UDC 004.891.032.26:629.7.01.066

A MODEL AND TRAINING METHOD FOR CONTEXT CLASSIFICATION IN CCTV SEWER INSPECTION VIDEO FRAMES

Moskalenko V. V. – PhD, Associate Professor of Computer Science department, Sumy State University, Sumy, Ukraine.

Zaretsky M. O. – Postgraduate Student of Computer Science department, Sumy State University, Sumy, Ukraine.

Moskalenko A. S. – PhD, Senior Lecturer of Computer Science department, Sumy State University, Sumy, Ukraine.

Panych A. O. – M. Eng., Lecturer of Computer Science department, Sumy State University, Sumy, Ukraine.

Lysyuk V. V. – M. Eng., Co-Founder of Molfar.AI, Molfar.AI sp. z o.o., Gdansk, Poland.

ABSTRACT

Context. A model and training method for observational context classification in CCTV sewer inspection vide frames was developed and researched. The object of research is the process of detection of temporal-spatial context during CCTV sewer inspections. The subjects of the research are machine learning model and training method for classification analysis of CCTV video sequences under the limited and imbalanced training dataset constraint.

Objective. Stated research goal is to develop an efficient context classifier model and training algorithm for CCTV sewer inspection video frames under the constraint of the limited and imbalanced labeled training set.

Methods. The four-stage training algorithm of the classifier is proposed. The first stage involves training with soft triplet loss and regularisation component which penalises the network's binary output code rounding error. The next stage is needed to determine the binary code for each class according to the principles of error-correcting output codes with accounting for intra- and interclass relationship. The resulting reference vector for each class is then used as a sample label for the future training with Joint Binary Cross Entropy Loss. The last machine learning stage is related to decision rule parameter optimization according to the information criteria to determine the boundaries of deviation of binary representation of observations for each class from the corresponding reference vector. A 2D convolutional frame feature extractor combined with the temporal network for inter-frame dependency analysis is considered. Variants with 1D Dilated Regular Convolutional Network, 1D Dilated Causal Convolutional Network, LSTM Network, GRU Network are considered. Model efficiency comparison is made on the basis of micro averaged F1 score calculated on the test dataset.

Results. Results obtained on the dataset provided by Ace Pipe Cleaning, Inc confirm the suitability of the model and method for practical use, the resulting accuracy equals 92%. Comparison of the training outcome with the proposed method against the conventional methods indicated a 4% advantage in micro averaged F1 score. Further analysis of the confusion matrix had shown that the most significant increase in accuracy in comparison with the conventional methods is achieved for complex classes which combine both camera orientation and the sewer pipe construction features.

Conclusions. The scientific novelty of the work lies in the new models and methods of classification analysis of the temporal-spatial context when automating CCTV sewer inspections under imbalanced and limited training dataset conditions. Training results obtained with the proposed method were compared with the results obtained with the conventional method. The proposed method showed 4% advantage in micro averaged F1 score.

It had been empirically proven that the use of the regular convolutional temporal network architecture is the most efficient in utilizing inter-frame dependencies. Resulting accuracy is suitable for practical use, as the additional error correction can be made by using the odometer data.

KEYWORDS: Sewer pipe inspection, convolutional neural network, error-correction output codes, Siamese network, Information-Extreme Learning, information criterion, LSTM, GRU.

ABBREVIATIONS

BB is a Building Block;
CNN is a Convolutional Neural Network;
GRU is a Gated Recurrent Unit;
LSTM is a Long Short-Term Memory;
MSCC is a Manual of Sewer Condition Classification;
PACP is a Pipeline Assessment Certification Program;
SLAM is a Simultaneous Localization and Mapping;
TCN is a Temporal Convolutional Network;
1D is a One-Dimensional space;
2D is a Two-Dimensional space.

NOMENCLATURE

D_v^{train} is the labeled frame sequences for training;

D_w^{test} is the labeled frame sequences for testing;

$I_{v,k}$ is a v-th set of ordered video frame sequences for training;

$L_{v,k}$ is a v-th set of ordered labels of video frame sequences for training;

V is a number of labeled frame sequences for training;

W is a number of labeled frame sequences for testing;

K_v is a size of v-th set for training;

K_w is a size of w-th set for testing;

N is a size of high-level feature set;

Z is a size of set of classes;

© Moskalenko V. V., Moskalenko A. S., Zaretsky M. O., Panych A. O., Lysyuk V. V., 2021
DOI 10.15588/1607-3274-2021-3-9

e_{ξ_1} is a ξ_1-th parameter which impacts on feature representation, $\xi_1 = \overline{1,\Xi_1}$;

f_{ξ_2} is a ξ_2-th parameter which impacts on efficiency of decision rules, $\xi_2 = \overline{1,\Xi_2}$;

TP_z is a numbers of true positives for decision rule of z-th class;

FP_z is a numbers of false positives for decision rule of z-th class;

FN_z is a numbers of false negatives for decision rule of z-th class;

b_z^* is a binary reference vector (center of optimal container) for class X_z^o;

d_z^* is a radius of optimal container for class X_z^o in Hamming distance units;

f(x) is a function describing the feature extractor;

x_a is a image randomly selected from the mini-batch;

x_{ep} is a nearest neighbour in the minibatch belonging to the same class;

C(x) is a function returning the image class;

x_{shn} is a sample image from the mini-batch which is the closest among the samples of opposite classes, but located further than hard negative sample;

e is a single column matrix, $e = [1, 1, ..., 1]^T$;

λ is a regularization coefficient;

$f_i(x)$ is a the value of sigmoid layer output i for input image;

$b_{z,i}$ is a value of i-th bit of the reference vector of the z-th class to which the image x belongs;

E_z is a Z-class information criteria, a function of the accuracy characteristics;

$\alpha_z^{(k)}$ is a false postive rate on k-th traning step for z-th class;

$\beta_z^{(k)}$ is a false negative rate on k-th traning step for z-th class;

$D_{1,z}^{(k)}$ is a true positive rate or sensitivity on k-th traning step for z-th class;

$D_{2,z}^{(k)}$ is a true negative rate or specificity on k-th traning step for z-th class;

d is a distance measure defining the hyperspherical container radii in the radial basis of Hamming space;

$\{d\}$ is a set of concentric radii of data distribution in class z with the centre b_z.

INTRODUCTION

Sewer pipes are critical infrastructure items which require frequent monitoring. The most widely used method for analysis of sewer pipe conditions involves the CCTV inspection of the pipes view to identify the defects and faults inside the pipe. Each of the detected defects and faults is assigned a standardised code in accordance with the applicable local standards, among which the most common are the British MSCC5 and American PACP6 or PACP7 [1].

The preparation of a report on the condition of inspected sewer pipes in accordance with the standards requires careful examination and detailed analysis of the collected CCTV inspection videos. The use of computer vision and machine learning techniques for CCTV inspection footage analysis can increase productivity and reduce costs [2].

To achieve the correct defect coding it is necessary to have information about the location, orientation, shape, severity and proximity of the defect to the upstream and downstream manholes and sewer line branches (laterals/service connections/taps). In turn, contextual data on the orientation and relative location of the inspection camera in the pipe is needed to extract such information. Such data, however, as a rule is not available in the explicit form. This makes observation context recognition a relevant task.

Orientation and relative position of the camera can be determined with the help of visual odometry or simultaneous localization and mapping (SLAM) methods [3]. However, CCTV inspection videos generally already contain superimposed distance readings measured by a mechanical odometer. Correct defect coding also does not require high degree of precision in respect of the camera optical axis angle and relative position in relation to the to the center of the pipe. Employing a computationally efficient frame sequence classifier to estimate the camera orientation instead of computationally complex SLAM algorithms is therefore more appropriate. In this case alphabet of classes can be expanded to include various non-standard situations which need to be processed correctly.

