This is an Accepted Manuscript version of the following article, accepted for publication in **CYBERNETICS AND SYSTEMS**.
Postprint of: Silva De Oliveira C., Sanin C., Szczerbicki E., Visual Content Learning in a Cognitive Vision Platform for Hazard Control (CVP-HC), CYBERNETICS AND SYSTEMS, Vol. 50, Iss. 2 (2019), pp.197-207, DOI: 10.1080/01969722.2019.1565116
It is deposited under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

Visual Content Learning in a Cognitive Vision Platform for Hazard Control (CVP-HC)

Caterine Silva de Oliveira[1], Cesar Sanin[1], and Edward Szczerbicki[2]

[1] *Department of Mechanical Engineering, University of Newcastle, Callaghan, NSW, Australia*

(Caterine.SilvaDeOliveira@uon.edu.au, cesar.sanin@newcastle.edu.au)

Address: ES320, Faculty of Engineering and Built Environment, University of Newcastle, Callaghan, NSW 2308, Australia;

[2] *Faculty of Management and Economics, Gdansk University of Technology, Gdansk, Poland*

(edward.szczerbicki@zie.pg.gda.pl)

Visual Content Learning in a Cognitive Vision Platform for Hazard Control (CVP-HC)

This work is part of an effort for the development of a Cognitive Vision Platform for Hazard Control (CVP-HC) for applications in industrial workplaces, adaptable to a wide range of environments. The paper focuses on hazards resulted from the non-use of personal protective equipment (PPE). Given the results of previous analysis of supervised techniques for the problem of classification of a few PPE (boots, hard hats and gloves extracted from frames of low resolution videos), which found the Deep Learning (DL) methods as the most suitable ones to integrate our platform, the objective of this paper is to test two DL algorithms: Single Shot Detector (SSD) and Faster Region-based Convolutional Network (Faster R-CNN). The testing uses pre-trained models on a second version of our PPE dataset (containing 11 classes of objects) and evaluates which of examined algorithms is more appropriate to compose our system reasoning.

Introduction and Background

Hazards are present in all workplaces environments and can result in injuries, illnesses, or death and their control is essential to ensure the safety of workers and occupational health (Safetycare Australia, 2015). In this context, monitoring of labourers daily activities and to identify any exposure to risks emerged as a need. The use of sensors data and computer vision technologies can give support to a fast and automated detection of potentially dangerous situations. This information might be utilized, for instance, to provide feedback and real time recommendations to avoid accidents, or to evaluate how programs and interventions are impacting particular safety problems and outcomes and help managing employees' behaviour to perform the work in a safe manner (Han & Lee,

MOST WIEDZY Downloaded from mostwiedzy.pl

2013). However, as indicated by Little et al. (2013), currently there is no such flexible system capable of performing well in different industrial environments and situations without the necessity of rewriting most of existing application code each time the circumstances or settings change. In terms of accuracy, the performance of available systems designed to attend a more comprehensive diversity of scenarios and applications when operating in real life is still limited (Chen, et al 2012). For this reason, building an automatic system capable to support safety management of a variety of scenarios subject to different settings and conditions at the same time as being specific and meaningful still remains a challenge.

This work is part of an effort for the development of a flexible Cognitive Vision Platform for Hazard Control (CVP-HC) for applications in industrial workplaces, attending a wide range of industrial environments (de Oliveira et al 2017). In this system, visual content, sensorial data and any context information, is collected through the platform and represented explicitly using the Set of Experience Knowledge Structure (SOEKS or SOE for short), grouped and stored as Decisional DNA (DDNA) (Sanín & Szczerbicki, 2005; Sanín & Szczerbicki, 2007, Shafiq et al 2014). The collected knowledge is used for reasoning and also to retrain the system from time to time, customizing the service according to each scenario and application, and improving its specificity.

This paper focuses on hazards resulted from the non-use of personal protective equipment (PPE). Given the results of previous analysis (de Oliveira et al. 2017) of supervised techniques for the problem of classification of a few PPE (boots, hard hats and gloves extracted from frames of low resolution videos), which found the Deep Learning (DL) methods as the most suitable ones to integrate our platform, the objective of this paper is to test two DL architectures Single Shot Detector (SSD) and Faster Region-based

MOST WIEDZY Downloaded from mostwiedzy.pl

Convolutional Network (Faster R-CNN) from pre-trained models on an second version of PPE dataset (containing more classes of objects), and evaluate which of these architectures is more appropriate to compose our reasoning system.

