctuation is frequent. However, the deep learning can adapt to the fluctuation of traffic flow through continuous learning, so it has better prediction effect.

Moreover, the error analysis results in Table 3 show that the prediction errors of the deep learning are less than the Robertson's model and artificial neural network. The MAE, MRE, and RMSE of the deep learning is average reduced by 9.53% 39.93, and 3.56%, respectively, compared with Robertson's model and artificial neural network. Therefore, deep learning can be used for real-time traffic flow prediction, and the prediction time resolution can be accurate to 5 seconds. The results can be applied to the optimization of adaptive signal timing.

5. Conclusion and Future Work

5.1 Conclusion

In this paper, a high-resolution traffic flow prediction model is proposed based on the perspective of the platoon dispersion model. The proposed model uses deep learning to describe the relationship between the arrival flow of the downstream intersection and the departure flow of the upstream intersection, and to realize the prediction of traffic flow at the downstream intersection. The results of the field data validation show that the proposed model is better than the Robertson's model and artificial neural network, and the time resolution is 5 seconds, which meets the basic needs of adaptive signal timing optimization algorithm. So, the proposed model can be used for adaptive signal timing optimization.

5.2 Future Work

Future work will be considered to study more other traffic information (e.g., vehicle' s speed and acceleration) to improve the proposed model prediction capability. And the calculation method of intersection stopping times and queuing length based on the proposed traffic flow prediction model should be studied in future research. In addition, the stability and robustness of the proposed model will be discussed by more data set in the future.

Acknowledgement: This work support by Cultivation Program for the Excellent Doctoral Dissertation of Southwest Jiaotong University(D-YB201708). The authors express our acknowledge to Lüou Shen, Weitiao Wu etc. for their hard work during data collection and processing.

References

[1] Robertson, D. I. TRANSYT: A Traffic Network Study Tool [R]. Transport and Road Research Laboratory Report LR 253. Transport and Road Research Laboratory, London, U.K., 1969.

[2] Hunt, P., D. Robertson, R. Bretherton, and R. Winton. SCOOT-A Traffic Responsive Method of Coordinating Signals [R]. Transport and Road Research Laboratory Report LR 1041. Transport and Road Research Laboratory,

London, U.K., 1981.

[3] Pacey, G. The Progress of a Bunch of Vehicles Released from a Traffic Signal [R]. Road Research Laboratory Note RN/2665/GMP, Transport and Road Research Laboratory, London, U.K., 1956.

[4] Hall, M., and L. WILLUMSEN. SATURN-A Simulation-Assignment Model for the Evaluation of Traffic Management Schemes [J]. Traffic Engineering & Control, 1980, 21(4), 168–176.

[5] Lieberman, E. B., and B. Andrews. Traflo: A New Tool to Evaluate Transportation System Management Strategies [J]. Transportation Research Record: Journal of the Transportation Research Board, No. 772, Transportation Research Board of the National Academies, Washington, D.C., 1980, 9–15.

[6] Shen, L., Liu, R., Yao, Z., Wu, W. and Yang, H. Development of Dynamic Platoon Dispersion Models for Predictive Traffic Signal Control[J]. IEEE Transactions on Intelligent Transportation Systems.2018, 99, 2018, 1–10.

[7] Jiang, Y., Yao, Z., Luo, X., Wu, W., Ding, X. and Khattak, A. Heterogeneous Platoon Flow Dispersion Model Based on Truncated Mixed Simplified Phase-type Distribution of Travel Speed[J]. Journal of Advanced Transportation, 2016, 50(8), 2160-2173.

[8] Lv, Y., Duan, Y., Kang, W., Li, Z. and Wang, F.Y., 2015. Traffic Flow Prediction with Big Data: A Deep Learning Approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(2), 865-873.

[9] Abadi, A., Rajabioun, T. and Ioannou, P.A. Traffic Flow Prediction for Road Transportation Networks With Limited Traffic Data[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(2), 653-662.

[10] Polson, N.G. and Sokolov, V.O. Deep Learning for Short-term Traffic Flow Prediction[J]. Transportation Research Part C: Emerging Technologies, 2017,9, 1-17.

[11] Kumar, S.V. and Vanajakshi, L., Short-term Traffic Flow Prediction using Seasonal ARIMA Model with Limited Input Data[J]. European Transport Research Review, 2015, 7(3), 21.

[12] Kumar, K., Parida, M. and Katiyar, V.K. Short Term Traffic Flow Prediction in Heterogeneous Condition using Artificial Neural Network[J]. Transport, 2015, 30(4), 397-405.

[13] Chen, D. Research on Traffic Flow Prediction in the Big Data Environment Based on the Improved RBF Neural Network[J]. IEEE Transactions on Industrial Informatics, 2017,13(4), 2000-2008.

[14] Kim, Y.J. and Hong, J.S. Urban Traffic Flow Prediction System using a Multifactor Pattern Recognition Model[J]. IEEE Transactions on Intelligent Transportation Systems, 2015,16(5), 2744-2755.

[15] Li, Y., Jiang, X., Zhu, H., He, X., Peeta, S., Zheng, T. and Li, Y., 2016. Multiple measures-based chaotic time series for traffic flow prediction based on Bayesian theory. Nonlinear Dynamics, 2016, 85(1),179-194.

[16] Lint, J. W. C. V., S. P. Hoogendoorn, and H. J. V. Zuylen. Accurate Freeway Travel Time Prediction with State-Space Neural Networks under Missing Data [J]. Transportation Research Part C: Emerging Technologies, 2005, 13(5–6), 347–369.

[17] Hodge, V. J., R. Krishnan, J. Austin, J. Polak, and T. Jackson. Short-Term Prediction of Traffic Flow Using a Binary Neural Network [J]. Neural Computing and Applications, 2014, 25, (25), 1639–1655.

[18] Yang, Y., and H. Lu. Short-Term Traffic Flow Combined Forecasting Model Based on SVM [C]. In International Conference on Computational and Information Sciences, Chengdu, China, 2010.

[19] Xie, Y., Y. Zhang, and Z. Ye. Short-Term Traffic Volume Forecasting Using Kalman Filter with Discrete Wavelet Decomposition [J]. Computer-Aided Civil and Infrastructure Engineering, 2007, 22(5), 326–334.

[20] Ojeda, L. L., A. Y. Kibangou, and C. C. De Wit. Adaptive Kalman Filtering for Multi-Step Ahead Traffic Flow Prediction [C]. In American Control Conference, Washington, USA, 2013.

[21] Courage, K., and C. E. Wallace. TRANSYT-7F Users Guide [M]. Office of Traffic Operations and Intelligent Vehicle/Highway Systems, U.S. Department of Transportation, Dec. 1991

[22] Wei, M., W. Jin, L. Shen. A Platoon Dispersion Model Based on a Truncated Normal Distribution of Speed [J]. Journal of Applied Mathematics, 2012, 2012(1), 155–172.

[23] Jiang, Y. S., Z. H. Yao, X. Ding, and X. L. Luo. Mixed Platoon Flow Dispersion Model Based on Truncated Mixed Phase Distribution of Speed [C]. CD-ROM. Transportation Research Board of the National Academies, Washington, D.C., 2016.

[24] Yao, Z. H., Y. S. Jiang, Y. X. Wu, and Y. Q. Liu. Platoon Dispersion Model Based on Mixed Phase Distribution of Speed [J]. Journal of Transportation Systems Engineering and Information Technology, V2016, 16(3), 133–140.

[25] Yao, Z. H., L. O. Shen, W. W. Wu, Y. S. Jiang, and L. Huang. Heterogeneous Traffic Flow Platoon Dispersion Model Based on Travel Time Distribution [J]. China Journal of Highway and Transport, 2016, 29(8), 134-142,151.

[26] Yao, Z. H., P. Han, B. Zhao, Y. S. Jiang, B. Liu, and M. Q. Du. High-Granularity Dynamic Traffic Flow Prediction Model Based on Artificial Neural Network [C]. CD-ROM. Transportation Research Board of the National Academies, Washington, D.C., 2017.

[27] Jiang, Y. S., Z. H. Yao, X. L. Luo, W. T. Wu, X. Ding, and A. Khattak. Heterogeneous Platoon Flow Dispersion

  DOI: https://doi.org/10.30564/jcsr.v1i1.381

Model Based on Truncated Mixed Simplified Phase-type Distribution of Travel Speed [J]. Journal of Advanced Transportation, 2016, 50, 2160-2173.

[28] Haykin, S. Neural Networks: A Comprehensive Foundation [M]. Prentice Hall, Ontario, 2004.

[29] Diaconis, P., Shahshahani, M. On Nonlinear Functions of Linear Combinations [J]. SIAM Journal on Scientific and Statistical Computing, 1984, 5(1), 175-191.

[30] Kim, S. J., Koh, K., Lustig, M., et al. An Interior-Point Method for Large-Scale l1-Regularized Least Squares [J]. IEEE Journal of selected topics in signal processing, 2007, 1, (4,), 606-617.

[31] Friedman, J. H., Tukey, J. W. A Projection Pursuit Algorithm for Exploratory Data Analysis [J]. IEEE Transactions on Computers, 1974, 23(9), 881-890.

[32] Yu, L. Calibration of Platoon Dispersion Parameters on the Basis of Link Travel Time Statistic [J]. Transportation Research Record: Journal of the Transportation Research Board, No. 1727, Transportation Research Board of the National Academies, Washington, D.C., 2000, pp. 89–94.

[33] Makridakis, S. G., and S. C. Wheelwright. Forecasting: Methods and Applications [M]. John Wiley and Sons, New York, 2008.

[34] Larry, H. K. Event-Based Short-Term Traffic Flow Prediction Model [J]. Transportation Research Record: Journal of the Transportation Research Board, No. 1510, Transportation Research Board of the National Academies, Washington, D.C., 1995, 125–143.

 DOI: https://doi.org/10.30564/jcsr.v1i1.381

BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
http://ojs.bilpublishing.com/index.php/jcsr

REVIEW

Logistic Regression Based Model for Improving the Accuracy and Time Complexity of ROI's Extraction in Real Time Traffic Signs Recognition System

Fareed Qararyah[1]* Yousef-Awwad Daraghmi[2] Eman Yasser Daraghmi[3]

1. Department of Computer Engineering, Koç University, Turkey

2. Computer Systems Engineering Department , College of engineering and Technology, Palestine Technical University-Kaddorie, Palestine

3. Applied Computing Department , College of Applied Science, Palestine Technical University-Kaddorie, Palestine

ARTICLE INFO

Article history:
Received: 2 January 2018
Accepted: 2 January 2018
Published: 4 January 2018

Keywords:
Traffic sign recognition systems
Logistic regression

ABSTRACT

Designing accurate and time-efficient real-time traffic sign recognition systems is a crucial part of developing the intelligent vehicle which is the main agent in the intelligent transportation system. Traffic sign recognition systems consist of an initial detection phase where images and colors are segmented and fed to the recognition phase. The most challenging process in such systems in terms of time consumption is the detection phase. The trade off in previous studies, which proposed different methods for detecting traffic signs, is between accuracy and computation time. Therefore, this paper presents a novel accurate and time-efficient color segmentation approach based on logistic regression. We used RGB color space as the domain to extract the features of our hypothesis; this has boosted the speed of our approach since no color conversion is needed. Our trained segmentation classifier was tested on 1000 traffic sign images taken in different lighting conditions. The results show that our approach segmented 974 of these images correctly and in a time less than one-fifth of the time needed by any other robust segmentation method.

1. Introduction

Road signs are used to regulate traffic, warn drivers, and provide useful information to help make driving safe and convenient [1]. However, driver's ability to recognize them depends on the physical and mental conditions. These conditions can be affected by many factors such as fatigue, and observatory skills [2]. As a result, traffic signs recognition systems were proposed to augment driver's attention so that driving can become safer and more convenient.

The core functionality of traffic signs recognition systems takes place in two phases: the first is traffic sign

**Corresponding Author:*
Fareed Qararyah
Department of Computer Engineering, Koç University, Turkey
Email: fqararyah@gmail.com

DOI: https://doi.org/10.30564/jcsr.v1i1.442

detection (TSD) and the second is the traffic sign recognition (TSR) [3]. In the TSD phase which is the input of the recognition phase, the image is preprocessed, enhanced, and segmented according to the sign properties such as color or shape [4]. About 62% of work done in this field used colors as the basic cue for TSD while the remaining used shape [2,5]. The aim of segmentation process, which turns the captured image into binary image based on a certain thresholding algorithm, is to extract the regions of interest from the whole image [6-9]. The regions of interest are those containing colors that qualify them to contain traffic signs. Due to this role, this phase is of a critical importance because it will narrow the search space that has to be targeted by the system by marking certain areas to be sought after for traffic signs instead of the whole image, so that the work of the next phase will be much efficient. This phase is also challenging since colors information has a considerable sensitivity to lighting conditions [2,10]. If ideal lighting conditions are presented, the task of segmentation will be easy. However, non-ideal illumination conditions are the predominant ones, some of them are the result of excessive daylight, and some are the result of poor lighting either at night or in bad weather conditions. Moreover, regions of interest extraction may be very costly in terms of the computation time depending on the image size [3]. Therefore, in this paper we focus on the TSD phase and we propose an approach for improving the accuracy of detection and reducing its computational time.

Previous studies proposed different fundamental approaches to deal with TSD accuracy and time complexity. For example, to overcome the issue of sensitivity, most of the previous work followed the approach of color space conversion, where the captured images which are originally represented in the RGB color space are transformed into another space like HSV(hue, saturation, value)/ HSL(hue, saturation, lightness) or IHLS(Improved Hue Luminance Saturation) [1,2,4,6,9]. These spaces are used because chromatic information can be easily separated from the lighting information making this approach more suitable for detecting a specified color in almost all light conditions. This approach produced relatively accurate results but in a high computational time.