A modern approach to the classification analysis of the sequence of video frames involves the use of deep neural networks. Important steps in the classification analysis of individual video frames or their sequences are the feature descriptions of both individual frames and the relationships between them, as it has a direct bearing on the effectiveness of the resulting decision rules.

Convolutional neural networks are still the most effective approach to image feature description at present. Where the analysis of time series is concerned, the undisputed leaders are recurrent and temporal convolutional networks, dilated versions of which provide a speed advantage without information loss [4]. The end results, however, depend not only on the architecture of the model, but also on the methods of machine learning and regularization employed. This is especially true where the labelled training dataset is small and variability of observations is high. The study of the effectiveness of different model architectures and training methods for specific applications thus remains a relevant task, as no single universal approach to data analysis exists and steady emergence of new research continually changes

© Moskalenko V. V., Moskalenko A. S., Zaretsky M. O., Panych A. O., Lysyuk V. V., 2021
DOI 10.15588/1607-3274-2021-3-9

the paradigm, forcing a rethink of the current body of knowledge and highlight new directions for further research.

The research goal is development of an effective deep learning model and its training method for recognizing the context of observations during CCTV sewer inspection. **The object of research** is the process of detection of temporal-spatial context during CCTV sewer inspections. **The subjects of the research** are machine learning model and training method for classification analysis of CCTV video sequences under the limited and imbalanced training dataset constraint.

1 PROBLEM STATEMENT

Let V sequences $D_v^{train} = \{ I_{v,k}, L_{v,k} \mid v = \overline{1,V};\ k = \overline{1,K_v}\}$ and W sequences $D_w^{test} = \{ I_{w,k}, L_{w,k} \mid w = \overline{1,W};\ k = \overline{1,K_w}\}$ are collected of labeled video frames for training and testing, respectively. Let the set $\{X_z^o \mid z = \overline{1,Z}\}$ is characterized by observation context in pipe, be given. In this case, the dataset is unbalanced, the minority class can contain twice as many samples as the majority.

Moreover, the structure of the vector of model parameters is known

$$g = < e_1,...,e_{\xi_1},...,e_{\Xi_1},\ f_1,...,f_{\xi_2},...,f_{\Xi_2} >, \quad (1)$$
$$\Xi_1 + \Xi_2 = \Xi .$$

In this case, the constraints $R_{\xi_1}(e_1,...,e_{\xi_1},...,e_{\Xi_1}) \le 0$, $R_{\xi_2}(f_1,...,f_{\xi_2},...,f_{\Xi_2}) \le 0$ are impose on parameters.

It is necessary to find by machine learning an optimal values of parameters g (1) which provide to achieve the maximum value of micro averaged F1 score for context classifier

$$F1 = \frac{2\sum_{z}^{Z} TP_z}{2\sum_{z}^{Z} TP_z + \sum_{z}^{Z} FP_z + \sum_{z}^{Z} FN_z}. \quad (2)$$

$$g^* = \arg\max_{G} \{F1(g)\}. \quad (3)$$

When the model functions in its inference mode, it is necessary to provide high confidence of classification of frame context on test images.

2 REVIEW OF THE LITERATURE

Early algorithms for CCTV sewer inspection video frames classification analysis employed edge and contour detection methods for feature description [5]. Such an approach, however, ignores a large amount of contextual information and necessitates particular attention to the design of a post-processing algorithm. An algorithm of this kind would require a large number of handcrafted parameters and conditions, which can lead to incompleteness of decision rules or contradictions between them due to the human error. Gabor filters offer a more flexible and theoretically sound approach to visual feature extraction [6]. However, models of this type are characterized by insufficient information capacity for computationally efficient description of contexts under conditions of complex defect and/or design features combinations.

Much progress in the field of visual data analysis had been achieved within the framework of the deep machine learning, based on hierarchical feature description. A distinguishing feature of hierarchical feature extractors is their higher information capacity in comparison to models with one hidden layer [7, 8]. At present deep convolutional neural networks are still considered the most effective for image feature description [8]. In the field of time-series analysis, the leading positions are occupied by recurrent and temporal convolutional networks. Their dilated versions provide a speed advantage without loss of information [4] in both customary, centered, and causal model output variants. However, some observational contexts or their parts are rare and can have significant intra-class variability, leading to imbalances and a scarcity of labeled samples corresponding to complex and irregular situations. This imposes limitations on the use of deep models sensitive to the volume and balance of labelled training data.

One of the ways to increase the efficiency of models with a limited amount of marked data is to use the ideas and methods of information theory and synthesis of decision rules within a geometric approach. An example of geometric approach methods are Siamese neural networks, where the fitness function makes use of constraints and relationships between the distances between samples of the same and different classes [9]. Siamese networks have shown the greatest efficiency in few-shot learning and meta-learning algorithms, but they are most commonly used for feature embedding.

Information theory and coding methods are a natural choice for increasing the resistance to noise, such as artifacts and limited visibility. For example, error-correcting output codes implement end-to-end pseudo-ensemble, increase the multi-class classification accuracy and Improve Probability Estimation for Adversarial Robustness of Deep Neural Networks [10].

However, the existing methods of binary class code selection does not take into account the internal structure of classes. Information-extreme machine learning methods provide optimization in the information sense of the decision rule parameters based on error-correcting output codes [11]. However, information-extreme learning does not provide end to end deep model learning mechanisms. Thus, the combination of ideas and methods of Siamese neural networks and information-extreme learning offers considerable promise for further improvement of data analysis models, in particular for context analysis in CCTV sewer inspections.

© Moskalenko V. V., Moskalenko A. S., Zaretsky M. O., Panych A. O., Lysyuk V. V., 2021
DOI 10.15588/1607-3274-2021-3-9

3 MATERIALS AND METHODS

Classification analysis of CCTV video frames in the simplest case can be performed by a single convolutional network. Such network can be trained both in the traditional way and as a part of Siamese or generative models. However, in the situations of loss of visibility or significantly close proximity of the camera to the pipe walls, images lose the large part of their useful context information. This necessitates the analysis of each frame in conjunction with the neighbouring frames to restore the context information. In general case context detection model will have the 2D convolutional neural network at the lower level for spatial feature extraction and 1D temporal network for analysis of cross-frame dependencies (Fig. 1).

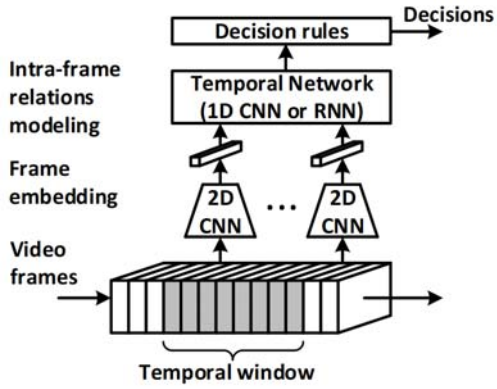

Figure 1 – Generalized architecture of context classifier model

For separate frame analysis the use of MobileNet general purpose convolutional network is proposed [9]. Only a convolutional backbone without fully connected layers is used [9]. Fig 2 a depicts the classic convolutional network variant and Fig 2b its modification used for research of the proposed training method.

Global Average Pooling is used for dimensionality reduction and a Dropout pseudo-ensemble with 50% of the input features dropping is used for regularization [5, 9]. Fully connected and sigmoid layers form the output feature set.