The rest of the paper is organized as follows: in the "Cognitive Vision Approach" section, a background about cognitive vision systems is introduced together with the overall architecture of the proposed system and the knowledge representation methodology implemented to facilitate the management of knowledge in the system. The section "Deep Learning" presents the concept of Deep Neural Networks, including the SSD and Faster R-CNN algorithms. The "Methodology" section explains the steps for creating the PPE Dataset, and describes the model training and evaluation stages. The section "Experimental Results" presents the performance for the SSD and Faster R-CNN algorithms. In the last section conclusions and future work are presented.

Cognitive Vision Approach

Computer vision techniques can be used to support automatic detection and tracking of workers indicating potential dangerous situations. Visual sensing facilities, such as video cameras, can monitor labourers' behaviour and conditions of the environment and the generated data (such as video sequences or digitized visual data) can be processed in powerful computers to generate inferences and predictions (Chen et al 2012). However, the accuracy of current computer vision systems when operating in real time, subject to change in illumination, backgrounds variations, occlusions and low camera resolutions still remains a challenge (Mosberger et al 2013). Furthermore, these technologies are commonly not scalable and lack adaptability to the wide industrial environments and situations existing. Consequently, they create case-based applications that work only for specific circumstances and any change in conditions or setting would result in rewriting

MOST WIEDZY Downloaded from mostwiedzy.pl

most of the application code (Zambrano et al 2015).

In this context, methods incorporating prior knowledge to mimic the human-like capabilities are gaining attention from the research community. One of the latest trends to achieve this goal is the joining of cognition and computer vision into cognitive computer vision. In cognitive vision systems knowledge and learning are central elements to reason about events and for the decision making process. The gathered knowledge (visual and contextual) can be used to retrain the system from time to time to customize the service for each workplace and application. Moreover, with explicit contextual information gathered from the existing settings it is possible to enhance the speed and accuracy of the detection algorithm and reduce scalability issues (Davis et al 1993).

Cognitive Vision Platform for Hazard Control (CVP-HC)

The overall architecture of the proposed CVP-HC is presented in Figure 1. The platform is composed by six layers: System Configuration, Central Reasoning, Experience Creation, Experience Validation, System Monitoring and Output Layer.