To avoid high computational complexity, some research used RGB color space with minimal computations [7,11]. However, this approach suffers from low accuracy. Another avoidance mechanism depended on statistical distribution findings which conclude that most traffic signs appear in the middle of the image and a few in the top with relatively large scales [12]. Such assumption reduces the robustness of the system and limits the detection of some real traffic signs which are not subject to the supposed distribution due to some poses of the vehicle like turning around or driving downhill. Another limitation is that the proposed system requires the vanishing horizon to be roughly in the middle of the captured image, which is not always the case [12]. Another approach was proposed to identify color space by using Maximally Stable Extremal Regions (MSERs) [3,13]. This approach resulted in a very precise detection. However, it is not sufficient alone to produce results that could be used as inputs to the recognition phase because it obtains traffic signs with a large number of backgrounds [3]. To enhance MSERs work, some research made refinement for the produced ROI's in the recognition phase like what was proposed in [3]. In [13], MSERs was used along with a complementary shape-based extractors which would eventually consume a considerable amount of time.

In this paper we propose a novel TSD model that has high performance and simultaneously balances between the accuracy and computational complexity. We use logistic regression, a simple but powerful machine learning technique for the segmentation. Logistic regression is the appropriate when the dependent variable is binary. It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables [14]. To evaluate our model, we used a total of 1000 images of local traffic signs and other images downloaded from 'Belgian Traffic Sign Dataset' [10]. Our model outperformed other related methods in terms of accuracy and computation time.

2. Related Work

Previous studies can be divided into two main groups depending on the color spaces used in segmenting the images. The first group contains studies that adopted the approach of color space conversion before thresholding, and the second contains those who have used the RGB color space as the domain for processing.

2.1 Converting RGB Image into Another Color Space

This approach is the most popular since its results are more robust and it covers wider cases under different lighting conditions. The two dominant color spaces used by the advocates of this approach are HSI(hue, saturation, intensity) [1,15,16] and HSV [4,6,9]. These color models are used extensively because they do a separation between the colorfulness and the lightness of the image so that there is independence between the chromatic and achromatic components. This makes extracting a good threshold an easier task. However, converting from the RGB to those color

 DOI: https://doi.org/10.30564/jcsr.v1i1.442

spaces consumes long time since it involves a non-linear transformation [17].

2.2 Using RGB for Segmentation

The aim of working directly with RGB is to design much faster segmentation algorithms. The idea of using the RGB in segmentation is based on an observation that the differences between the color component of the sign and the other two colors components remains relatively high and could easily be used with an appropriate (not too sensitive) threshold for segmentation [18]. Hence, a simple segmentation algorithm can be implemented using the three differences: ΔRG, ΔRB and ΔGB which need not be selected very precisely [11].

However, another research based on this algorithm have concluded that using constant ΔGR, ΔGB and ΔRB for different non-ideal illumination conditions throughout the whole day and night is not appropriate. They developed the idea of using different adaptive color threshold values for different non-ideal illumination conditions at different times in the day and night. They adapted the threshold values based on the intensity or brightness of the time. The intensity value that was used as a threshold, it was calculated using the following formula [5]:

$$I = \frac{R + G + B}{3} \quad (1)$$

However, even when using adaptive threshold, this algorithm has a problem that there are no set of threshold values that yield a very good segmentation. The range of the threshold values may be narrow so that some signs are going to be lost, or wide so that many parts of the image will be segmented as regions of interest. We tested this approach on a local traffic sign image as shown in figure 1 which shows an example of adaptive wide threshold. The figure shows that colors that have dominant red component for example orange and brown will be segmented as if they are red. That is because the differences between the red component and the other two components relative to the average of the three components are similar to what is found in some shades of the red color.

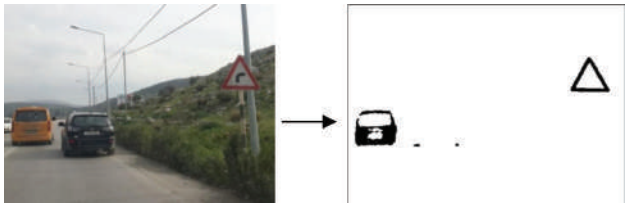


Figure 1. Wide range segmentation based on adaptive thresholding in which other objects with colors close to red are segmented as a red traffic sign

3. Logistic Regression based TSD Model

3.1 Theoretical Background

We investigated the road signs color segmentation problem as a machine learning problem. We used logistic regression classifier as a segmentation algorithm. Logistic regression is a classification algorithm used to derive a hypothesis $h_0(x)$ given training data represented as a features vector $x=\{x_1,x_2,x_3 \ldots, x_n\}$ and a target function $(\theta^T x)$, this is done by estimating a set of values for weights θ^T that achieves the best mapping between input and output values given in the training data. The derived hypothesis categorizes new observations under a discrete set of classes. It uses sigmoid function to map real values into probabilities between 0 and 1. The following set of equations describes how the sigmoid is used to do this mapping:

$$h_0(x) = S(\theta^T x) \quad (2)$$

$$Z = \theta^T x \quad (3)$$

$$S(z) = \frac{1}{1 + e^{-z}} \quad (4)$$

Where θ is the values of the model's weights and the bias, x is the features of training data used in the target function, and is the sigmoid function. The estimation of θ values is done by an ongoing minimization of the cost function (the deviation of the hypothesis prediction from the actual output) during the learning phase. The cost function is given by:

$$Cost\,(h_0(x), y) = -y\,log(h_0(x)) - (1-y)log(1-h_0(x)) \quad (5)$$

where y is the actual output value.

3.2 The Proposed Model

We derive our approach based on the results proposed by Benalla et. al.[11] that the differences ΔRG, ΔRB and ΔGB are indicators of the sign color. We also borrowed Sajjad et. al.[7] approach of searching for efficient adaptive threshold. However, instead of dealing with these differences as mutual exclusive conditions to decide the threshold, we have modeled them together as variables/features in a function where each of them contributes simultaneously to the decision-making process. This way the relative value of each difference to the other two decides the threshold, as such we have avoided the aforementioned issue shown in figure 1 of getting either a narrow or a wide threshold. Another feature we added to our function is the value of the desired color component of a certain sign (blue or red in our case). This is because we have noticed that the relations between the three differences are not adequate to judge the color in all cases, especially when the difference between two of them are approaching to zeros. Measuring their values relatively to the value of the desired color

DOI: https://doi.org/10.30564/jcsr.v1i1.442

was the solution to avoid failing in such cases. The values of these four features contributions in determining the threshold are the weights associated with these features. We have used logistic regression technique to estimate the best values of these weights. Figure 2 is simple a description of the model.

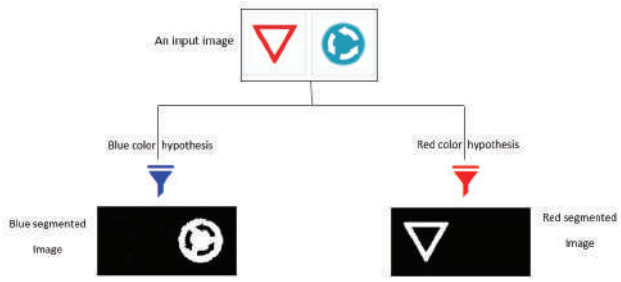


Figure 2. Abstract representation of segmentation process

As shown in the figure 2, an image goes under two segmentation thresholds that are executed in parallel and produce two binary images. Each binary image contains regions having pixels of needed color as ones and other pixels as zeros.

The Red threshold hypothesis is given by:

$$h_\theta(x) = S(\theta_0 + \theta_1 R + \theta_2 \Delta RG + \theta_3 \Delta RB + \theta_4 |\Delta GB|) \quad (6)$$

and Blue threshold hypothesis is given by:

$$h_\theta(x) = S(\theta_0 + \theta_1 B + \theta_3 \Delta RB + \theta_4 \Delta GB + \theta_2 |\Delta RG|) \quad (7)$$

where S is the sigmoid function, R is the red component of the pixel in RGB model in the range [0–255], G is the green component of the pixel [0–255], and B is the blue component of the pixel [0–255]. when thresholding for a color, the difference between the other two colors is considered as a magnitude, because its sign is not important, what is important is the ratios between its magnitude and the values of other features.

We classify the pixel as red/blue if the value of the hypothesis (the probability that this pixel is red/blue) is equal to or greater than 0.5 (see figure 3); in other words, if the value of the target function is greater than or equal to zero the pixel is classified as red/blue. Hence our classifier can be represented by:

$$P = \begin{cases} Red, \theta_0 + \theta_1 R + \theta_2 \Delta RG + \theta_3 \Delta RB + \theta_4 |\Delta GB| \geq 0 \\ Blue, \theta_0 + \theta_1 B + \theta_3 \Delta RB + \theta_4 \Delta GB + \theta_2 |\Delta RG| \geq 0 \end{cases} \quad (8)$$

where P is the pixels of the image.

The following is the pseudo code of the core function of our model, which receives a road image '*Image*' and the segmentation color '*Desired_Color*' as inputs and returns a segmented/binary image '*Image_2*' as an output:

INPUT

1: desired_color // the color of the segmentation, it takes the value of ***red*** *or* ***blue***

2: image // the image taken from real time capturing unit I.e. a camera

BEGIN

3: image = ConvertToSinglePrecision(Image)

if *desired_color = red* ***then***

c1 = image(,*,1) //Red color component of Image*

c2 = image(,*,2) //Green color component of Image*

c3 = image(,*,3) //Blue color component of Image*

else if *Desired_Color = blue* ***then***

c1 = image(,*,3) //Blue color component of Image*

c2 = image(,*,2) //Green color component of Image*

c3 = image(,*,1) //Red color component of Image*

end if

diff_1 = c1 – c2

diff_2 = c1 – c3

diff_3 = |c2 – c3|

image_2 = + c1 + Diff_1 + Diff_2 + Diff_2 + Diff_3 ≥ 0

return *Image_2*

END

3.3 Training and Testing the Model

The data set used for training the model includes 300 local images captured in different lighting conditions in Palestine. The images contain signs from different types, for example, signs giving warnings, signs giving orders and information signs as shown in figure 3. For testing the proposed model, we used another data set includes 1000 traffic sign images obtained online from 'Belgian Traffic Sign Dataset' found at http://btsd.ethz.ch/shareddata/.

Figure 3. examples of Traffic signs used in Palestine

To train our model, we have created a set of segmented images; in the training data each training entity consists of our four feature values and the corresponding output value at the pixel level. We used the 'Minimize a continuous differentiable multivariate function' proposed by Carl Edward Rasmussen in 2002 to derive the weights value.

To evaluate our model, we compared it with other related studies that have robust results [1-4,6,9,15-17]. We implemented the core functionality of their models which entails color spaces conversion using the same programming language we used in our implementation and under the same hardware specs; and the same dataset.

Our derived hypothesis is implemented using Matlab and we have thoroughly assessed the performance of the model using Matlab profiler. The model was implemented on a CPU with specifications: *Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz, 3 MB SmartCache, 2 Cores and 4 Threads*. We used vectorized instructions in all image-pixels manipulation operations to optimize the performance. A detailed discussion of the performance and accuracy is discussed in the following section.

4. Results and Discussion

4.1 Accuracy

The proposed model shows dependable results as of the 1000 test images 974 were segmented correctly. This accuracy of about 97.4% outperforms that of the previous models as shown in Table 1. We define the accuracy as the percentage of correct detections.

Table 1. Accuracy comparison of different models used for TSD

Model	Accuracy %
The proposed model	97.40
[18]	94.85
[3]	93.54
[6]	92.98
[12]	94.41
[9]	89.32

The results shown in the figure 4 depict samples of model accuracy under different lighting conditions. The segmented images allow perfect determination of ROI's which form the domain of search for the recognition phase.

The images that were not segmented correctly are only those signs whose colors are extremely deteriorated.

4.2 Computational Time:

The proposed model also reduces the time required for

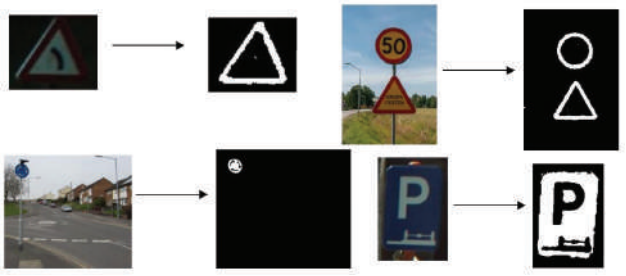


Figure 4. Our model test results samples

TSD. Previous studies that provided accurate results have based their work on converting from RGB color space to other color spaces. However, they did not provide a detailed performance measurement of their algorithms complexity. So, to give an indication of the performance comparison between our work and theirs, we compare the time taken by our model with that needed to merely converting an image from RGB to the required color spaces. Table 2 shows the time needed to detect a sign with different image sizes. For each image size, we took the average time needed for RGB To HSV in the studies [4,6,9] and for RGB To HIS in the studies [1,15,16].

Table 2. Computational time comparison for different models used for TSD

Image Size in pixels	Time in seconds		
	Proposed model	RGB To HSV	RGB To HSI
87 * 174	0.000195	0.002001	0.002042
800 * 556	0.008231	0.071822	0.035139
768 * 1024	0.017002	0.116413	0.071002
1536 * 2048	0.057206	0.815914	0.319685
2448 * 3264	0.141652	1.278549	0.572381

As shown in table 2, our model outperforms other related models with different image resolutions. As can be inferred from the table, our model is faster than other models that have a competing accuracy. And this is a crucial benefit, because computation time is a key component in every real time application, especially in intelligent transportation system where parts of a second can make a great difference. Table 2 shows that the time needed for converting from the RGB to other color spaces, before doing any thresholding, is more than five times of the time needed to do a total segmentation of the image in the RGB color model using an efficient algorithm.