Image classifier model's decision rules contain the rounding layer which produces the binary coded representation and radial-basis function defining the object's belonging to a certain class, with classes separated by hyper-spherical containers in binary Hamming space. Each hyper-spherical container is defined by the binary reference vector (container center) and container radius in Hamming distance units. In this case radial-basis membership function $\mu_z(b)$ for N-dimensional binary vector b is

$$\mu_z(b) = 1 - \sum_{i=1}^{N} b_i \oplus b^*_{z,i} / d^*_z, \qquad (4)$$

where b^*_z – binary reference vector (center of optimal container) for class X^o_z; d^*_z – radius of optimal container for class X^o_z in Hamming distance units.

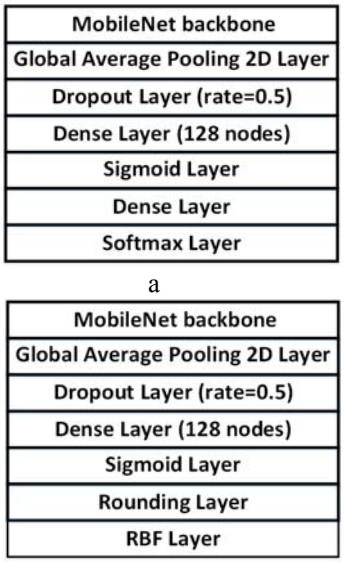

Figure 2 – Architecture of classic and modified variants of the convolutional network for frame-by-frame classification : a – baseline image classifier model structure; b – proposed image classifier model structure

Temporal network can be implemented using one of the popular model architectures, such as 1D dialated non-causal convolutional network, 1D dialated causal convolutional network, Recurrent neural networks with Long Short-Term Memory (LSTM), Recurrent neural networks with Gated Recurrent Units (GPU).

The Base Block (BB) convolutional temporal networks is depicted in Fig. 3 The first layer of the BB is a 1D dilated convolution with kernel size k = 3 where the dilation factor is doubled for each subsequent BB, i.e. 1, 2, 4, 8.

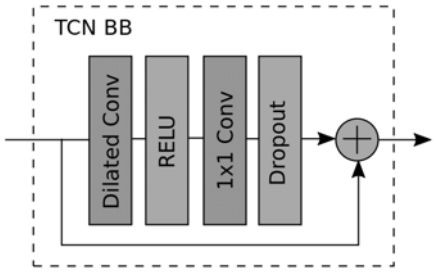

Figure 3 – Basic block of Temporal Convolutional Network

Fig. 4a gives an illustration of the receptive field (black arrows) of an output activation from a single stage with three stacked BBs with regular convolutions. Regular convolutions have a receptive field that expands equally wide to the right as it does to the left. This means that it looks as far into the future (right) as it looks into the past (left). Thus the current frame context will be clarified using the information in both past and future frames.

© Moskalenko V. V., Moskalenko A. S., Zaretsky M. O., Panych A. O., Lysyuk V. V., 2021
DOI 10.15588/1607-3274-2021-3-9

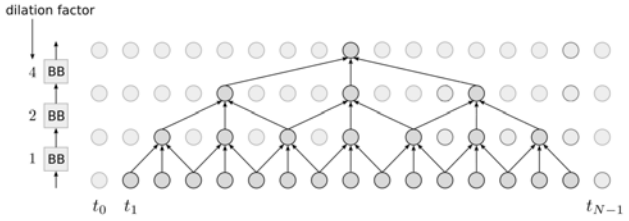

Figure 4 – Temporal Network with Non-causal (regular) convolutions

Fig. 5 illustrates temporal neural network with casual convolutions which amplify the forecast productivity nearer to the right edge. In this case, the curt frame context will be completely defined on the basis of the characteristics and interrelationships of the preceding frames.

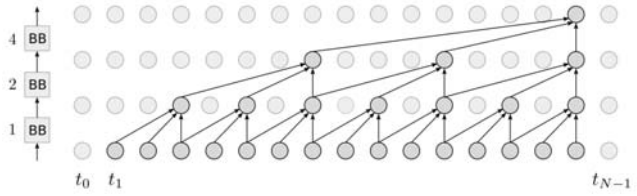

Figure 5 – Temporal Network with causal convolutions

LSTM has memory blocks connected by successive layers, and it enables the network to selectively memorize the input training data through a unique three-gate structure (Fig. 6a). The line across the top of the diagram is the cell state c, and represents the internal memory of the unit. The line across the bottom is the hidden state h, and the i, f, o, and g gates are the mechanism by which the LSTM works around the vanishing gradient problem. During training, the LSTM learns the parameters for these gates. Instead of the input, forget, and output gates in the LSTM cell, the GRU cell has two gates, an update gate z, and a reset gate r (Fig. 6b). The update gate defines how much previous memory to keep around and the reset gate defines how to combine the new input with the previous memory. There is no persistent cell state distinct from the hidden state as in LSTM.

To compare and trace the changes in productivity as a function of the proposed solutions, training will be performed in stages. First the single-frame detection model will be trained in a traditional manner and with the proposed training method without taking the neighbouring frames into account. Then the best trained model is chosen and its feature extractor, the layers located up to and including the sigmoid layer, is used for frame embedding into the temporal detection model. Every neural network type will be trained with both traditional and the proposed training method. The capacity hyperparameter, responsible for the network size will be grid-optimized for each model.

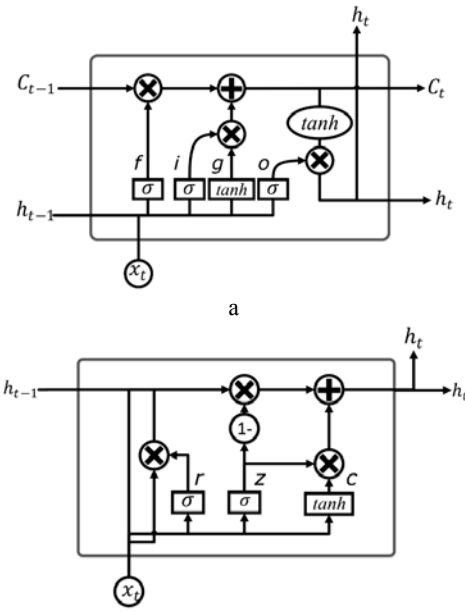

Figure 6 – Cell of Recurrent Neural Network: a – LSTM cell; b – GRU cell

Traditional training method involves an addition of Dense layer with Softmax output normalization, error backpropagation and cross-entropy loss function, such as Adam, to the feature extractor.