MOST WIEDZY Downloaded from mostwiedzy.pl

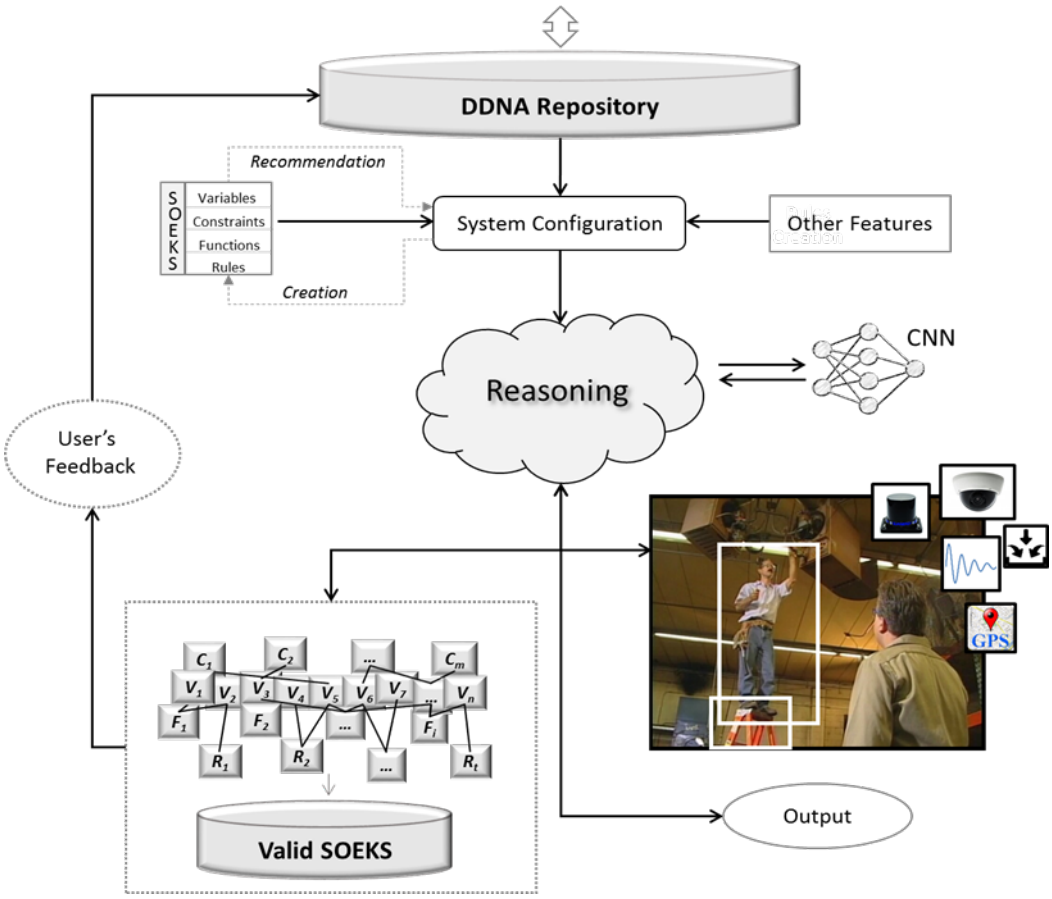

Figure 1. Overall architecture of the Cognitive Vision Platform for Hazard Control (CVP-HC).

The *System Configuration layer* consist of the selection (or creation) of the attributes according to requirements of the organization, and configuration of extra functionalities such as frame, experience creation and learning rate. In the *Experience Creation layer* attributes are synchronized according to their timestamp and used to create experiences. During *Experience Validation* experiences are created and compared among each other and the most redundant ones are pruned. Some of the remaining experiences are used to query the user and check if given solution is satisfactory. Once validated the SOE is stored in the decisional DNA repository for reuse and sharing. The *Central Reasoning layer* is the intelligence of the whole system. It is composed by a Deep Neural Network arranged in a hierarchical structure to support detection of attributes, location and relationship among them and its interpretation inside the given context. The Deep Neural Network

MOST WIEDZY Downloaded from mostwiedzy.pl

enhances the gained experience from formal decision events and transforms it into new knowledge. In summary, it uses SOEs as input and produces enhanced knowledge in SOEs format to compose the DDNA. This DDNA is a base for predictions according to experienced knowledge and learnt knowledge (due to DDNA and CNN respectively). The *Monitoring* layer represents the monitoring of workers' activities and feeds the reasoning with visual and contextual information to be structured, explicitly represented and processed. Finally, in the *Output layer*, when a hazard or risky situation is identified by the system, an alert message is shown with details and recommendations of action to be taken.

Set of Experience Knowledge Structure (SOEKS) and Decisional DNA (DDNA)

SOEKS is a knowledge representation structure designed to gather and store formal decision events in an explicit form. SOEKS or in short SOE is based on four fundamental elements of decision-making actions: variables, functions, constraints and rules. The most basic element of SOEKS is the variable, which is usually used to represent knowledge in an attribute-value form, following the traditional approach for knowledge representation. Functions, constraints, the rules are different ways of establishing relationships among variables. Functions define relations between dependent variables and a set of input variables; consequently, functions are used to build multi-objective goals. Similarly, constraints are functions, but they act as a way to limit possibilities, restricting the set of possible solutions and controlling the performance of the system in relation to its goals. Lastly, rules are associations that work in the universe of variables, expressing condition-consequence connections as "if-then-else" and are used to represent inferences and conditions under which the system should be implemented (Sanín & Szczerbicki, 2005).

MOST WIEDZY Downloaded from mostwiedzy.pl

In our platform, the experiences are grouped according to the areas of decision categories – the Decisional DNA (DDNA). A SOEKS works as a gene that guides decision-making and is a portion of an organization's DDNA. This gene belongs to a decisional chromosome from a certain category or type. A group of chromosomes from different categories (e.g. safety decisions, human resources decisions and product development decisions) comprise the DDNA of the given organization. DDNA is used in out Cognitive Platform to solve the scalability issues found in current vision-based approaches by introducing an experience-based approximation to recognize events defined by the user using production rules, adaptable to different work environment conditions, clients and situations (Sanín & Szczerbicki, 2007).

Deep Learning

Neural networks have been proved to be a very powerful Machine Learning technology and applied both in binary and multi-class problems. The first functional networks with many layers (Deep Neural Network DNN) was proposed in late 19ths. Nevertheless, at that time the computers didn't have enough processing power to successfully handle the work required by large neural networks (Ivakhnenko et al 1967). The main trigger for the renewed interest in neural networks and their learning capabilities was the backpropagation algorithm that accelerated the training of multi-layer networks (Werbos, 1974). Another method known as Dropout used to reduce overfitting (a very common issue in deep neural networks) contributed to major improvements over other regularization methods (Srivastava et al 2014).

To learn a wide range of possible hazards that workers may be exposed to in industrial environments such as when accessing controlled zones without authorization, crossing yellow lines, not respecting safe distances from machines and areas, not wearing

MOST WIEDZY Downloaded from mostwiedzy.pl

the required personal protective equipment (PPE), among others, a model with a large learning capacity is needed. Deep Neural Networks constitute one such class of models. Such supervised method was the first artificial pattern recognizer to achieve human-competitive performance on certain tasks (Ciresan et al 2011).