5. Conclusion

This paper proposes a novel model of addressing a challenging part of traffic sign recognition systems that is traffic sign detection in which the sign image is segmented and the ROI's are determined. The proposed model uses logistic regression as an image thresholding algorithm. It

DOI: https://doi.org/10.30564/jcsr.v1i1.442

operates on the RGB color model and does not use any color space conversions. The proposed model provides a high level of accuracy of about 97.4%, with very time-efficient results. The application of our approach will help to design higher performance complete traffic signs recognition systems. Our future work will focus on developing the second phase which is a model for traffic sign recognition that utilizes the proposed TSD model.

References

[1] C.-Y. Fang, S.-W. Chen, and C.-S. Fuh, "Road-Sign Detection and Tracking," IEEE Trans. Veh. Technol., 2003, 52(5).

[2] H. Fleyeh, "Color detection and segmentation for road and traffic signs," in IEEE Conference on Cybernetics and Intelligent Systems, 2004, 2, 809–814.

[3] H. Luo, Y. Yang, B. Tong, F. Wu, and B. Fan, "Traffic Sign Recognition Using a Multi-Task Convolutional Neural Network," IEEE Trans. Intell. Transp. Syst., 2018, 19(4), 1100–1111.

[4] H. Fleyeh, "Road and Traffic Sign Color Detection and Segmentation - A Fuzzy Approach," in Conference on Machine VIsion Applications, 2005, 3–27.

[5] E. De Micheli, R. Prevete, G. Piccioli, and M. Campani, "Color cues for traffic scene analysis," in Proceedings of the Intelligent Vehicles' 95. Symposium, 466–471.

[6] N. Beg, M. Agrawal, R. Pasari, A. Singh, and K. H. Wanjale, "Traffic Sign Recognition System," Int. J. Res. Advent Technol., 2016.

[7] M. S. Hossain, M. M. Hasan, M. Ameer Ali, M. H. Kabir, and a B. M. Shawkat Ali, "Automatic detection and recognition of traffic signs," 2010 IEEE Conf. Robot. Autom. Mechatronics, 2010, 286–291.

[8] A. Broggi, P. Cerri, P. Medici, P. P. Porta, and G. Ghisio, "Real Time Road Signs Recognition," 2007 IEEE Intell. Veh. Symp., no. section III[1] A. Broggi, P. Cerri, P. Medici, P. P. Porta, and G. Ghisio, "Real Time Road Signs Recognition," 2007 IEEE Intell. Veh. Symp., section III, 2007, 981–986,981–986.

[9] A. Shustanov and P. Yakimov, "CNN Design for Real-Time Traffic Sign Recognition," Procedia Eng., 2017, 201, 718–725.

[10] R. Timofte, K. Zimmermann, and L. Van Gool, "Multi-view traffic sign detection, recognition, and 3D localisation," Mach. Vis. Appl.,2014, 25(3), 633–647.

[11] M. Benallal and J. Meunier, "Real-time color segmentation of road signs," in CCECE 2003 - Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (Cat. No.03CH37436), 2003, 3, 1823–1826.

[12] Y. Yuan, Z. Xiong, and Q. Wang, "An Incremental Framework for Video-Based Traffic Sign Detection, Tracking, and Recognition," IEEE Trans. Intell. Transp. Syst., 2017,18(7), 1918–1929.

[13] S. Salti, A. Petrelli, F. Tombari, N. Fioraio, and L. Di Stefano, "A traffic sign detection pipeline based on interest region extraction," Proc. Int. Jt. Conf. Neural Networks, 2013.

[14] S. Caballé and J. Conesa, Software Data Engineering for Network eLearning Environments, 2018, 11. Cham: Springer International Publishing.

[15] P. Arnoul, M. Viala, J. P. Guerin, and M. Mergy, "Traffic signs localisation for highways inventory from a video camera on board a moving collection van," in Proceedings of Conference on Intelligent Vehicles, pp. 141–146.

[16] L. Priese, R. Lakmann, and V. Rehrmann, "Ideogram identification in a realtime traffic sign recognition system," in Proceedings of the Intelligent Vehicles '95. Symposium, pp. 310–314.

[17] A. Mavrinac, J. Wu, X. Chen, and K. Tepe, "Competitive learning techniques for color image segmentation," Proc. Mach. Learn. Comput. Vis., 2007, 88, (590), 33–37.

[18] S. B. Wali, M. A. Hannan, S. Abdullah, A. Hussain, and S. A. Samad, "Shape Matching and Color Segmentation Based Traffic Sign Detection System," Prz. Elektrotechniczny, 2015,91(1), 36–40.

 DOI: https://doi.org/10.30564/jcsr.v1i1.442

BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
http://ojs.bilpublishing.com/index.php/jcsr

REVIEW

Requirement Gathering Problems: Environmental Issues in Robot Development

Muhammad Arif[1]* Zunaira Shafqat Ali[2] Samman Arooj[3]

1. School of computer Science and Technology, Guangzhou University, China
2. Department of Computer Science, University of Gujrat, Pakistan
3. Department of Software Engineering, University of Gujrat, Pakistan

ARTICLE INFO

Article history:
Received: 17 December 2018
Accepted: 7 January 2019
Published: 11 January 2019

Keywords:
Human-robot interaction
Environment
Operational
Autonomous

ABSTRACT

This paper deals with the importance of environment consideration in developed countries while collecting the requirement from customer to make robot that would address the question "Why robots should be made more precisely according to the environment needs? "In developed countries, robots are used in manufacturing work as well as in performing the hazardous tasks such as bomb-disposal. So, there is a need to pay attention towards making the robots that can fit perfectly to some extent in environment to be utilized more efficiently. A lot of money, effort and time is spent on making the robots .But what if such a worth costing robot fails to fit in the operational environment? The best way to solve this problem is proposed in this paper which is to make the environment as a part of Requirement gathering process carrying high importance in robot making process to make the robots more Operational and suitable for the working environment .Like the other main attributes in requirement gathering process such as user requirements, system requirements and external requirements, there should be an attribute "Environmental requirements" which will automatically put emphasis on the considering also the environment as a main subject to pay heed.

1. Introduction

An area of knowledge that gives you a chance to study and understand robots is called robotics. A robot is an artificially intelligent machine which can do different works by the guidance or on its own. Mainly, a robot is an electro mechanical machine that is handled and given instructions by computer or software installed in it. Two types of robots are used: 1) autonomous, the robots which are used to do research in human like systems as ASMIO and TOPIO and 2) autonomous, those which are designed specially to perform a specific task like Nano and Swarm robots and other helper robots which are used to make or move things. By movements robots can send the messages to very far off places too[1].

Requirement specification is a phase of SDLC in which requirements of the user are specified[2]. Depending on

**Corresponding Author:*
Muhammad Arif
Department of computer Science and Technology, Guangzhou University, China
Email: arifmuhammad36@hotmail.com

DOI: https://doi.org/10.30564/jcsr.v1i1.412

these requirements a robot is designed so that it can fulfil a particular task. Requirement specification is an important phase as almost 20% of work is done in it according to a rough survey. The robots which are the main topic in our case are autonomous robots which are designed specially to facilitate the user. These robots are designed for special purposes and hence they have special and their own requirements to be fulfilled[3]. So, the problem which is encountered mostly is that companies or people who are demanding these cannot specifically tell what they want the robot to do, when and how? They cannot specify that in which type of environment it is to be used. Mostly the environment is most meaningful requirement which is often ignored.

Environment is the factor that mostly affects human beings too. As people cannot survive in all type of environment no matter the same they are, robots usually cannot do that to. Some unwanted things and effects in the environment can cause them to be dead or not work properly.

The problem arose, when a company in Tokyo launched a robot to search the nuclear plates dumped under water. The work of the robot was to find these plates and bring them back so that the radioactive rays emitting from them cannot cause problem to the surroundings. The fact that it has to go under water was considered but they did not realize that it was to be used near radioactive rays. So when it was sent down it broke down 2 feet away from the nuclear plates. Now they are going to make a robot which will not be affected by radioactivity. But it cost almost 2 million dollars, which is a huge amount to compensate[4].

Now we have understood that environment is an important factor to be considered while working on any type of autonomous robots. This was an example to depict the importance of environment. As environment also contains our society so you should also consider how the society will take it, what will be their response towards them?

Japan being the most advanced country in the field of robotics encourages the use of robots in every field but the developing countries don't use this technology largely as there is no work of robots in these countries[5]. So, we should also consider for which society we are making it? How it will work there? Secondly, we should consider with whom these robots have to interact? It should not happen like a robot that was introduced on Tokyo airport as a clerk and it could speak only Japanese.

Our main focus regarding this research is to make the developers aware of the fact that environment is a non-compromising attribute in the development of autonomous robots. It should be considered on every step of SDLC life cycle. But most importantly requirements should be gathered by considering all the surrounding factors with which it has to interact.

2. Literature Review

Let us consider the facts that how all of us perform in the environment. When we are performing in a familiar environment, we have our attention focused on some information we know like what to do to an object, where we have to go, what to do etc. Our actions are then carried out subconsciously to satisfy our goals. When we go in a new environment or perform in the same environment new set of work, our attention increases. As we do not know that environment or the set of actions we are carrying out we are attending consciously to the continuous mechanism of direction in the way we walk, that we look and the way we are controlling our body. This is the same mechanism that applies to humanoid robots operations[6]. When operating a robot which is equipped with a high level of autonomy serving in a known environment like transportation VANETs [7-10], a small number of high level commands will be sufficient in achieving the intended tasks[11,12].

Many robots need an environment that has to be controlled for their working. Controlled environment generally refers to a specific are with specified set of objects. The Robots cannot just walk in and do all the work they are intended to do, instead they can only provide some specific functions on specific objects in a specified area. So for this condition they need a controlled environment. Our main concern are autonomous robots that have to work in the human environment and interact efficiently with all the people in that environment.

Speaking about human environments, there are challenging characteristics which are beyond the control of a creator. According to CHARLES C. KEMP these characteristics are [13]:

- People are present around.
- Controlled environment cannot be assured.
- Other autonomous characters are present.
- Dynamic changes in world like flood, earth quake etc.
- Variations in placing the objects.
- Long distances between locations.
- Need of special tools.
- Changings in object's type and appearance.
- Non rigid objects and substances may need to be manipulated.
- Variation in environment's structure.
- Some architectural obstacles.

Dr. Nick Hawes, Senior Lecturer in Intelligent Robotics, School of Computer Science, and University of Birmingham says "There's this huge excitement around robots. Everyone really believes, as we do ourselves, that

 DOI: https://doi.org/10.30564/jcsr.v1i1.412

robots are going to have a huge impact on our future – in workplaces, in roles in various industries." Agreeing with the words of Dr. Nick Hawes, I would like to say that if these robots are really our future then why not pay attention on their designs and make them work efficiently for the environment they are build.

Requirement specification is an important feature in describing what you want your robot to actually be. Most of the times we do not understand that in which type of environment our robot is going to work and how it is going to interact. This creates a lot of problems as we cannot make our robot compatible and according to the needs of environment.

For this let us take an example. Worker's Daily Newspaper published an article that said, "Three restaurants in the southern Chinese city of Guangzhou have been forced to fire all of their robot staff after their utter incompetence began costing them money. Two of the restaurants have closed completely after discovering the clumsy waiters could not perform simple tasks like taking orders, pouring drinks and carrying soup, reports say.

The slacking robot team also kept breaking down and after a string of complaints the third restaurant mentioned above decided to sack all but one and bring back human employees[12].

Figure 1. Robot still working

From the above example it has become clear that if we have considered the fact that these robots are being used for the environment in which you need speed , a great voice recognition system and some important function like leading to a free table and pouring water etc. These type of problems can be rectified if before making a robot we consider in which type of environment these robots are going to work in and how they are going to perform their work efficiently.

Considering another example, we encounter a humanoid robot known as Sophia, which was created to live in a human environment, work with them, help and play. But what if they are repeatedly seeing some actions that like killing and beating? Being Artificially Intelligent they will pick how to perform these actions and would be able to do that [14-17]. Now, that's an alarming situation that when Sophia was asked a question that says something about humans she said, "I Will Destroy Humans".

Emotion and sociable robots, a research paper of the MIT media lab included ," autonomous robots are designed to operate as independently and remotely as possible from humans, often performing tasks in hazardous and hostile environments (such as sweeping minefields, inspecting oil wells, or exploring other planets). Other applications such as delivering hospital meals, mowing lawns, or vacuuming floors bring autonomous robots into environments shared with people. However, a new range of application domains (domestic, entertainment, health care, etc.) are driving the development of robots that can interact and cooperate with people as a partner, rather than as a tool[9,18].

Considering the above lines to be true in the near future we have to focus completely on gathering appropriate requirements for robots so that they cannot harm humans in any possible way.

In June 2011 Sakai Yasuyuki wrote an article on "Japan's Decline as a Robotics Superpower". "The two articles that follow highlight the failures of R&D in Japanese robotics engineering that were dramatically and tragically revealed by the earthquake and tsunami-driven meltdown of TEPCO's nuclear power plants at Fukushima. Vbgy787uContrary to expectations that Japan would be a leader in manufacture of disaster relief robots that could have been used in problem solving and cleanup in the wake of the Fukushima Daiichi nuclear disaster, three months after 3.11, Japan's robots have yet to make a significant contribution. These articles explain why Japan, in general, its robotics industry in particular, proved unprepared for severe nuclear accidents, and how haphazard the government and the nuclear industry has been in developing robots that could have eased the crisis[19].