The modified method consists of 4 stages necessary to create a binary feature description used to form the information-extreme decision rules (Fig. 7)

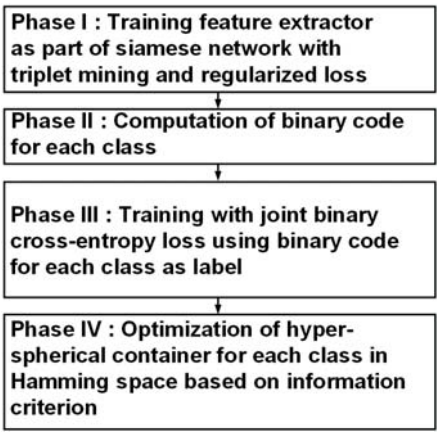

Figure 7 – Proposed training method stages

The first stage involves training with soft triplet loss and regularization component for penalising the binary code output rounding error. The mode input receives a mini-batch with M images of every class. The loss function is calculated as

$$L = -\log \frac{\exp(\left\| f(x_a) - f(x_{ep}) \right\|)}{\exp(\left\| f(x_a) - f(x_{ep}) \right\|) + \exp(\left\| f(x_a) - f(x_{shn}) \right\|)} + \\ + \lambda (f(x_a)^T (e - f(x_a)) + f(x_{ep})^T (e - f(x_{ep})) + \\ + f(x_{shn})^T (e - f(x_{shn}))), \quad (5)$$

© Moskalenko V. V., Moskalenko A. S., Zaretsky M. O., Panych A. O., Lysyuk V. V., 2021
DOI 10.15588/1607-3274-2021-3-9

$$x_{ep} = \arg \min_{x:C(x)=C(x_a)} \left\| f(x_a) - f(x) \right\|, \quad (6)$$

$$x_{shn} = \arg \min_{\substack{C(x)\neq C(x_a) \\ x: \\ \|f(x_a)-f(x)\|>\|f(x_a)-f(x_p)\|}} \left\| f(x_a) - f(x) \right\|, \quad (7)$$

The next phase is needed to determine the binary class code in accordance with the error-correction output codes principles but also accounting for intraclass and interclass relationships. The training set of Z classes $\{x_{z,s} \mid z = \overline{1,Z},\ s = \overline{1,n_z}\}$ containing n_z samples of z-class, is encoded with a binary representation $\{b_{z,s,i} \mid z = \overline{1,Z},\ s = \overline{1,n_z},\ i = \overline{1,N}\}$ with dimensionality N. The binary coding of input image $x_{z,s}$ is achieved by rounding the i sigmoid layer output to the whole number.

$$b_{z,s,i} = \begin{cases} 1, & \text{if } f_i(x_{z,s}) > 0.5; \\ 0, & \text{otherwise.} \end{cases}$$

Binary reference vector b_z for class z can be calculated by [rank-wise] comparison of frequency of binary ones in z class with the background binary ones frequency in the training set

$$b_{z,i} = \begin{cases} 1, & \text{if } \frac{1}{n_z}\sum_{s=1}^{n_z} b_{z,s,i} > \frac{1}{Z}\sum_{c=1}^{Z}\frac{1}{n_c}\sum_{s=1}^{n_c} b_{c,s,i}; \\ 0, & \text{otherwise.} \end{cases}$$

Reference vector b_z for class z is used as a sample label in further training with Joint Binary Cross Entropy Loss, which is calculated for each training sample x as

$$L = -\sum_{i=1}^{N}\left(b_{z,i}\log f_i(x) + (1-b_{z,i})\log(1-f_i(x))\right).$$

The last stage of machine learning is related to the optimization of container radius by information criterion to account for the boundaries of deviation of binary representation of observations in each class from the corresponding reference vectors.

$$E_z^* = \max_{\{d\}} E_z(d). \quad (8)$$

Entropy criteria for a bi-alternative evaluation system (Z = 2) and equally probable hypothesis, representing the most statistically difficult case, can be expressed as a function of accuracy characteristics as follows [11]

$$E_z^{(k)} = 1 + \frac{1}{2}\left(\frac{\alpha_z^{(k)}(d)}{\alpha_z^{(k)}(d) + D_{2,z}^{(k)}(d)} \log_2 \frac{\alpha_z^{(k)}(d)}{\alpha_z^{(k)}(d) + D_{2,z}^{(k)}(d)} + \right.$$
$$+ \frac{\beta_z^{(k)}(d)}{D_{1,z}^{(k)}(d) + \beta_z^{(k)}(d)} \log_2 \frac{\beta_z^{(k)}(d)}{D_{1,z}^{(k)}(d) + \beta_z^{(k)}(d)} +$$
$$+ \frac{D_{1,z}(d)}{D_{1,z}^{(k)}(d) + \beta_z^{(k)}(d)} \log_2 \frac{D_{1,z}(d)}{D_{1,z}^{(k)}(d) + \beta_z^{(k)}(d)} +$$
$$\left. + \frac{D_{2,z}^{(k)}(d)}{\alpha_z^{(k)}(d) + D_{2,z}^{(k)}(d)} \log_2 \frac{D_{2,z}^{(k)}(d)}{\alpha_z^{(k)}(d) + D_{2,z}^{(k)}(d)} \right). \quad (9)$$

4 EXPERIMENTS

The current article considers all the training stages and their corresponding results obtained on the dataset provided by Ace Pipe Cleaning, Inc.

Class alphabet used to detect the observation context without taking into account the content of the neighbouring frames contains 10 main context classes (Table 1). This alphabet contains ignore, side and connection classes. Samples for these classes are easy to collect and label, however they do not provide complete certainty as to the camera orientation, as it is not clear which pipe wall camera is facing (Fig. 8).

Class alphabet used to detect the observation context with taking into account the content of the neighbouring frames contains 11 main context classes (Table 2). Temporal features have to provide complete certainty as to the camera orientation, hence there is no ignore class and side and connection classes have been replaced by Right, Left, Top, Bottom, Right connection, Left connection, Top connection and Bottom connection. Semi Left, Semi Right, Semi top, Semi bottom classes are replaced by the corresponding Right, Left, Top, Bottom. Temporal window for consideration of the neighbouring frames is set to 128 frames. This window was selected as a multiple of 2 and was experimentally selected as being close to optimal for various models. Parsing labelled video files $\{D_v^{train}\}$ and $\{D_w^{test}\}$ ensures the formation of variable quantity of samples for each class..

Prior unsupervised learning of the upper convolutional layers on unlabelled samples from the intended usage domain is aimed at increasing the subsequent supervised machine learning efficiency. It is worthwhile considering the influence the parameters of the growing sparse coding neural gas algorithm used in unsupervised learning have on the results of supervised learning. Table 1 presents the machine learning results and quantity Nc of generated convolutional filters (neurons) as a function of the parameter ν, which characterises the accuracy of coverage of the training set by the convolutional filters.

Before training, the entire dataset is balanced by applying augmentation to minor classes (0–5% change in scale, ±5% rotation, ±5% change in brightness).

Each model has a number of hyperparameters which define its configuration and capacity. Optimal hyperparameters are first selected for each architecture

© Moskalenko V. V., Moskalenko A. S., Zaretsky M. O., Panych A. O., Lysyuk V. V., 2021
DOI 10.15588/1607-3274-2021-3-9

and a comparison of results obtained from various architectures with these parameters. Models are trained during 60 epochs. The hyperparameters are selected with a view to avoid the noticeable overtraining effect.

Table 1 – Class set for the single-frame context classifier model

Designation of class	Number of examples	Names of context	Description
X_1^o	5000	Forward	Camera pointing forward along the pipe
X_2^o	3000	Side	Camera is pointing at the pipe wall facing left, right, down or up, when it is hard to understand which part of the pipe (top, bottom, left of right) the camera is pointing at
X_3^o	1000	Semi right	Incomplete right turn, orientation of the camera can be clearly determined from a single frame
X_4^o	1000	Semi left	Incomplete left turn, orientation of the camera can be clearly determined from a single frame
X_5^o	320	Semi top	Incomplete turn upwards, orientation of the camera can be clearly determined from a single frame
X_6^o	180	Semi bottom	Incomplete turn downwards, orientation of the camera can be clearly determined from a single frame
X_7^o	2100	Connection	Another pipe connecting to the main
X_8^o	500	Manhole	Point of entry into the pipe for inspection, from the manhole to the beginning of the pipe proper
X_9^o	152	Collapse	Collapsed pipe, further movement forward is impossible
X_{10}^o	100	Ignore	Situations to be ignored for processing purposes

Backbone of Mobilenet with the capacity coefficient set to 0.25 and input resolution set to 160x160 pixels was used as a single-frame feature extractor. Temporal convolutional network has 1 stage configuration with 7 BBs. Quantity of feature channels in convolutional filters of the first layer is set as C=128, to match the dimensionality of image embedding. The last layer of the temporal network is connected to the Dense layer, which contains 128 nodes with sigmois activation function. GRU and LSTM recurrent networks contain one layer with 128 units. States of each unit are submitted to the Dense layer which also contains 128 nodes with sigmoid activation function. After the sigmoid layer of either model two additional output layers analogous to those depicted in Fig. 2 are used, dependent on the chosen training method.