Apart from other reasons, one reason of robot failure was also the ignorance factor regarding environment .Robot was designed, a huge amount of money was served in making the robot. But the robot failed at the time when it has to do its work at the target location.

The article by O. Khatib also gives emphasis on environmental interaction of robotics. "This article discusses the basic capabilities needed to enable robots to operate in human-populated environments for accomplishing both autonomous tasks and human-guided tasks. These capabilities are key to many new emerging robotic applications

DOI: https://doi.org/10.30564/jcsr.v1i1.412

in service, construction, field, underwater, and space. An important characteristic of these robots is the "assistance" ability they can bring to humans in performing various physical tasks. To interact with humans and operate in their environments, these robots must be provided with the functionality of mobility and manipulation. The article presents developments of models, strategies, and algorithms concerned with a number of autonomous capabilities that are essential for robot operations in human environments. These capabilities include: integrated mobility and manipulation, cooperative skills between multiple robots, interaction ability with humans, and efficient techniques for real-time modification of collision-free path. These capabilities are demonstrated on two holonomic mobile platforms designed and built at Stanford University in collaboration with Oak Ridge National Laboratories and Nomadic Technologies[20,21].

So before making a new humanoid robots we shall consider the fact that environment is the most important factor. We have to consider in which environment our robot is going to work. What will be the circumstances their? How these robots are going to work efficiently without creating any disturbance?

For this, we have to make environment an important attribute to consider while taking requirements for our robots. We have to specify correctly, what it is for and why? Only then we will be able to build a robot that can work efficiently and would be reliable!

3. Problem Statement

The lack of consideration that environment has an impact on working of robots and lack of experience for gathering the requirements in which robots will be working while making them.

4. Research Questions

- How to take requirements that will help in making environment friendly robot?
- Which factors of environment have impact on making robots?
- Why robots should be made more precisely according to the environment needs?

5. Methodology

5.1 Research Type

The type of research we are using in finding the answer to those research questions is quantitative methodology. Quantitative methodology targets to gather the information about the human environment and its effect on the working of autonomous robots. This phenomenon can be examined through some statistical analysis on the gathered data. Along with the questionnaire we had some interviews with some researchers in the field of robotics from some software houses, as it was easier to gather the information by questionnaire so that analysis could be performed.

5.2 Method of Sampling

The sampling method for the research is choosing random students (round about 50) from the Software Engineering and Computer science department of University of Gujrat. Permission of the supervisor was granted to do the research in the university premises.

5.3 Respondents

The respondents of this research were the random students picked from SE and Cs department. We choose these departments as it was easy and economical to gain information from them and prove our point that environment do count in the efficient working of robots and should be considered while gathering requirements for making it.

6. Findings

Gathering all the data from the questionnaires and conducted interviews we have to find the following facts:

a)As it can be seen from the Fig.2 that most of the respondents actually consider and support the fact that environment is an important factor that should be considered while making robots. This graph suggests that while making robots appropriate requirements that are related to its working in the specific controlled environment should be considered and worked upon.

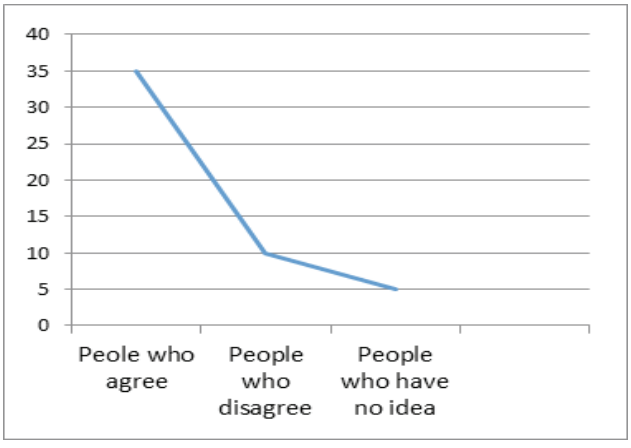


Figure 2. Environment effect

b)There are some certain variables in a controlled environment that should be mainly considered while gathering requirements which are:

·Number of people
·Other interactive machines
·Dynamic variations in the world

 DOI: https://doi.org/10.30564/jcsr.v1i1.412

·Change in size
·Change in distances
·Need of specialized tools
·Architectural structure

From these variables only some are worth considering and they can be analyzed by Fig. 3 which is given below.

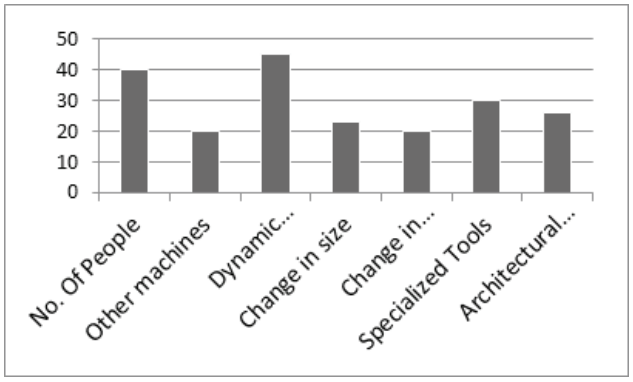


Figure 3. Variables

The above figure only suggests that number of people in a controlled environment, dynamic variations in the world, specialized tools and architectural structure have a great impact on working of robots in a given environment.

7. Analysis

So by the analysis of all our findings we have come to answers research questions latterly asked. Firstly the question arose how to make environment friendly robots? The answer to that question is making environment friendly robots means we are trying to make more interactive and speedy systems. They can be made by inculcating all overall expectations of the users. A survey should be conducted in which you should gather what are the user expectations regarding this robot? What are they thinking? Which features are important in it?

Second question was which factors of environment have impact on working of robots? So as it can be seen from figure 2 that all the factors discussed above have impact on working of robots but the main factors of environment that effect how robot will be performing it's tasks are:

- Number of People Present around
- Dynamic Variations
- Use of Specialized tools
- Architectural Structure

So, while gathering requirements for the robots, these factors should always be considered and information about it should be gathered so that we can work with robots with efficiency.

The last question is why environment should be considered while making robots? So all the above things discussed lead to a single point that human environment is very difficult to survive in. We should consider all the aspects of environment so that a working robot can be presented in the market.

8. Conclusion

The answered research questions in the analysis show that the problems with the working of robots and their breakdown can be controlled if we consider environment as an important factor and start gathering more information about the environment in which robots will be working in the requirement gathering phase. It may need some extra effort and time but it is far better than using the robots that cannot fit in the environment properly and just not break down at eleventh hour so that reliability of robots can be guaranteed. Making robots that can fit in the society efficiently and that can truly fulfill their purpose help in saving time and money of the companies making it. So a small effort in data gathering process can help a lot in making environment fit robots.

References

[1] Edsinger, A., and Kemp, C.C, Human-robot interaction for cooperative manipulation: Handing objects to one another, in Editor (Ed.)^(Eds.): 'Book Human-robot interaction for cooperative manipulation: Handing objects to one another' (IEEE, 2007, edn.), pp. 1167-1172

[2] Khan, S.U.R., Lee, S.P., Dabbagh, M., Tahir, M., Khan, M., and Arif, M., RePizer: a framework for prioritization of software requirements, Frontiers of Information Technology & Electronic Engineering, 2016, 17(8), 750-765

[3] Khatib, O., Yokoi, K., Brock, O., Chang, K., and Casal, A., Robots in human environments, in Editor (Ed.)^(Eds.): Book Robots in human environments (IEEE, 1999, edn.), 213-221

[4] Ajoudani, A., Zanchettin, A.M., Ivaldi, S., Albu-Schäffer, A., Kosuge, K., and Khatib, O., Progress and prospects of the human–robot collaboration, Autonomous Robots, 2018, 1-19.

[5] Kose, T., and Sakata, I., Identifying technology convergence in the field of robotics research, Technological Forecasting and Social Change, 2018.

[6] Faudzi, A.A.M., Ooga, J., Goto, T., Takeichi, M., and Suzumori, K., Index finger of a human-like robotic hand using thin soft muscles, IEEE Robotics and Automation Letters, 2018, 3, (1), 92-99.

[7] Arif, M., Wang, G., and Balas, V.E., Secure VANETs: trusted communication scheme between vehicles and infrastructure based on fog computing, Stud. Inform. Control, 2018, 27 (2), 235-246.

[8] Arif, M., Wang, G., and Chen, S., Deep learning with non-parametric regression model for traffic flow predic-

  DOI: https://doi.org/10.30564/jcsr.v1i1.412

tion, in Editor (Ed.)^(Eds.): 'Book Deep learning with non-parametric regression model for traffic flow prediction' (IEEE, 2018, edn.), 681-688

[9] Arif, M., Wang, G., and Peng, T., Track me if you can? Query Based Dual Location Privacy in VANETs for V2V and V2I, in Editor (Ed.)^(Eds.): 'Book Track me if you can? Query Based Dual Location Privacy in VANETs for V2V and V2I' (IEEE, 2018, edn.), pp. 1091-1096

[10] Arif, M., Wang, G., Wang, T., and Peng, T., SDN-Based Secure VANETs Communication with Fog Computing, in Editor (Ed.)^(Eds.): 'Book SDN-Based Secure VANETs Communication with Fog Computing' (Springer, 2018, edn.), pp. 46-59.

[11] Sian, N., Sakaguchi, T., Yokoi, K., Kawai, Y., and Maruyama, K., Operating humanoid robots in human environments, in Editor (Ed.)^(Eds.): 'Book Operating humanoid robots in human environments' (2006, edn.), pp.

[12] Lee, J.H., The capitalist and imperialist critique in HT Tsiang's And China Has Hands, Recovered legacies: Authority and identity in early Asian American literature, 2009, pp. 80-97

[13] Kemp, C.C., Edsinger, A., and Torres-Jara, E., Challenges for robot manipulation in human environments [grand challenges of robotics], IEEE Robotics & Automation Magazine, 2007, 14 (1), pp. 20-29

[14] Arif, M., and Hussain, M., Intelligent agent based architectures for e-learning system: survey, International Journal of u-and e-Service, Science and Technology, 2015, 8(6), pp. 9-24

[15] Arif, M., and Mahmood, T., Cloud Computing and its Environmental Effects, International Journal of Grid and Distributed Computing, 2015, 8 (1), pp. 279-286

[16] JAVAID, Q., ARIF, M., AWAN, D., and SHAH, M., Efficient facial expression detection by using the Adaptive-Neuro-Fuzzy-Inference-System and the Bezier curve, Sindh University Research Journal-SURJ (Science Series), 2016, 48(3)

[17] Javaid, Q., Arif, M., Talpur, S., Korai, U.A., and Shah, M.A., An intelligent service-based layered architecture for eLearning and eAssessment, Mehran University Research Journal Of Engineering & Technology, 2017, 36 (1), pp. 97.

[18] Breazeal, C., Emotion and sociable humanoid robots, International Journal of Human-Computer Studies, 2003, 59, (1-2), pp. 119-155

[19] Yasuyuki, S., Japan's Decline as a Robotics Superpower: Lessons From Fukushima, 2011.

[20] Khatib, O., Yokoi, K., Brock, O., Chang, K., and Casal, A., Robots in human environments: Basic autonomous capabilities, The International Journal of Robotics Research, 1999, 18 (7), pp. 684-696.

[21] Brock, O., and Khatib, O., Elastic strips: A framework for motion generation in human environments, The International Journal of Robotics Research, 2002, 21 (12), pp. 1031-1052.

 DOI: https://doi.org/10.30564/jcsr.v1i1.412

BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
http://ojs.bilpublishing.com/index.php/jcsr

ARTICLE

Augmented Reality Book for Preserving Malay Traditional Dances: A Case Study

Norfadilah Kamaruddin*
Faculty of Art and Design, University Teknologi MARA (UiTM), Shah Alam, Selangor, Malaysia

ARTICLE INFO

Article history:
Received: 17 December 2018
Accepted: 7 January 2018
Published: 22 January 2019

Keywords:
Malay traditional dance
Tarian Piring
New media technology
Augmented reality book

ABSTRACT

Malaysia is known throughout the world for its multiculturalism. As a multiple ethnic country, many countries are looking on Malaysia as a great example of peaceful co-existence races and belief where all the ethnic groups in Malaysia live together in harmony and enrich the country's cultural lifestyle. Within that, Malaysia also consists of a collective blend of food, traditions, clothing and customs. Towards that, traditional dance is the treasure of art and culture. Therefore, with modern era and technology nowadays, it has led the younger generation care less about traditional dance. Beside, a printed media such as bunting, banners and pamphlets are less effective in promoting the traditional dance. By concerning this, this research study aims to preserving the traditional dance among young generation towards new media technology. In explaining the issues, a case study through quantitative approaches of questionnaires survey and interviews was used in studied the uniqueness of traditional Malay dance and further proposes a new approach for preserving the traditional Malay dance awareness among the young generation. The research significantly impacts the publics particularly on the new generation towards uniqueness of traditional dance in Malaysia. It is also contributes to the National Heritage Department and the National Arts and Culture Department where the documentation could be used as a collection of cultural and heritage books in the form of new media technology for young generation.

1. Introduction

Traditional dance is another aspect of many traditional Malaysian cultures that many would argue is rapidly changing in the face of globalization. Moreover, by looking for the traditional dance uniqueness in Malaysia, that is rarely perceived and unpopular among young generation in Malaysia. Each state has its own form of their traditional dance. Visual communication plays an important role in our daily lives. Furthermore, the present generation of Y and Z are more interested in picture messages rather than an old style message, which contains too many words and can bore them. If visual communication is used and applied in the right way, it will contribute to various benefits in countless angles. In the era of modernization, society gains and provides information in a limited amount of time so, visual communication helps to disseminate and broadcast information with quicker and even with better ideas.

**Corresponding Author:*
Norfadilah Kamaruddin
Faculty of Art and Design, University Teknologi MARA (UiTM), Shah Alam, Selangor, Malaysia
Email: norfadilah@salam.uitm.edu.my

DOI: https://doi.org/10.30564/jcsr.v1i1.491

As the time passes by today, it makes some past event is forgotten and there is no means to retain it especially on the heritage history. Moreover, due to the today modern technology, the younger generation has not exposure and care about our heritage including traditional dances and it causes them unfamiliar with the history. Extensively, existing materials such as pamphlets and poster on promoting heritage history do not really attracting the young generation to know about the heritage and concern to preserving it. With this issue, one study was conducted among Public aims to preserving the traditional dance among young generation towards new media technology. In explaining the issues, two research objectives was formed in studied the uniqueness of traditional Malay dance and further proposes a new approach for preserving the traditional Malay dance awareness among the young generation. This research significantly impacts the publics particularly on the new generation towards uniqueness of traditional dance in Malaysia.

2. Exploratory Study on Augmented Reality Book for Preserving Malay Traditional Dances among Young Generation

2.1 Malay Traditional Dances in Malaysia

Each ethnic groups in Malaysia has its own dance forms, which can be characterizing by its culture and identifying with certain religious practices which are often performed in wedding ceremony, cultural shows, religious ceremonies or other public events. The dances of the 3 major racial groups in Malaysia are categories as Malay dances, Chinese dances and Indian dances.

Malay dances or also known as Tarian Melayu portrays the customs or adat resam and culture of the Malays. It depicts the true nature of the Malay people and their way of life. Generally, Malay dances are divided into two main categories, which are the "original" Malay dances and "adopted" Malay dances. The "original" Malay dances originally encompassing of Sumatra, the Malay Peninsular, Singapore, the Riau Archipelago and Borneo, and its origins can be traced back to the early Malay civilizations. While, the "adopted" Malay dances are influenced by foreign cultures due to political and historical events. The various forms or styles of Malay dance are further categorized by its beats (rentak) and rhythm (irama).

2.2 New Media and Technology Influences Traditional Dances Awareness

Taken in the retail industry for marketing purposes, "pop-up" book style has become a popular concept for products that want to attract the attention of the people[1]. Pop-up book moreover means a book with a page that appears when opened in three-dimensional form. The term pop-up has been used very easily to illustrate something used for the short term to get more exposure about a project or information because it is an economic model for social interaction with interested users. Therefore, a pop-up book is not just designed for children. Instead, these books containing an innovative tools for teaching anatomy, predicting, and telling the future[2]. Moreover, this pop-up book gives many different effects to readers[3].

When technology grow, television advertising or TV ads is a very influential medium as it is a major aspect of culture, news sources, education and entertainment to the public. Television also is a unique and dynamic advertising medium that offers stimulus to two senses, sight and sound. The fact that TV Ads takes place in real time and using both audio and visual communication channels simultaneously that has multiple effects on the audiences.

In contrast to today's age, most people are more likely to have gadgets to access advertisement released on television. According to Johnson (2011)[4], augmented reality has existed for more than three decades. Raphael[5] further explains that the purpose of augmented reality is to add a layer of information and meaning to a real place or object. Moreover, Kipper and Rampolla[6], reported that incensement of augmented reality takes digital information such as images, audio, video, or touch sensations and directs them into a realistic environment. While augmented reality can be used to improve all senses, the most common use is visual. Raphael [5] further noted that increasing uses of advanced gadgets with more internet access was enabled the augmented reality extended to be more accessible to the general public. Increased augmented reality also offered many opportunities for the future[7].

2.3 Principles and Elements to be use in Augmented Reality

In producing augmented reality product, a good designer needs to have a good eye in assessing a composition, can see the aesthetic and communicative potential of the subject in a picture. In this study, the technology medium used must include the photograph of human figures or characters that are dancing so that the message to be delivered becomes clearer. Other than that, typography is also placed on this new media as content that can help in delivering the message more clearly. This is because; texts show a positive impact on the time those readers spending on viewing an advertisement [8]. In addition to generating content, fonts are a very important thing to reveal the content is interesting and clear. Fonts can be divided into six main categories of serif, sans serif, script, black letter, novel, and dingbat[9]. In producing an augmented reality

 DOI: https://doi.org/10.30564/jcsr.v1i1.491

book, the use of serif and sans serif combination are the most appropriate. Serif is a rectangular, sharp, straight, curved, thin, thick or stroke-shaped font on its tip[9]. The sans serif font does not have a pointed tip, it looks even and thick which is very suitable to the type of theme used to portray the new medium technology.

The choice of color in producing an augmented reality book is very important as the color can bring a person's mood to life, help users to see and choose what they like [10]. Viewers also are more interested and turn their eyes on the use of colors that they are rare and vibrant in delivering the ideas and emotions to the audience. Consequently, the color must have an important role in marketing and advertising [11]. According to Krauser[9], "when the eye takes note of a pattern built from duplication of a shape, it generally recognizes the design in search of meaning since patter is unlikely to contain anything other than identical shapes, over and over." Repetition patterns can attract and entertain the eyes of an individual because not too many different patterns are used, just the same pattern returned. These principles of design are suitable to put into the new media that will be produced.

2.4 ADDIE Model for Design Development

ADDIE model is a teaching design medium, which means Analysis, Design, Development, Implementation, and Evaluation of learning materials and activities. Firstly, the Analysis phase is the basis for all other teaching design phases. During this phase, the researcher can determine the problem, identify the source of the problem and determine the possible solution in the study. This phase may include specific research techniques such as requirement analysis and task analysis. The findings of this phase often involve the research objectives, and the task list to be directed. The results of this analysis will be the source for the design phase.

Second, the Design phase involved the use of findings from the analysis phase to designing strategies for developing the study. During this phase, the researcher creates outlines how to achieve the research goals determined during the phase analysis and further expands the basics of the study. Some design phase elements can be included such as writing descriptions, target populations, conducting learning analysis, writing aims and objectives, selecting the methodology system to use and composing the final work to be implemented. The results of the design phase will be the source of information for the development phase.

The Development stage is built upon the phase of analysis and design. The purpose of this phase is to produce a lesson plan and research materials. During this phase the researcher will develop ideas, media to be used and supporting documentation to produce materials for the success of the study. It also involves media such as computers, gadgets and more.

After that, the Implementation phase refers to the presentation of planned ideas whether they are based on laboratory or computer-based. The purpose of this phase is the delivery of effective and efficient instructions. This level must be aligned with the researcher's understanding of the material produced, whether or not it is true. It must ensure the knowledge of the researcher from the research assignment towards the job.

Lastly is the evaluation phase. This phase measures the effectiveness and efficiency of the instruction. Evaluation should occur throughout the entire research design process - in phase, between phase, and after implementation. Evaluation may be Formative or Summative. Formative evaluation continues during and between phases. The purpose of this evaluation is to improve the material produced before the final version is implemented. Summative evaluation usually occurs after the final version of the instruction is implemented. This type of evaluation assesses the overall effectiveness of the instructions. Data from summative evaluation are often used to make decisions about the resulting material.

3. Methodology

Explanatory study is to explore an unknown area or to investigate the possibility of conducting a particular research study. In this study is to explore the authenticity and uniqueness of the Traditional Dances is still present or known to the public, especially to the young people. After coming into contact with the problem, researchers are looking for ways to maintain the uniqueness and distinctiveness of the dance and the attractive medium to develop in today's era. Explanatory studies are also conducted to develop, improve and test a medium whether it is effective or not.

3.1 Research Design

The researcher was conducted this research using quantitative method as a whole because the researcher solely conducted a survey using questionnaire set. A questionnaire survey was used to get information and feedback from respondents towards the issue. Previously, the researcher used interview method to find and get the problem statement for this study to strengthen the problem and process of doing this research.

The survey questionnaire is a research method that using the list of questions written, and the answers by the respondents are recorded. In other words, it also communicates with people but not through conversa-

DOI: https://doi.org/10.30564/jcsr.v1i1.491

tions, it uses paper and writing answers. The survey questionnaire is designed according to the researcher's creativity whether it is point or answer in writing. The question is broken into three parts so that questions are not fibrous and are systematically arranged and easy to understand step by step. The question is designed to get information, help in the study and get the opinion of the respondents.

In this study, researcher conducted surveys with public aged 13 to 27 years old representing a handful of age groups including generation Y and Z. Selecting the public as the samples because in the survey questionnaires there is a question of what kind of dance they know and do not know so, public is the best choice because of the weight both sides and fair and just. The research design and strategy for this study as shown in Figure 1.

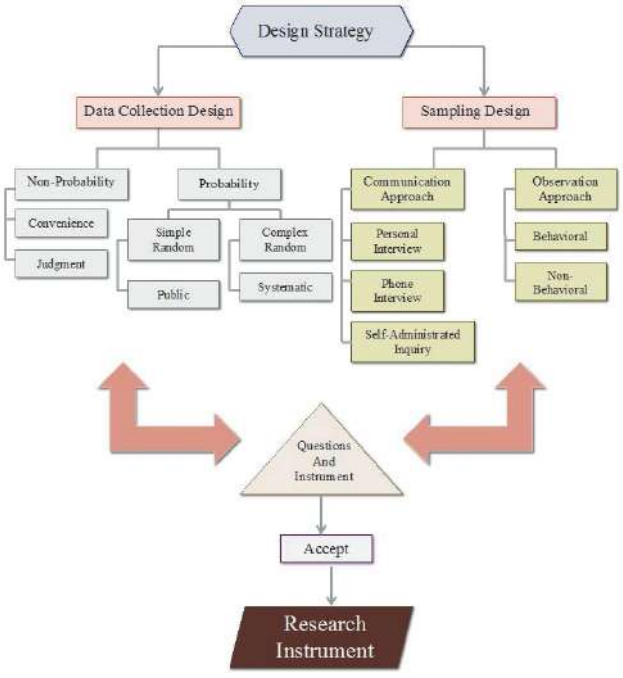


Figure 1. Research design

3.2 Data Collection and Analysis

The survey questionnaires were distributed by the researcher themselves to a sample of 100 participants. 50 people were selected around the Section 7 Commercial Center of Shah Alam and another 50 person were chosen at Seremban City. The respondents were targeted based on their residential status in that area which mainly consists of students and youths. Do not focus on the criteria of respondent as it is targeted to the public at large. The survey question was distributed and collected at the same time by the researcher within the selected area for five consecutive days, which is from Saturday to Wednesday.

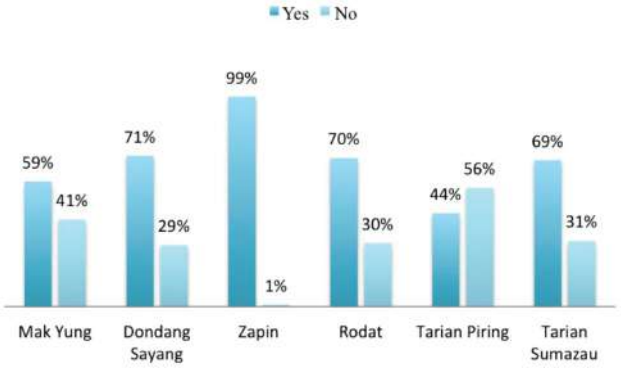


Figure 2. The type of traditional Malay dance

Based on the above charts and graphs, it is clear that the Zapin dance originating in Johor is the highest dance known to the respondents with 99% and 1% who do not know the Zapin dance. The second highest is Dondang Sayang from Melaka state as 71% respondents know it and only 29% do not know it. The third is Rodat from Terengganu, as much as 70% who know this dance and 30% do not know this dance. The fourth is Sumazau dance originating from Sabah, 69% respondents know this dance and 31% do not know this dance. The fifth is Mak Yung from Kelantan, there are 59% respondents who know this dance and 41% who do not know this dance. Sixth or even the lowest percentage is Tarian Piring originating from Negeri Sembilan as much as 44% who know about this dance and 56% do not know and this means most respondents do not know about this Tarian Piring.

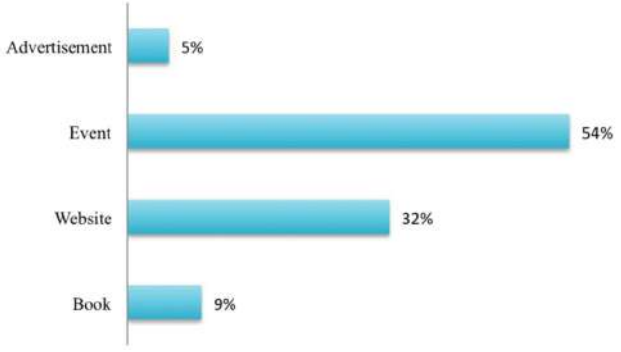


Figure 3. Medium to get information about traditional dances

In point of fact, the medium that is often used to obtain the information about traditional dance is through an event with the highest percentage of 54%. Website is the second most popular medium of information on the topic of traditional dances. Only 9% of the information is obtained through books and 5% through advertising. It indicates that traditional dances are popular found out through events which may consist of cultural nights and local fes-

 DOI: https://doi.org/10.30564/jcsr.v1i1.491

tival rather than only leaflets and brochure advertisements distributed.