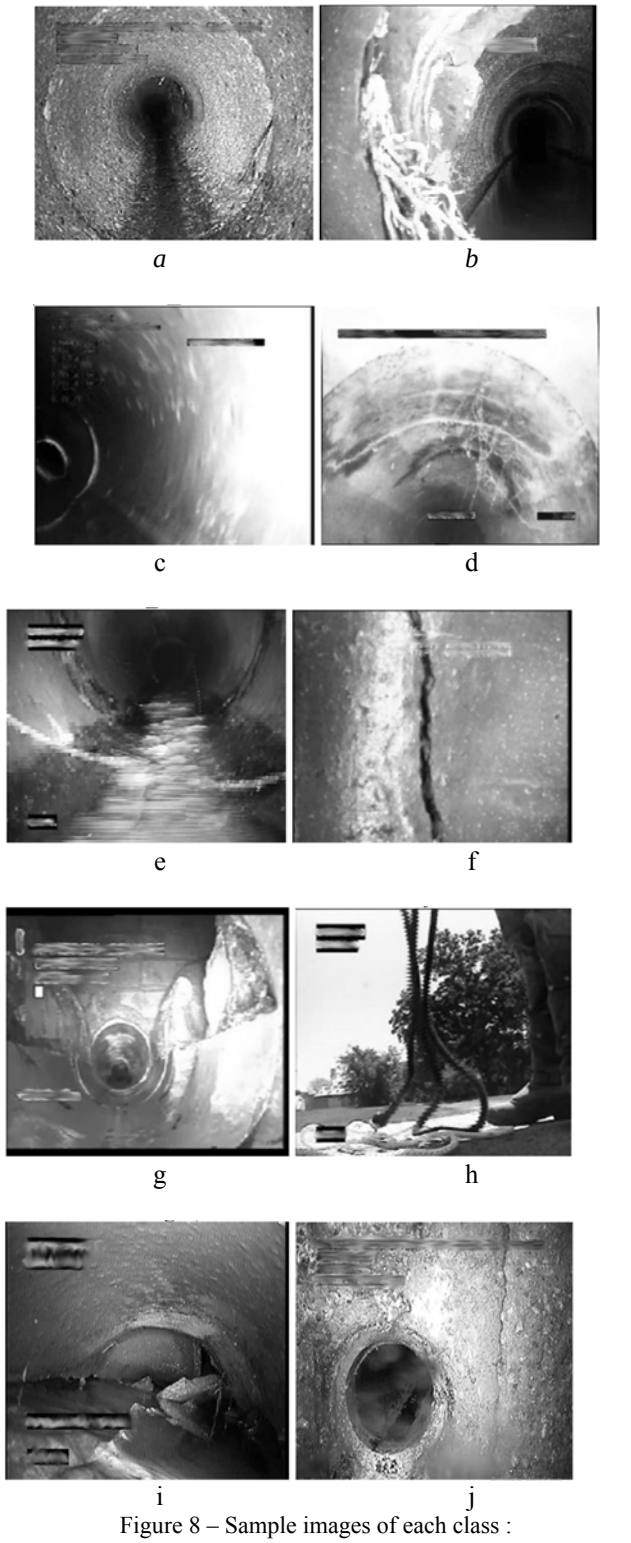

Figure 8 – Sample images of each class : a – class X_1^o ; b – class X_2^o ; c – class X_3^o ; d – class X_4^o ; e – class X_5^o ; f – class X_6^o ; g – class X_7^o ; h – class X_8^o ; i – class X_9^o ; j – class X_{10}^o

© Moskalenko V. V., Moskalenko A. S., Zaretsky M. O., Panych A. O., Lysyuk V. V., 2021
DOI 10.15588/1607-3274-2021-3-9

Table 2 – Class set for context classifier model accounting for neighbouring frames content

Designation of class designation	Number of examples	Names of context	Description
X_1^o	4500	Forward	Camera pointing forward along the pipe
X_2^o	1500	Right	Incomplete or complete right camera turn
X_3^o	1420	Left	Incomplete or complete left camera turn
X_4^o	400	Up	Incomplete or complete camera turn upwards
X_5^o	180	Down	Incomplete or complete camera turn downwards
X_6^o	500	Right con-nection	Connecting pipe on the right
X_7^o	500	Left con-nection	Connecting pipe on the left
X_8^o	150	Top con-nection	Connecting pipe at the top
X_9^o	75	Down con-nection	Connecting pipe at the bottom
X_{10}^o	500	Manhole	Point of entry into the pipe for inspection, from the manhole to the beginning of the pipe proper
X_{11}^o	170	Collapse	Collapsed pipe, further movement forward is impossible

Fig. 9 depicts the change in F1 score on the test dataset during the classifier training with the baseline and proposed (Fig. 7) methods. The test dataset was created by selecting 15% of the samples from each class represented in Table 1.

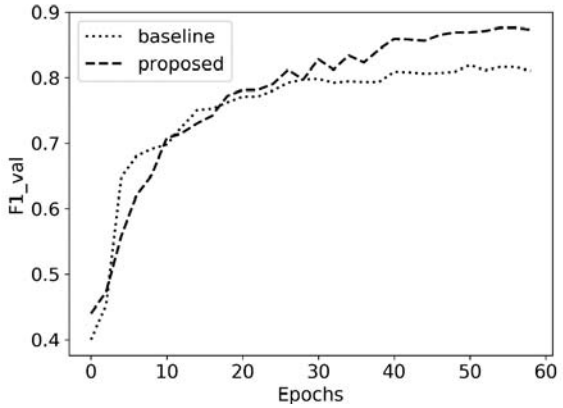

Figure 9 – Dependency of F1 score on test dataset from training epochs for single frame context classifier without accounting of neighbouring frames

Analysis of Fig. 9 shows that up to 30 epochs both training methods are performing with similar efficiency. However, the third phase (Fig. 7) of the proposed method, employed after 30th epoch, increases accuracy by 6%. Hence the feature extractor trained with the proposed training method is used for embedding into the high-level neural network.

Fig. 10 depicts the change in the F1 score on the test dataset when training the frame classifier which accounts for the neighbouring frames. Training is done with the baseline method but with varying architectures of the temporal network. Test dataset is created in the same way, but selecting 15% of the samples of each class represented in Table 2.

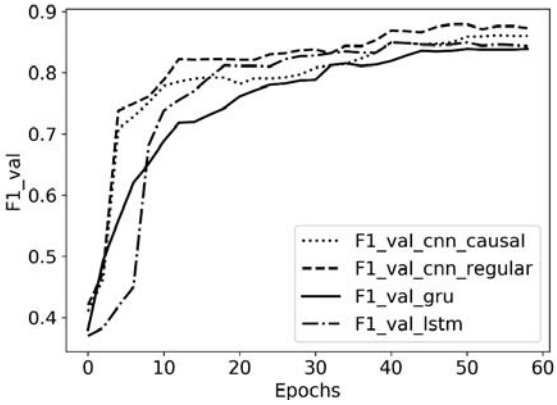

Figure 10 – Dependency of F1 score on test dataset from quantity of training epochs when training with the baseline method for context classifier with accounting of neighbouring frames

Analysis of Fig. 10 shows that the accuracy of recurrent networks during the initial training epochs increased slower than that of the convolutional networks. However, during the final stages the accuracy effectively reached a plateau with the results for different networks converging to virtually indistinguishable values. Amongst the convolutional network regular rather than causal architecture shown the best result. F1 score for the regular temporal network reached 0.87, whereas F1 for LSTM and GRU model stands at 0.843 and 0.839 respectively.

Fig. 11 shows the change in test dataset F1 score when training the classifier which accounts for the neighbouring frames trained with the proposed (Fig. 7) method using different temporal network architectures. Test network was formed in the same way by selecting 15% of the samples of each class illustrated in Table 2.

Analysis of Fig. 11 shows that accuracy of recurrent networks in the case is also inferior to those of the convolutional networks. Regular structure model had likewise shown to be the leader among the convolutional networks. However, the maximal F1 score for the model with the regular structure trained with the proposed method is 92%, which exceeds the baseline training method results on the same network by 4%.

© Moskalenko V. V., Moskalenko A. S., Zaretsky M. O., Panych A. O., Lysyuk V. V., 2021
DOI 10.15588/1607-3274-2021-3-9

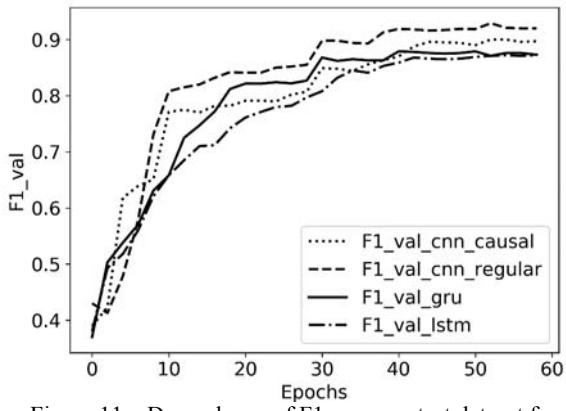

Figure 11 – Dependency of F1 score on test dataset from quantity of training epochs when training with the proposed method for context classifier with accounting of neighbouring frames

Thus the advantage of using a convolutional model for observation context analysis and the proposed multi-phase training method had been established. It was also empirically proven that the regular convolutional temporal network architecture is the most efficient in utilising interframe dependencies. Resulting accuracy is suitable for practical use, as the additional error correction can be made by using the odometer / distance counter data.

6 DISCUSSION

It is worth considering not just the aggregated metrics but also the confusion matrix to properly evaluate the efficiency of the proposed approach. A normalised confusion matric derived from the model with regular TCN trained with the conventional approach is presented in Fig. 12.

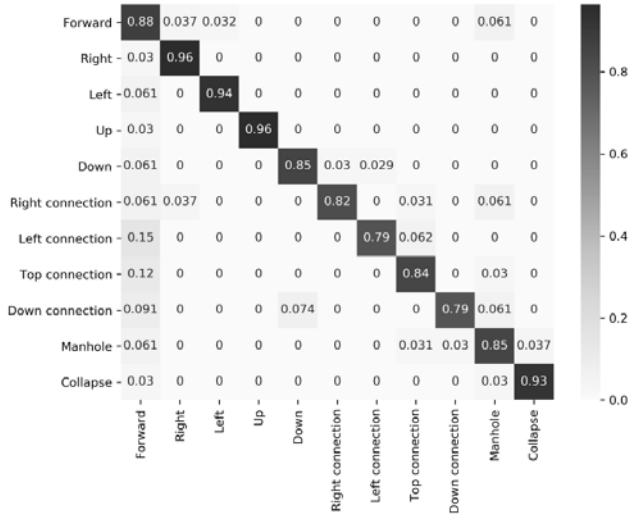

Figure 12 – Normalized confusion matrix for the optimal context classifier based on regular TCN trained with the conventional approach

Analysis of the confusion matric presented in Fig. 12 shows that the lowest accuracy corresponds to the complex classes : Right Connection, Left Connection, Top Connection and Manhole. These classes combine both camera orientation and distinctive structural elements of the sewer system.

Fig. 13 depicts the normalized confusion matrix for the context classifier based on the regular TCN trained with the proposed approach.

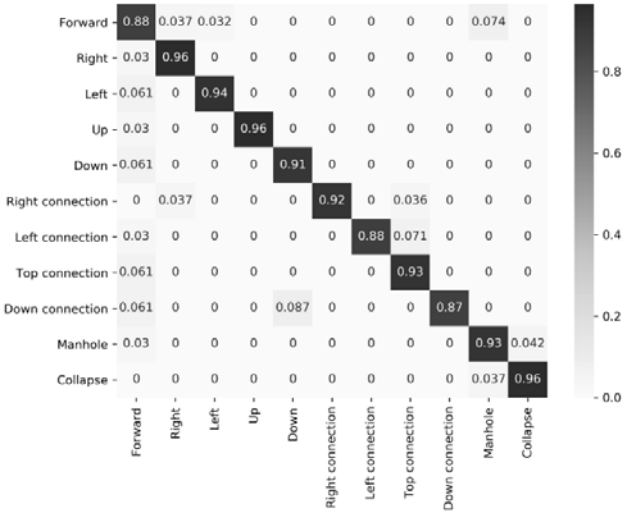

Figure 13 – Normalized confusion matrix for the optimal context classifier based on regular TCN trained with the proposed multi-phased approach

Analysis of Fig. 13 shows that the *forward* class is intersecting with almost all the other classes. In comparison with the conventional training methods, the proposed solution ensures increase in detection accuracy for complex context classes, such as Right Connection, Left Connection, Top Connection and Manhole.

Despite the improvement in accuracy, training on the same data is still not error-free. Some connections can be missed and for connections in 10:00–02:00 sector the camera orientation can be confused with the connections in the 02:00–05:00 and 07:00–10:00 sectors. This can be due to the insignificant difference between the specific degrees of partial camera turn. However, the impact of the erroneous context classification at the intersection of the context classes on the effectiveness of sewer pipe inspection is outside the scope of this study.

CONCLUSIONS

The scientific novelty of the work lies in the new models and methods of classification analysis of spatial and temporal context for automation of CCTV sewer inspections under the limited training dataset and data imbalance constraints.

Model contains a 2D convolutional network which produces feature embedding as an output of the sigmoid layer and 1D temporal convolutional regular network with the sigmoid and [rounding] output layers with radial-basis

© Moskalenko V. V., Moskalenko A. S., Zaretsky M. O., Panych A. O., Lysyuk V. V., 2021
DOI 10.15588/1607-3274-2021-3-9

decision rules created according to the error-correcting output codes and information-extreme learning.

The proposed method includes 4 stages : training with soft triplet loss and regularization component for penalising the binary code output rounding error; determination of the binary class code in accordance with the error-correcting output codes principles but also accounting for intraclass and interclass relationships; optimization of container radius by information criterion to account for the boundaries of deviation of binary representation of observations in each class from the corresponding reference vectors.

Training results obtained with the proposed method were compared with the results obtained by conventional training method, with a resulting 4% improvement in micro averaged F1 score metric. Confusion matrix analysis had shown that the biggest improvement in accuracy is observed for the classes which combine both the camera orientation and the distinctive structural features of the sewer system.

The practical significance of the achieved outcomes is an increase in accuracy of the classification analysis of temporal and spatial context during the CCTV sewer inspections under conditions of limited labelled training data availability and uncertainty related to the arbitrary observation conditions.