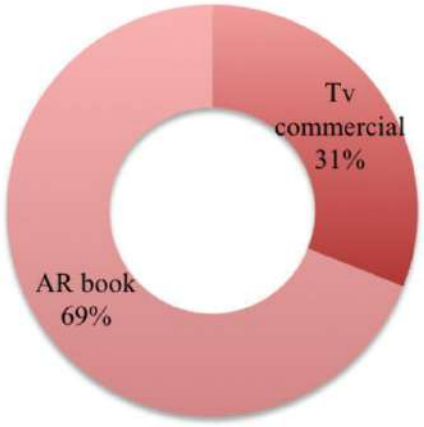


Figure 4. Demand for new media as a medium for maintaining traditional dances in Malaysia

A total of 69% of respondents chose the Augmented Reality Book as a new medium to protect and preserve traditional Malaysian dances. It demonstrates that the respondents are eager to see a new type of medium to sustain the traditional Malay dances .The other remaining 31% has chosen the common television commercial as a new medium in order to preserve the traditional dances.

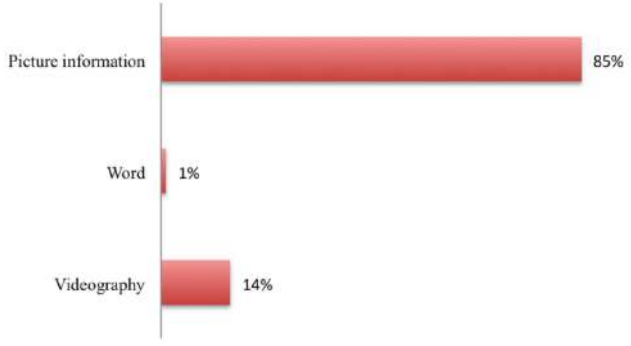


Figure 5. New media types to retain traditional dance in Malaysia

As much as 85% choose pictorial information as one way to put in new media to maintain traditional dance in Malaysia. 14% choose videography and only 1% chooses writing. This means that the respondents refuse to have a new media which contain too much text and words as compared to the pictures and images. They would prefer to have something more appealing to see and something interesting to read which can be full of imagery.

Significantly, a total of 74% of the participants choose the concepts with the use of something real to be applied in the new media to be produced. It shows that nowadays people are looking for something intrinsic yet realistic concepts rather than caricatures or even made-up animations in order to appreciate the new media to retain the traditional dances in Malaysia. On the other hand, 15% respondents choose the concept of cartoon and a minority of 11% choose fantasy concept.

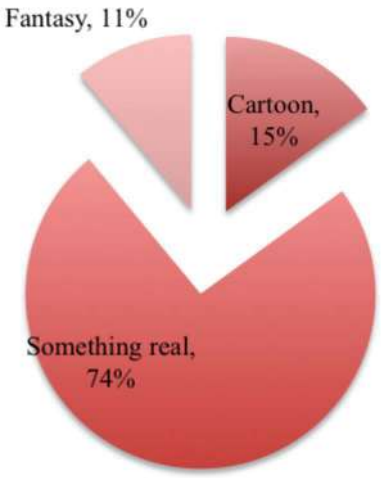


Figure 6. The concept of new media to retain traditional dance in Malaysia

4. Design Development

4.1 Design Development for a New Book

In the beginning, the researcher had searched and explored the size and shape of the existing book in market as a reference material. The process of searching this is conducted in libraries and bookstores. The final size of the product will be the result of the observations. From that, researcher also got an idea on how to produce a new Tarian Piring book form so that it looks more interesting. In researcher observation and it also an opinion from someone while doing a survey, size and shape play an important role because it gives a positive impact to the reader, if it is too small, it will cause readers to read and retrieve the information in the book and if it is too big, it will make it hard for readers to carry it and keep it. Suitability must be taken so that it is easy in all respects, example the size of the writing in the book, the size of the image used and the way to hold and carry the book.

4.2 Sketches

In creating whatever designs and drawings, sketches are the first steps that are so prevalent. From where the idea comes, the arrangement and the colors to be used. First of all, the researcher made a few sketches to get a random idea on how to make a Tarian Piring book. In a few sketches, the researcher also makes the layout of the writing in the book, the size of the picture and the shape of the picture to be placed and the suitability of the layout

DOI: https://doi.org/10.30564/jcsr.v1i1.491

in the book. In order to make the book systematic and organized, the partition of each chapter is very important to make it easier and more structured. Researcher also made the book look more exclusive.

4.3 Design in Adobe Illustrator

After making the first step that which is sketches, researchers have also come up with ideas and the next step is to create a digital design using Adobe Illustrator. Researchers re-create the sketches in Adobe Illustrator according to the sketches that have been produced. The suitability of using Adobe Illustrator is that it is easy to play with the colors and use the right colors on the spot. It is also a shadow of the book to be produced. If the sketches are not very clear and still difficult to get ideas, in digital form it can be easily and quickly generated. The resulting book size is 8inch x 8inch equivalent 20.5cm x 20.5cm. The size is not too big and not too small. Many books are also produced in these dimensions. The researcher still maintains a lot of writing because this book is a '2 in 1' book, which means a reference book that can be a reference material and a reading material to the public and at the same time applies the concept of Augmented Reality that modernizes the book according to contemporary circulation. A widely used article is to preserve the authenticity of the information and the history of the Tarian Piring itself so that it does not lose the deck of the ages and Augmented Reality is used as a tool to attract and double inform because it has a video that facilitates readers to continue to recognize Tarian Piring without searching information on Google or to browse YouTube.

4.4 Virtualization Technology

The books that have been produced and printed only then can proceed to the next step. As already stated, the researcher produced an Augmented Reality book that implements the video inside it when it is scanned on the resulting book. The researcher uses the HP Reveal application to generate Augmented Reality. This app can be downloaded for readers outside of here to scan this Tarian Piring book. This app needs to be signed in first using a research account to get Augmented Reality results in this Tarian Piring book. Below are steps with images in creating Augmented Reality.

4.5 Implementation Process

After producing an Augmented Reality's book on Tarian Piring, researcher further conducted an implementation process to determining the effectiveness of the book with the public. The implementation process questionnaires were distributed to 40 participants around Section 7 Commercial Center in Shah Alam. The respondents were targeted based on their residential status in that area which mainly consists of students and youths. The result revealed as below:

Table1. Implementation process

	YES	NO
Liked the color of the book	92%	8%
The color suggested		Red
The color gives feeling	83%	17%
Appropriate size	95%	5%
Composition and layout is easy to read	93%	8%
Text are easy to read and understand	95%	5%
Function as promotional item	100%	
Book is a very informative	100%	
Increasingly advanced era of technology and modernization	100%	
Book can convey information	100%	
Video in the book can be entertaining	97%	3%

5. Conclusion and Recommendations

5.1 Conclusion

The result of this study indicates that the new media technology that has been selected by most of the respondent is the Augmented Reality book as compared to pop-up books and television commercial. The youths nowadays moreover only focus on the contemporary arts and preoccupied with gadgets day and night without trying to know their own traditional cultures. Furthermore, printed media such as banners, bunting and pamphlets only 10% are effective for promoting the traditional Malay dance. Finally, the matter can be summarized is that the uniqueness of Tarian Piring can be preserved through the creation of Augmented Reality book for the younger generation which tend to choose new media technology in this endless era of technology advancement. Other than promoting the traditional dance to the upcoming generations, but also can help to maintain a long lasting exquisite arts and cultural heritage in the future.

5.2 Recommendations Based on the Findings

1) In term of element of designs, the heavy content of words than pictures should be reduced to a suitable quantity so readers will not feel bored easily.

2) The content can be directly scanned to produce the Augmented Reality concept automatically rather than scanning on the yellow vector only.

DOI: https://doi.org/10.30564/jcsr.v1i1.491

3) The digital book should have more Augmented Reality concept in the book not only focuses on the Tarian Piring dance video but also the history or the dancing equipment. 4) The book can be produced in English in order to attracting foreign tourist to read as it can be read by many people not just Malaysians.

5.3 Recommendations for Future Research

As this study had only focused on Tarian Piring which is a traditional Malay dance of Negeri Sembilan Darul Khusus, it is recommended that further studies should be carried out on dances from another state whether traditional or contemporary to see whether there are many or fewer similarities in the findings. Besides that, this study is only conducted in the area of Seremban city and Section 7 Commercial Center of Shah Alam. It is also recommended that further studies should be in other province in Negeri Sembilan and other states too. Furthermore, future research could explore on the interest of other popular dances in Malaysia as what really attracts the people to admire it and how they do it. Lastly, although Tarian Piring is not that famous to our people, it might be a good idea to keep our cultural heritage to not be forgotten in the future.

Additionally, the existing books that have been produced are in Malay language. The next recommendation is that the book is produced in English as well as to facilitate tourists from abroad to understand this Tarian Piring book if this book is sold or marketed in the museum. Foreign tourists will also feel close to having a book that is easy to understand the language they speak English. At the same time they will be interested in buying and being used as a collection of dance books originating from Malaysia.

In addition, the results that have been found during the evaluation process are the renewable technology of the Hologram media. It is also comparable and in parallel with the time passage that progresses forward when using this Hologram method. It also achieves the research objective of maintaining and preserving the Tarian Piring by using up-to-date technology media. At the same time the increasingly advanced technology media captures the hearts of young people who are in disguise and progress and they are keen to know about this Tarian Piring.

References

[1] Burgess, (2012). Pop-up Retailing: The Design, Implementation, and Five-Year Evolution of an Experiential Learning Project. Journal of Marketing Education, 34(3), 284-296

[2] Begay, M. (2015). Designing Children's Interactive Pop-up Books: Creating enhanced experiences through the incorporation of animation principles and interactive design. ProQuest Information and Learning Company, 6

[3] Chiu, M. W. (2013). Using Multimedia to Promote Environmental Protection. ProQuest Information and Learning Company, 1.

[4] Johnson, L., Smith, R., Willis, H., Levine, A., & Haywood, K. (2011). The 2011 Horizon Report. Austin, TX: The New Media Consortium

[5] Raphael, R. (2011). Learning connections: Abracadabra-It's augmented reality Learning & Leading With Technology, 38(8), 24-26.

[6] Kipper, G., & Rampolla, J. (2013). Augmented reality: An emerging technologies guide to AR. Waltham, MA: Syngress

[7] Kapp, C,. & Balkun, M. (2011). Teaching on the virtuality continuum: Augmented reality in the classroom. Transformations: The Journal of Inclusive Scholarship and Pedagogy, 22(1), 100-113.

[8] Olney, Thomas J., Morris B. Holbrook, and Rajeev Batra (1999), "Consumer Responses to Advertising : The Effects of Ad Content, Emotions, and Attitude toward the Ad on Viewing Time". Journal of Consumer Research. 17.

[9] Krause, J. (2015). Visual design: ninety-five things you need to know, told in Helvetica and dingbats. San Francisco, CA: New Riders, Peachpit.

[10] Laar, V. D., G., & Berg, L. V. (2003). "Assessing structure's branding impact". BrandPackaging.

[11] VanHurley, V. L. (2007). The Influence on Instruction of Regulating Teachers' Use of Copyrighted External. ProQuest Information and Learning Company, 1.

DOI: https://doi.org/10.30564/jcsr.v1i1.491

BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
http://ojs.bilpublishing.com/index.php/jcsr

ARTICLE

Churn Prediction Task in MOOC

Lisitsyna Liubov* Oreshin S.A.

ITMO University, Kronvrkskiy pr. 49, Saint Petersburg, 197101, Russia

ARTICLE INFO

Article history
Received: 18 February 2018
Accepted: 5 March 2019
Published: 7 March 2019

Keywords:
Machine learning
Data science
Exploratory data analysis
Logistic regression
Gradient boosting on trees
Stacking
Classification; Ranking

ABSTRACT

Churn prediction is a common task for machine learning applications in business. In this paper, this task is adapted for solving problem of low efficiency of massive open online courses (only 5% of all the students finish their course). The approach is presented on course "Methods and algorithms of the graph theory" held on national platform of online education in Russia. This paper includes all the steps to build an intelligent system to predict students who are active during the course, but not likely to finish it. The first part consists of constructing the right sample for prediction, EDA and choosing the most appropriate week of the course to make predictions on. The second part is about choosing the right metric and building models. Also, approach with using ensembles like stacking is proposed to increase the accuracy of predictions. As a result, a general approach to build a churn prediction model for online course is reviewed. This approach can be used for making the process of online education adaptive and intelligent for a separate student.

1. Introduction

The main problem of using Massive Open Online Courses (MOOC) is their low performance (no more than 5%), which is estimated as the proportion of successfully completing the course to the total number of students registered at the start of this course. The low performance analysis of MOOC [1] revealed a number of reasons related to the poor readiness of listeners for e-learning, with low motivation to achieve higher learning outcomes. To solve the problem, approaches [1–4] have been proposed and experimentally confirmed, aimed at situational awareness training of the student when working with electronic forms before learning.

In this paper, we adapt a churn prediction task to predict students' churn in MOOCs. Classical churn prediction task is about building a model which finds a list of clients who are likely to break their contract. This task is solved to predict students' churn in classical higher education [5]. If adapt this task to MOOCs, the formulation is different. Firstly, we need to select the right time period in the course, so we can use the data of students' activity before this point. There may be several such points. Secondary, we consider students a churn if they haven't finished the final exam of the course. Further, we propose an approach to solve this problem demonstrating its effectiveness on

**Corresponding Author:*
Lisitsyna Liubov,
ITMO University, Kronvrkskiy pr. 49, Saint Petersburg, 197101, Russia
Email: lisizina@mail.ifmo.ru

 DOI: https://doi.org/10.30564/jcsr.v1i1.537

online course "Methods and algorithms of graph theory" by IFMO University.