REFERENCES

1. Moradi S., Zayed T., Golkhoo F. Review on Computer Aided Sewer Pipeline Defect Detection and Condition Assessment, *Infrastructures*, 2019, Vol. 4, No. 1: 10. DOI: 10.3390/infrastructures4010010.
2. Myrans J., Everson R. , Kapelan Z. Automated detection of fault types in CCTV sewer surveys, *Journal of Hydroinformatics,* 2018, Vol. 21, No. 1, pp. 153–163. DOI: 10.2166/hydro.2018.073.
3. He M., Zhu Ch., Huang Q. et al. A review of monocular visual odometry, *The Visual Computer*, 2020, Vol. 36, No. 2, pp. 1053–1065. DOI: 10.1007/s00371-019-01714-6.
4. Lim B., Zohren S. Time-series forecasting with deep learning: a survey, *Philosofical Transactions of the Royal Society A*, 2021, Vol. 379, Issue 2194, P. 14. DOI: 10.1098/rsta.2020.0209.
5. Syahrian N. M., Risma P., Dewi T. Vision-Based Pipe Monitoring Robot for Crack Detection Using Canny Edge Detection Method as an Image Processing Technique, *Kinetik*, 2017, Vol. 2, No. 4, pp. 243–250. DOI: 10.22219/kinetik.v2i4.243.
6. Czimmermann T., Ciuti G., Milazzo M. et al. Visual-Based Defect Detection and Classification Approaches for Industrial Applications – A SURVEY, *Sensors*, 2020, Vol. 20, No. 5: 1459. DOI: 10.3390/s20051459.
7. Cheng J. C. P., Wang M. Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques, *Automation in Construction*, 2018, Vol. 95, pp. 155–171. DOI: 10.1016/j.autcon.2018.08.006.
8. Panella F., Boehm J., Loo Y. et al. Deep learning and image processing for automated crack detection and defect measurement in underground structures, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2018, Vol. XLII-2, pp. 829–835. DOI: 10.5194/isprs-archives-xlii-2-829-2018.
9. Zhan H., Shi B., Duan L.-Y. et al. DeepShoe: An improved Multi-Task View-invariant CNN for street-to-shop shoe retrieval, *Computer Vision and Image Understanding*, 2019. Vol. 180, pp. 23–33. DOI: 10.1016/j.cviu.2019.01.001.
10. Zhang B. , Tondi B. , Lv X. et al. Challenging the Adversarial Robustness of DNNs Based on Error-Correcting Output Codes, *Security and Communication Networks,* 2020, Vol. 2020: 8882494. DOI: 10.1155/2020/8882494.
11. Moskalenko V., Moskalenko A. , Korobov A. et al. The Model and Training Algorithm of Compact Drone Autonomous Visual Navigation System, *Data*, 2019, Vol. 4, No. 1: 4. DOI: 10.3390/data4010004.

Received 23.04.2021.
Accepted 24.08.2021.

УДК 004.891.032.26:629.7.01.066

МОДЕЛЬ І МЕТОД НАВЧАННЯ ДЛЯ КЛАСИФІКАЦІЙНОГО АНАЛІЗУ КОНТЕКСТУ КАДРІВ ВІДЕОІНСПЕКЦІЇ СТІЧНИХ ТРУБ

Москаленко В. В. – канд. техн. наук, доцент кафедри комп'ютерних наук, Сумський державний університет, Суми, Україна.

Зарецький М. О. – аспірант кафедри комп'ютерних наук, Сумський державний університет, Суми, Україна.

Москаленко А. С. – канд. техн. наук, старший викладач кафедри комп'ютерних наук, Сумський державний університет, Суми, Україна.

Панич А. О. – магістр інженерії, асистент кафедри комп'терних наук, Сумський державний університет, Суми, Україна.

Лисюк В.В. – магістр інженерії, співзасновник компанії Molfar.AI sp. z o.o., Гданськ, Польща.

АНОТАЦІЯ

Актуальність. Розроблено та досліджено модель та метод навчання для класифікації контекстів спостереження на кадрах відеоінспекції стічних труб. Об'єктом дослідження є процес виявлення просторово-часового контексту під час інспекцій стічних труб. Предметом дослідження є модель та метод машинного навчання для класифікаційного аналізу кадрів відеоінспекції в умовах обмеженого та незбалансованого набору розмічених навчальних даних.

Мета дослідження – розроблення ефективних моделі і методу машинного навчання для класифікаційного аналізу контексту відеокадрів інспекції стічних труб в умовах обмеженого обсягу та незбалансованості розміченого навчального набору даних.

Методи дослідження. Запропоновано чотирьох етапний алгоритм навчання класифікатора. Перший етап полягає у навчанні з нормалізованою триплетною функцією втрат і регуляризуючою складовою, яка штрафує за помилку округлення вихідного сигналу до двійкового подання. Наступний етап полягає у визначенні двійкового коду для кожного класу для

© Moskalenko V. V., Moskalenko A. S., Zaretsky M. O., Panych A. O., Lysyuk V. V., 2021
DOI 10.15588/1607-3274-2021-3-9

реалізації кодів, що виправляють помилки, але з урахуванням внутрішньокласових та міжкласових відношень. Отриманий еталонний двійковий вектор для кожного класу потім використовується як цільова мітка під час наступного етапу навчання з бінарною крос-ентропійною функцією втрат. Останній етап машинного навчання пов'язаний з оптимізацією параметрів правила прийняття рішень за інформаційним критерієм для визначення допустимих меж відхилення двійкового подання спостережень кожного класу від відповідного еталонного вектора. Розглядається 2D згортковий екстрактор ознак у поєднанні з темпоральною мережею для аналізу міжкадрових залежностей. Розглядаються варіанти з 1D згорткова мережа з дірявими регулярними згортками, 1D згорткова мережа з дірвими причинно-наслідковими згортками, рекурентна мережа LSTM та рекурентна мережа GRU. Порівняння ефективності моделей проводиться на основі мікро усередненої F1-міри, обчисленої на тестовому наборі даних.

Результати. Результати, отримані за набором даних, наданим Ace Pipe Cleaning, Inc, підтверджують придатність моделі та методу для практичного використання, оскільки отримана точність дорівнює 92%. Порівняння результатів навчання із запропонованим методом та традиційним методом показало перевагу на 4% за мікро-усередним значенням F1-міри. Подальший аналіз матриці помилок показав, що найбільш суттєве підвищення точності порівняно зі традиційними методами досягається для складних класів, які поєднують як орієнтацію камери, так і особливості конструкції стічної труби.

Висновки. Наукова новизна роботи полягає у нових моделях та методах класифікаційного аналізу просторово-часового контексту для автоматизації відеоінспекції стічних труб в умовах обмеженого обсягу та незбалансованості розмічених навчальних даних. Результати навчання, отримані за пропонованим методом, порівнюються з результатами, отриманими за допомогою традиційного методу класифікаційного аналізу зображень. Запропонований метод продемонстрував перевагу на 4% за мікро-усередненим значенням F1-міри.

Емпірично було доведено, що темпоральна мережа на основі 1D згорткової мережі з дірявими регулярними згортками є найбільш ефективною для аналізу міжкадрових залежностей. Отримана точність забезпечує придатність отриманих моделей для практичного використання, оскільки додаткове виправлення помилок можна реалізувати на основі даних одометра.

КЛЮЧОВІ СЛОВА: інспекція стічних труб, згорткова нейронна мережа, коди з самокорекцією помилок, сіамська нейронна мережа, інформаційно-екстремальне навчання, інформаційний критерій, LSTM, GRU.

УДК 004.891.032.26:629.7.01.066

МОДЕЛЬ И МЕТОД ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИОННОГО АНАЛИЗА КОНТЕКСТА КАДРОВ ВИДЕОИНСПЕКЦИИ СТОЧНЫХ ТРУБ

Москаленко В. В. – канд. техн. наук, доцент кафедры компьютерных наук, Сумской государственный университет, Сумы, Украина.

Зарецкий Н. А. – аспирант кафедры компьютерных наук, Сумской государственный университет, Сумы, Украина.

Москаленко А. С. – канд. техн. наук, старший преподаватель кафедры компьютерных наук, Сумской государственный университет, Сумы, Украина.

Паныч А. А. – магистр инженерии, ассистент кафедры компьютерных наук, Сумской государственный университет, Сумы, Украина.

Лысюк В. В. – магистр инженерии, сооснователь компании Molfar.AI sp. z o.o., Гданск, Польша.

АННОТАЦИЯ

Актуальность. Разработаны и исследованы модель и метод обучения для классификации контекстов наблюдения на кадрах видеоинспекции сточных труб. Объектом исследования является процесс распознавания пространственно-временного контекста во время инспекций сточных труб. Предметом исследования является модель и метод машинного обучения для классификационного анализа кадров видеоинспекции в условиях ограниченного и несбалансированного набора размеченных обучающих данных.

Цель исследования – разработка эффективных модели и метода машинного обучения для классификационного анализа контекста видеокадров инспекции сточных труб в условиях ограниченного объема и несбалансированности размеченного обучающего набора данных.

Методы исследования. Предложено четырехэтапный алгоритм обучения классификатора. Первый этап заключается в обучении с нормализованной триплетного функцией потерь и регуляризирующей составляющей, которая штрафует за ошибку округления выходного сигнала к двоичному представлению. Следующий этап заключается в определении двоичного кода для каждого класса для реализации кодов, исправляющих ошибки, но с учетом внутрикласовых и межклассовых отношений. Полученный эталонный двоичный вектор для каждого класса затем используется как целевая метка во время следующего этапа обучения с бинарной кросс-энтропийной функцией потерь. Последний этап машинного обучения связан с оптимизацией параметров правила принятия решений за информационным критерием для определения допустимых пределов отклонения двоичного представления наблюдений каждого класса от соответствующего эталонного вектора. Рассматривается 2D сверточный экстрактор признаков в сочетании с темпоральной сетью для анализа межкадровых зависимостей. Рассматриваются варианты 1D сверточной сети с дырявыми регулярными свертками, 1D сверточной сети с дырявыми причинно-следственными свертками, рекуррентная сеть LSTM и рекуррентная сеть GRU. Сравнение эффективности моделей производится на основе микро усредненной F1-меры, которая вычисляется на тестовом наборе данных.

Результаты. Результаты, полученные на наборе данных, предоставленным Ace Pipe Cleaning, Inc, подтверждают пригодность модели и метода для практического использования, так как полученная точность равна 92%. Сравнение результатов обучения за предложенным методом с результатами за традиционным методом показало преимущество на 4% за микро-усредненным значением F1-меры. Дальнейший анализ матрицы ошибок показал, что наиболее существенное

© Moskalenko V. V., Moskalenko A. S., Zaretsky M. O., Panych A. O., Lysyuk V. V., 2021
DOI 10.15588/1607-3274-2021-3-9

повышение точности по сравнению с традиционными методами достигается для сложных классов, которые объединяют как ориентацию камеры, так и особенности конструкции сточной трубы.

Выводы. Научная новизна работы заключается в новых модели и методе классификационного анализа пространственно-временного контекста для автоматизации видеоинспекции сточных труб в условиях ограниченного объема и несбалансированности размеченных обучающих данных. Результаты обучения, полученные по предлагаемому методу, сравниваются с результатами, полученными с помощью традиционного метода классификационного анализа изображений. Предложенный метод продемонстрировал преимущество на 4% за микро-усредненным значением F1-меры.

Эмпирически было доказано, что темпоральная сеть на основе 1D сверточной сети с дырявыми регулярными свертками является наиболее эффективной для анализа межкадровых зависимостей. Полученная точность обеспечивает пригодность полученных моделей для практического использования, поскольку дополнительное исправление ошибок можно реализовать на основе данных одометра.

КЛЮЧЕВЫЕ СЛОВА: инспекция сточных труб, сверточная нейронная сеть, коды с самокоррекцией ошибок, сиамская нейронная сеть, информационно-экстремальное обучение, информационный критерий, LSTM, GRU.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Defect Detection and Condition Assessment / S. Moradi, T. Zayed, F. Golkhoo // Infrastructures. – 2019. – Vol. 4, No. 1: 10. DOI: 10.3390/infrastructures4010010.
2. Myrans J. Automated detection of fault types in CCTV sewer surveys / J. Myrans, R. Everson, Z. Kapelan // Journal of Hydroinformatics. – 2018. – Vol. 21, No. 1. – P. 153–163. DOI: 10.2166/hydro.2018.073.
3. A review of monocular visual odometry / [M. He, Ch. Zhu, Q. Huang et al.] // The Visual Computer. – 2020. – Vol. 36, No. 2. – P. 1053–1065. DOI: 10.1007/s00371-019-01714-6.
4. Lim B. Time-series forecasting with deep learning: a survey / B. Lim, S. Zohren // Philosofical Transactions of the Royal Society A. – 2021. – Vol. 379, Issue 2194. – P. 14. DOI: 10.1098/rsta.2020.0209.
5. Syahrian N. M. Vision-Based Pipe Monitoring Robot for Crack Detection Using Canny Edge Detection Method as an Image Processing Technique / N. M. Syahrian, P. Risma, T. Dewi // Kinetik. – 2017. – Vol. 2, No. 4. – P. 243–250. DOI: 10.22219/kinetik.v2i4.243.
6. Visual-Based Defect Detection and Classification Approaches for Industrial Applications – A SURVEY / [T. Czimmermann, G. Ciuti, M. Milazzo et al.] // Sensors. – 2020. – Vol. 20, No. 5: 1459. DOI: 10.3390/s20051459.
7. Cheng J. C. P. Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques / J. C. P. Cheng, M. Wang // Automation in Construction. – 2018. – Vol. 95. – P. 155–171. DOI: 10.1016/j.autcon.2018.08.006.
8. Deep learning and image processing for automated crack detection and defect measurement in underground structures / [F. Panella, J. Boehm, Y. Loo et al.] // The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. – 2018. – Vol. XLII-2. – P. 829–835. DOI: 10.5194/isprs-archives-xlii-2-829-2018.
9. DeepShoe: An improved Multi-Task View-invariant CNN for street-to-shop shoe retrieval / [H. Zhan, B. Shi, L.-Y. Duan et al.] // Computer Vision and Image Understanding. – 2019. – Vol. 180. – P. 23–33. DOI: 10.1016/j.cviu.2019.01.001.
10. Challenging the Adversarial Robustness of DNNs Based on Error-Correcting Output Codes / [B. Zhang, B. Tondi, X. Lv et al.] // Security and Communication Networks. – 2020. – Vol. 2020: 8882494. DOI: 10.1155/2020/8882494.
11. The Model and Training Algorithm of Compact Drone Autonomous Visual Navigation System / [V. Moskalenko, A. Moskalenko, A. Korobov et al.] // Data. – 2019. – Vol. 4, No. 1: 4. DOI: 10.3390/data4010004.

© Moskalenko V. V., Moskalenko A. S., Zaretsky M. O., Panych A. O., Lysyuk V. V., 2021
DOI 10.15588/1607-3274-2021-3-9