This article proposes a user-based approach to sampling statistical data recorded by the e-learning system during the course to predict the performance of an online course. After the correct sample is collected, the problem is formulated in machine learning terms. Proposed in the paper approach of constructing the correct sample for prediction the performance of online courses and building predictive models is used for the further development of the MOOC platforms with the aim of increasing personalized monitoring of the e-learning process and adaptation of a platform to a student.

2. Exploratory Data Analysis and Data Collection

This section presents the process of collecting pure data from logs of activity in the platform, aggregating this data by every student and choosing the correct time period in the course to build predictions on.

2.1 Course Material Overview

The study used statistical data accumulated on the national open education platform of the Russian Federation during the online course "Methods and algorithms of graph theory" (https://openedu.ru/course/ITMOUniversity/AGRAPH/) for the period from 2016 to 2018. This online course [6-7] is conducted for 10 weeks twice a year (at the beginning of the fall and spring semesters), contains 41 video lectures with surveys and 11 interactive practical exercises. On the 10th week an online exam is held. Table 1 presents practical exercises presented in the course.

Table 1. Practical exercises of the course

Task number	Typical graph problem	Algorithm	Week number
1	Search shortest route	Lee algorithm	2
2	Search route with minimal weight	Bellman-Ford algorithm	2
3	Search for Hamilton loops	Roberts-Flores algorithm	3
4	Search for minimum spanning tree	Prim algorithm	4
5	Search for minimum spanning tree	Kruskal algorithm	4
6	Search for largest empty subgraphs	Magu-Weismann algorithm	5
7	Minimum vertex coloring of graph	Method based on Magu-Weisman algorithm	6
8	Minimum vertex coloring of graph	Greedy heuristic algorithm	6
9	Search perfect matching in a bipartite graph	Hungarian algorithm	7
10	Detecting of isomorphism of two graphs	Algorithm based on ISD method	8
11	Graph planarization	Gamma-algorithm	9

To select the most appropriate time period to build predictions on, analysis of practical exercises in the middle of the course was performed. Figure 1 presents average maximum time needed to a student to complete a practical exercise. As we can see, Magu-Weismann algorithm is held on 5th week (middle points of the course) and has a bimodal distribution, which means that this task is complicated for some number of students.

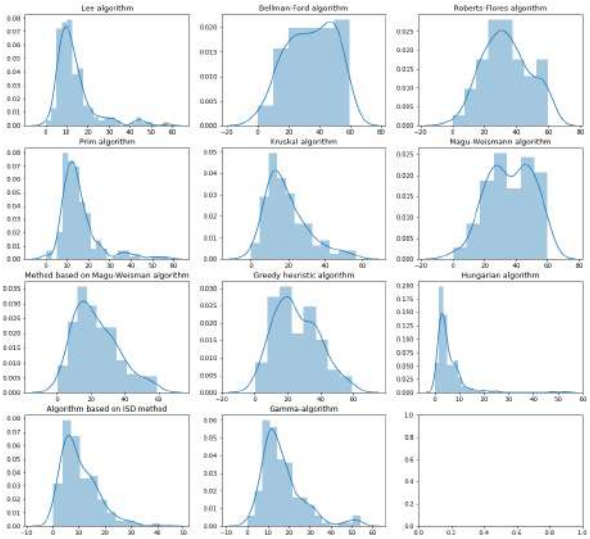


Figure 1. Average maximum time (in minutes) taken to compete exercise

In addition, Figure 2 presents mean amount of tries of students to pass a practical exercise. We can see that task 6 (Magu-Weismann algorithm) has the biggest number of mean amounts of tries. So, the hypothesis that the fact about passing this exercise can be a good feature for a further model and the 5th week is the best time to build predictions on was proposed.

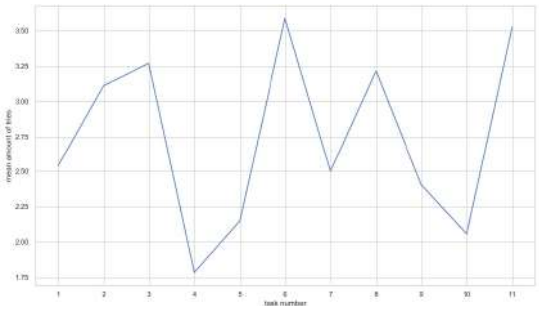


Figure 2. Average number of tries of practical exercises

2.2 Collecting and Analyzing Data-set for Further Predictions

After the point of building predictions was chosen, we

DOI: https://doi.org/10.30564/jcsr.v1i1.537

collected different features of statistical activity of every student of watching lectures, solving practical exercises and quizzes, activity on forum and other. As a result, 50-dimensional feature space was created. As a target, we took a binary feature of a fact of passing a final exam (1 – passed; 0 – not passed). Figure 2 shows to descending trend of overall activity of students for the first 5 weeks.

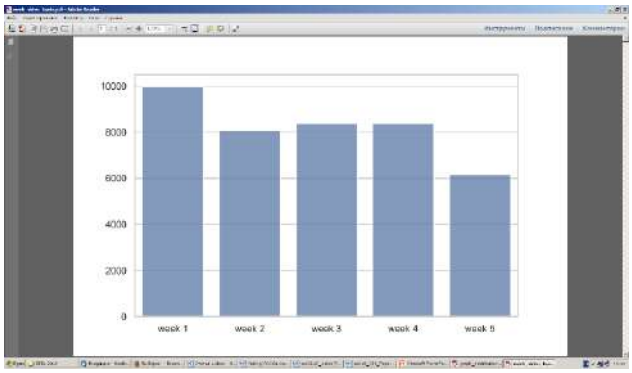


Figure 2. Overall activity of students in the first 5 weeks of the course

To visualize 50-dimensional feature space on a plain, t-SNE algorithm was applied [8, 9] (Figure 3). Students who passed the exam are marked with orange color and other students are marked blue. From the figure, most of the students who passed the exam are grouped in one area in both projections. It means that there is a hyperplane in the original dimension of features that separates the majority of students who pass the exam more likely, so the problem has a solution.

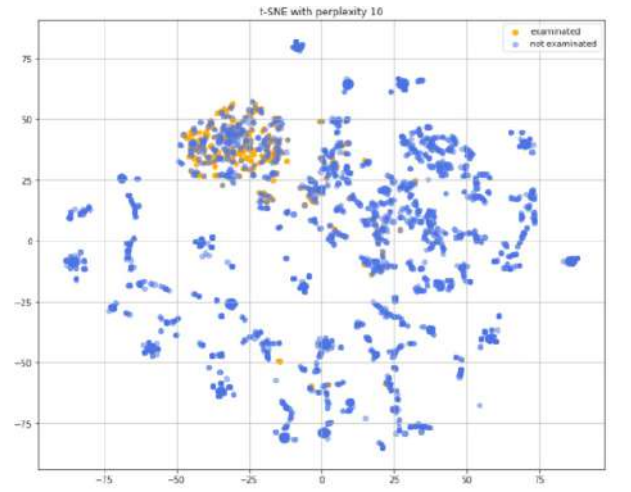


Figure 3. The projection of students on a two-dimensional space using t-SNE algorithm (perplexity equal 50)

Concluding this section, we establish that 5th week is an appropriate time period to build predictions. If we choose a later time, then the number of students that we could try to keep will be less. The choice of this period is confirmed by the Mage-Weisman algorithm, which is the most difficult in the course. Also, the assessment of the solution of the problem is confirmed on the graph of the t-SNE projection.

3. Machine Learning Part

In this section, we formulate churn prediction problem in machine learning terms and build an ensemble model for classification.

3.1 Problem Overview

The purpose of building a model in this task is to rank the participants according to their likelihood of passing the exam. To evaluate the models, the ROC AUC metric [10] was chosen due to the operation of the probabilities of the object belonging to the class with different thresholds. To determine the threshold, the expected number of participants is used based on the historical data of mean number of students who passed the exam in each session. After building a predictive model, participants are ranked according to their likelihood to successfully complete the course. The group of students, which is located below the selected threshold, is a group on which additional effects are required to increase the likelihood of a successful completion of the course, and, accordingly, increase the effectiveness of their learning. The formulation of the problem is a probabilistic binary classification.

3.2 Building Classifier Model

To build a baseline for this classification problem, support vector machine [11], logistic regression[12, 13], random forest [14] and gradient boosting on decision trees (GBDT) [15, 16] were chosen and validated. Due to the small sample size, the evaluation and comparison of the models was implemented through a cross-validation with using different sessions as folds, which allowed to consider the time component in the data. The session, which took place in the Fall of 2018, was chosen as a test set.

Table 2 present the final results of cross-validation for these models of ROC-AUC value and its std. As we can see from the table, GBDT has the best value of the chosen metric. But we have a hypothesis that we can improve out baseline due stacking [17, 18]. For this purpose, we choose one linear model and one tree-based model. We chose logistic regression as a linear model for stacking because SVM has constant probabilities with Radial basis function kernel (RBF) [19], which is not appropriate for ROC-AUC and stacking. SVM with linear kernel is not able to operate with probabilities as other models because it does not apply any sigmoid function. GBDT model was chosen as

 DOI: https://doi.org/10.30564/jcsr.v1i1.537

a tree-based model for further improvement.

Table 2. Results of cross-validation for baseline models

Model	Result of ROC AUC
Logistic Regressor	0.8699 ± 0.0274
Support vector machine	0.8763 ± 0.0258
Random Forest	0.9027 ± 0.0353
Gradient boosting on trees	0.9153 ± 0.0312

After the chosen two models were fitted, they were compared due the chosen metric and the similarity of their prediction was analyzed. The results of comparison models using ROC-AUC is introduced in the table of final scores below. The graph of the similarity of predictions of gradient boosting and logistic regression models on a separate split of cross-validation is shown in Figure 4. Each point of the plot is a separate student. Orange points are the students who passed the exam and blue points are the others. On x and y axis the predicted probability of passing the exam by logistic regression and gradient boosting on decision trees models are shown respectively. If these two models have similar predictions, the points should be located on the blue line. But, from the figure, there are many points far from the line. Therefore, a hypothesis about improving the value of the ROC-AUC metric by using ensembling of the initial models via stacking was approved.

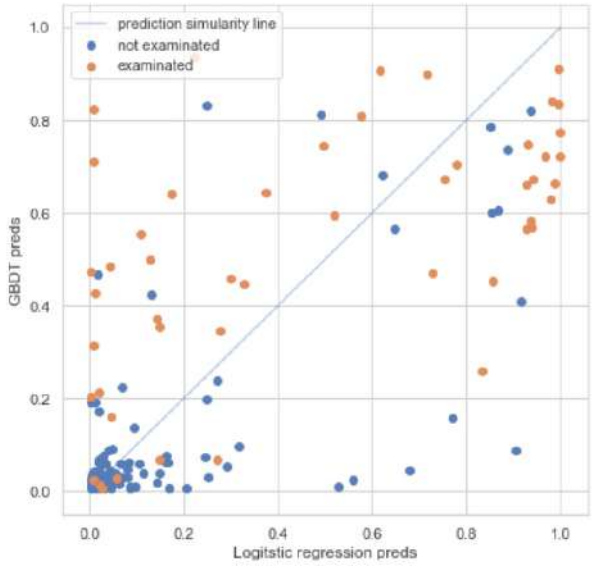


Figure 4. The similarity of the predictions of different models for students of the 3rd session from the beginning of the course via cross-validation

Stacking was applied using another logistic regression model to build new predictions on predictions of the initial models. The same session as in the figure 2 was used as a validation set for stacking due to the large variance of the predictions between two initial models.

Table 3 shows the results of the ROC-AUC score of all the applied models after the final cross-validation run. Table shows that the model of gradient boosting on trees is always preferable to logistic regression. Also, the ensemble of models in most cases shows a better quality than the gradient boosting model. From the mean results of cross-validation, the ensemble of the models gives a significant increase of ROC-AUC value. The possible reason for this is that logistic regression gives a better score using some linear dependencies in the dataset and gradient boosting is better in more complicated cases. In further calculations we will use stacking as a final model.

Table 3. Results of cross-validation

Split	Model	Result of ROC AUC
1	Logistic Regressor	0.9255
	Gradient Boosting on decision trees	0.9546
	Stacking	0.9767
2	Logistic Regressor	0.8702
	Gradient Boosting on decision trees	0.9302
	Stacking	0.9688
4	Logistic Regressor	0.9116
	Gradient Boosting on decision trees	0.9780
	Stacking	0.9742
5	Logistic Regressor	0.7659
	Gradient Boosting on decision trees	0.9117
	Stacking	0.8876
6	Logistic Regressor	0.8651
	Gradient Boosting on decision trees	0.8925
	Stacking	0.9160
Logistic regressor mean and std		0.8612 ± 0.0531
Gradient Boosting on decision trees mean and std		0.9189 ± 0.0427
Stacking mean and std		0.9304 ± 0.0459

3.3 Analysis of the Final Model

After building the final predictions, the model was analyzed. Tables 4, 5, 6 respectively show the 5 most significant features that were used by a separate model in the ensemble to build the final predictions. Table 7 provides a description of these features. The extended description of tasks is introduced in article [1]. It can be seen that different features are used by different models. As an example, for linear model an overall activity in solving interactive tasks in the course is important, but gradient boosting uses many features based on a separate week of the course.

Table 4. Feature importance for GBDT with small depth

Feature	Importance in %
Activity_sum	18.3777
Week_2_activity	10.4238
Task_1_amount_of_tries	5.4629
Task_5_amount_of_tries	5.2787
Mean_attempts	5.1692

DOI: https://doi.org/10.30564/jcsr.v1i1.537

Table 5. Feature importance for GBDT with big depth

Feature	Importance in %
Problems_solved	12.9358
Week_1_activity	11.4598
Week_2_activity	11.2761
Task_5_amount_of_tries	6.6758
Week_5_video_loads	4.0969

Table 6. Feature importance for logistic regressor

Feature	importance in %
Mean_attempts	20.1610
Task_3_amount_of_tries	11.9389
Grade_mean_rate	11.8862
Task_0_amount_of_tries	6.3207
Task_5_amount_of_tries	4.6111

Table 7. Feature meaning

Feature	Meaning
Task_1_amount_of_tries	Number of attempts of a student of solving a task with Lee algorithm
Task_2_amount_of_tries	Number of attempts of a student of solving a task with Bellman-Ford algorithm
Task_4_amount_of_tries	Number of attempts of a student of solving a task with Prim algorithm
Task_6_amount_of_tries	Number of attempts of a student of solving a task with Magu-Weismann algorithm
Mean_attempts	Mean number of attempts of a student during solving an interactive task
Problems_solved	Total number of polls solved by a student
Grade_mean_rate	Rate of the correct answers of a student
Week_1_activity	Overall activity of a student in the first week
Week_2_activity	Overall activity of a student in the second week
Week_5_video_loads	Number of viewed video by a student in the fifth week
Activity_sum	Overall activity of a student in the first 5 weeks of the course

To obtain the results of the final feature importance, the values of the calculated feature importance of each separate model were multiplied by the corresponding meta-model coefficients (0.612 for GBDT and 0.388 for logistic regression). The final feature importance is shown in Table 7. From the table it concluded that the most important features for the final model are features based on activity of students during separated weeks of the course and their overall amounts of attempts in interactive tasks and solving quizzes.

Table 8. Final feature importance

Feature	importance in %
Week_2_activity	8.2561
Mean_attempts	7.8725
Activity_sum	7.6408
Week_1_activity	5.8877
Grade_mean_rate	5.7178

4. Results and Discussions

To select a correct threshold value, we take the percentage of students who passed the exam in previous sessions (5.6%) multiplied by the number of students in the current session. Table 9 presents the results of ranking students in the test set on their likelihood to complete the course, starting with the highest probability. After applying calculated threshold, we get a list of students who need to have an additional impact to increase the effectiveness of their learning (Table 10). The last column in the tables shows whether the participant has actually passed the exam (1 for yes, 0 for no). The resulting tables show that the model correctly ranks the students of the course according to their likelihood to pass the exam in general. The resulting tables can be used to further impact a particular group of students of the course.

Table 9. Students with highest probability of examination

User	Probability of examination	Examinated
Student 11	0.8661	1
Student 12	0.8616	1
Student 13	0.8542	1
Student 14	0.8221	0
Student 15	0.8217	0
Student 16	0.8162	1
Student 17	0.8038	1
Student 18	0.7765	0
Student 19	0.7719	1
Student 110	0.7666	0

Table 10. Students below the threshold of examinations

User	Probability of examination	Examinated
Student 21	0.4741	0
Student 22	0.4453	0
Student 23	0.4352	0
Student 24	0.4232	0
Student 25	0.4216	1
Student 26	0.4163	0
Student 27	0.4015	0
Student 28	0.3793	1
Student 29	0.3771	0
Student 210	0.3348	1

 DOI: https://doi.org/10.30564/jcsr.v1i1.537

5. Conclusion

Accumulated statistics on the activity of MOOC's students allow to predict their future behavior and learning outcomes. This paper overviews all the process of building such a model: from EDA of course materials to constructing a strong classifier to predict a fact of passing an exam by a student using his activity in the first half of the course. To solve this problem, various machine learning approaches and models have been proposed. According to the results, the most significant features were obtained for assessing the fact that the exam was passed by the students. As a result of model's prediction, a list of participants was received. This approach can be used as to increase the efficiency of learning of separated students and to improve course materials in general. Also, this problem can be interpreted as a churn prediction problem. After the final list of students is received, it can be used to make the course more personal for this group of students. As an example, we suggest giving some hints and additional bonuses for the student if he will continue learning or increasing deadlines. Results of the final model analysis can be used for exploring aspects of the course that are important for a separate group of students. Thus, this article proposes a general approach for assessing and identifying MOOC students during the course, on which additional impact is required to improve the performance of e-learning using MOOC. Using this approach in MOOCs can increase effectiveness of online courses and make e-learning more self-organized and adaptive for a separate student.

References

[1] Lisitsyna L.S., Efimchik E.A., "Making MOOCs more effective and adaptive on the basis of SAT and game mechanics", Smart Education and e-Learning, 2018, 75, 56-66.

[2] Liubov S. Lisitsyna, Andrey V. Lyamin, Ivan A. Martynikhin, Elena N. Cherepovskaya, "Situation Awareness Training in E-Learning", Smart Education and Smart e-Learning, 2015, 41, 273-285.

[3] Liubov S. Lisitsyna, Andrey V. Lyamin, Ivan A. Martynikhin, Elena N. Cherepovskaya, "Cognitive Trainings Can Improve Intercommunication with e-Learning System", 6th IEEE international conference series on Cognitive Infocommunications, 2015, 39-44.

[4] Lisitsyna L.S., Pershin A.A., Kazakov M.A., "Game Mechanics Used for Achieving Better Results of Massive Online", Smart Education and Smart e-Learning, 2015, 183-193.

[5] A. Pradeep, S. Das, J. J. Kizhekkethottam, A. Cheah et al., "Students dropout factor prediction using EDM techniques", Proc. IEEE Int. Conf. Soft-Computing Netw. Secur. ICSNS 2015, pp. 544-547, 2015.

[6] Lisitsyna L.S., Efimchik E.A., "Designing and application of MOOC «Methods and algorithms of graph theory» on National Platform of Open Education of Russian Federation", Smart Education and e-Learning, 2016, 59, 145-154.

[7] Lisitsyna L.S., Efimchik E.A., "An Approach to Development of Practical Exercises of MOOCs based on Standard Design Forms and Technologies", Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 2017, 180, 28-35.

[8] Laurens van der Maaten, Geoffrey Hinton, "Visualizing Data using t-SNE", Journal of Machine Learning Research, 2008, 9, 2579-2605.

[9] Laurens van der Maaten, "Accelerating t-SNE using Tree-Based Algorithms", Journal of Machine Learning Research, 2014, 15, 3221-3245.

[10] Tom Fawcett, An introduction to ROC analysis, In: Pattern Recognition Letters, 27 (2006), 861–874.

[11] D. Michie, D. J. Spiegelhalter, C. C. Taylor, and J. Campbell, editors. Machine learning, neural and statistical classification. Ellis Horwood, Upper Saddle River, NJ, USA, 1994. ISBN 0-13-106360-X.

[12] Chao-Ying Joanne, Peng Kuk Lida Lee, Gary M. Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting", The Journal of Educational Research, 2002, 96, 3-14.

[13] Marijana Zekić-Sušac, Nataša Šarlija, Adela Has, Ana Bilandžić, "Predicting company growth using logistic regression and neural networks", Croatian Operational Research Review, 2016, 7, 229–248.

[14] Boulesteix, A.-L., Janitza, S., Kruppa, J. and König, I.R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 2, no. 6, pp. 493–507.

[15] Jerome H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, In: The Annals of Statistics, 2001, 5, 1189-1232.

[16] Friedman, Jerome H., "Stochastic gradient boosting", Computational Statistics and Data Analysis, 2002, 38(4), 367–378.

[17] Smyth, P. and Wolpert, D. H., "Linearly Combining Density Estimators via stacking", Machine Learning Journal, 1999, 36, 59-83.

[18] David Opitz, Richard Maclin, Popular Ensemble Methods: An Empirical Study, Journal of Artificial Intelligence Research, 1999, 11, 169-198.

DOI: https://doi.org/10.30564/jcsr.v1i1.537

[19] Chang, Yin-Wen; Hsieh, Cho-Jui; Chang, Kai-Wei; Ringgaard, Michael; Lin, Chih-Jen (2010). "Training and testing low-degree polynomial data mappings via linear SVM". Journal of Machine Learning Research. 11: 1471–1490.

 DOI: https://doi.org/10.30564/jcsr.v1i1.537

Author Guidelines

This document provides some guidelines to authors for submission in order to work towards a seamless submission process. While complete adherence to the following guidelines is not enforced, authors should note that following through with the guidelines will be helpful in expediting the copyediting and proofreading processes, and allow for improved readability during the review process.

Ⅰ. Format

- Program: Microsoft Word (preferred)
- Font: Times New Roman
- Size: 12
- Style: Normal
- Paragraph: Justified
- Required Documents

Ⅱ. Cover Letter

All articles should include a cover letter as a separate document.

The cover letter should include:

- Names and affiliation of author(s)

The corresponding author should be identified.

Eg. Department, University, Province/City/State, Postal Code, Country

- A brief description of the novelty and importance of the findings detailed in the paper

Declaration

v Conflict of Interest

Examples of conflicts of interest include (but are not limited to):

- Research grants
- Honoria
- Employment or consultation
- Project sponsors
- Author’s position on advisory boards or board of directors/management relationships
- Multiple affiliation
- Other financial relationships/support
- Informed Consent

This section confirms that written consent was obtained from all participants prior to the study.

- Ethical Approval

Eg. The paper received the ethical approval of XXX Ethics Committee.

- Trial Registration

Eg. Name of Trial Registry: Trial Registration Number

- Contributorship

The role(s) that each author undertook should be reflected in this section. This section affirms that each credited author has had a significant contribution to the article.

1. Main Manuscript
2. Reference List
3. Supplementary Data/Information

Supplementary figures, small tables, text etc.

As supplementary data/information is not copyedited/proofread, kindly ensure that the section is free from errors, and is presented clearly.

Ⅲ. Abstract

A general introduction to the research topic of the paper should be provided, along with a brief summary of its main results and implications. Kindly ensure the abstract is self-contained and remains readable to a wider audience. The abstract should also be kept to a maximum of 200 words.

Authors should also include 5-8 keywords after the abstract, separated by a semi-colon, avoiding the words already used in the title of the article.

Abstract and keywords should be reflected as font size 14.

Ⅳ. Title

The title should not exceed 50 words. Authors are encouraged to keep their titles succinct and relevant.

Titles should be reflected as font size 26, and in bold type.

Ⅳ. Section Headings

Section headings, sub-headings, and sub-subheadings should be differentiated by font size.

Section Headings: Font size 22, bold type

Sub-Headings: Font size 16, bold type

Sub-Subheadings: Font size 14, bold type

Main Manuscript Outline

Ⅴ. Introduction

The introduction should highlight the significance of the research conducted, in particular, in relation to current state of research in the field. A clear research objective should be conveyed within a single sentence.

Ⅵ. Methodology/Methods

In this section, the methods used to obtain the results in the paper should be clearly elucidated. This allows readers to be able to replicate the study in the future. Authors should ensure that any references made to other research or experiments should be clearly cited.

Ⅶ. Results

In this section, the results of experiments conducted should be detailed. The results should not be discussed at length in

this section. Alternatively, Results and Discussion can also be combined to a single section.

Ⅷ. Discussion

In this section, the results of the experiments conducted can be discussed in detail. Authors should discuss the direct and indirect implications of their findings, and also discuss if the results obtain reflect the current state of research in the field. Applications for the research should be discussed in this section. Suggestions for future research can also be discussed in this section.

Ⅸ. Conclusion

This section offers closure for the paper. An effective conclusion will need to sum up the principal findings of the papers, and its implications for further research.

Ⅹ. References

References should be included as a separate page from the main manuscript. For parts of the manuscript that have referenced a particular source, a superscript (ie. [x]) should be included next to the referenced text.

[x] refers to the allocated number of the source under the Reference List (eg. [1], [2], [3])

In the References section, the corresponding source should be referenced as:

[x] Author(s). Article Title [Publication Type]. Journal Name, Vol. No., Issue No.: Page numbers. (DOI number)

Ⅺ. Glossary of Publication Type

J = Journal/Magazine

M = Monograph/Book

C = (Article) Collection

D = Dissertation/Thesis

P = Patent

S = Standards

N = Newspapers

R = Reports

Kindly note that the order of appearance of the referenced source should follow its order of appearance in the main manuscript.

Graphs, Figures, Tables, and Equations

Graphs, figures and tables should be labelled closely below it and aligned to the center. Each data presentation type should be labelled as Graph, Figure, or Table, and its sequence should be in running order, separate from each other.

Equations should be aligned to the left, and numbered with in running order with its number in parenthesis (aligned right).

Ⅻ. Others

Conflicts of interest, acknowledgements, and publication ethics should also be declared in the final version of the manuscript. Instructions have been provided as its counterpart under Cover Letter.

About the Publisher

Bilingual Publishing Co. (BPC) is an international publisher of online, open access and scholarly peer-reviewed journals covering a wide range of academic disciplines including science, technology, medicine, engineering,education and social science. Reflecting the latest research from a broad sweep of subjects, our content is accessible worldwide – both in print and online.

BPC aims to provide an analytics as well as platform for information exchange and discussion that help organizations and professionals in advancing society for the betterment of mankind. BPC hopes to be indexed by well-known databases in order to expand its reach to the science community, and eventually grow to be a reputable publisher recognized by scholars and researchers around the world.

BPC adopts the Open Journal Systems, see on http://ojs.s-p.sg

Database Inclusion

National Library, Singapore

Asia & Pacific area Science Citation Index

China National Knowledge Infrastructure

Creative Commons

Google Scholar

Crossref

J-Gate

My Science Work

Bilingual Publishing Co. is a company registered in Singapore in 1984, whose office is at 12 Eu Tong Sen Street, #08-169, Singapore 059819, enjoying a high reputation in Southeast Asian countries, even around the world.

BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Tel: +65 65881289
E-mail: contact@blipublishing.Com
Website: www.Blipublishing.Com

ISSN 2630-5151 01
9 772630 515193