Recently, Deep Convolutional Neural Networks (CNN, or ConvNets) have significantly improved image classification and object detection to a wide range of domains improving the accuracy and speed of early Deep Neural Networks (Girshick et al 2015; Krizhevsky et al 2012; Sermanet et al 2014).

Faster Region-based Convolutional Network (Faster R-CNN)

Fast R-CNN is a ConvNets method that has been proposed to reduce the complexity of multi-stage algorithms. It is single-stage training algorithm that combined learns to classification of object and refine their spatial locations. This method employs several innovations to improve training and testing speed at the same time it increases detection accuracy. Comparing the results of Fast R-CNN to other algorithms, it achieves state-of-the-art mAP (Mean Average Precision) on PASCAL VOC2007 (VOC - Visual Object Classes), VOC2010, and VOC2012 datasets and faster training and testing compared to R-CNN, SPPnet (Spatial Pyramid Pooling) (Girshick, 2015).

Faster R-CNN is a combination of Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals and Fast-R-CNN detector that uses these proposed regions (Ren et al 2017).

Single Shot Detector (SSD)

The Single Shot Detector (SSD) method has been proposed recently to detect objects in

MOST WIEDZY Downloaded from mostwiedzy.pl

images using a single deep neural network. Experimental results on the PASCAL VOC, COCO (Common Objects in Context - large scale object detection, segmentation, and captioning), and ILSVRC (ImageNet Large Scale Visual Recognition Competition) datasets confirm that SSD has competitive accuracy to state-of-art methods and is much faster, while providing a unified framework for both training and inference. It makes the monolithic and relatively simple SSD model a useful building block for larger systems that employ an object detection component (Liu et al 2016).

Methodology

In this section, (i) the process of creating the dataset is explained; and (ii) the training and evaluation process is described.

Dataset

For the creation of the dataset 30 videos of surveillance cameras of industries have been downloaded from the internet. Frames were extracted from these videos, totalizing 19,303 images codified in JPG format. These images were filtered and the ones not containing any worker wearing a personal protective equipment manually removed. After filtering, the remaining 9,029 have been used for annotation of gloves, safety boots, hard hats, earmuffs, eye protector (such as goggles and glasses), face masks, headlamp, high visibility clothes, respirators, safety harness and welder masks. These objects result in a total of 11 different classes. The process of creation of the dataset is illustrated in Figure 2.

MOST WIEDZY Downloaded from mostwiedzy.pl

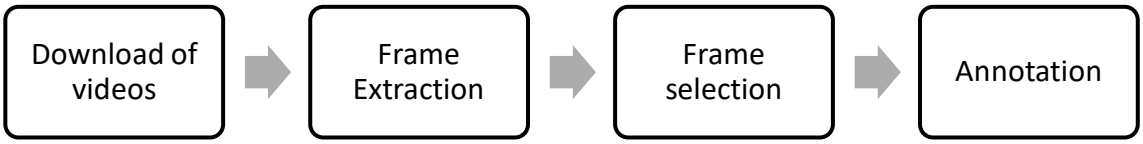

Figure 2: Image processing flow.

The annotation is at this stage still an ongoing work and for the tests of examined methods a total of 1878 annotations have been used. For the annotation process, LabelImg, a graphical image annotation tool is being used. An example of an annotated image (selected regions and labels) is shown if Figure 3.

Figure 3: Example of annotated image.

Training and Evaluation

The experiment is performed through Tensorflow Object Detection API (Application Programming Interface) (Huang et al 2017). To accelerate the process, we performed training and evaluation on Google Cloud ML Engine.

To be able to train a model using TensorFlow API, a few steps must followed. Firstly, the training and evaluation annotations in XML files in PASCAL VOC format and images of the dataset is used to create TFRecords, the Tensorflow standard data

MOST WIEDZY Downloaded from mostwiedzy.pl

format. In addition a label map containing all classes must is created in .pbtxt format. Finally, the object detection pipeline is configured.

The Google Cloud ML (Machine Learning) Engine has been configured with a cluster with five training jobs and three parameters servers. The analysis of the results and comparison among the algorithms presented in next section is based, essentially, on Detection quality (map) and learning rate along the time visualized over a Tensorflow board.

Experimental Results

Training a state of the art object detector from scratch can take days, even when using multiple GPUs (Graphics Processing Unit). So, in order to speed training process, parameters from object detectors trained on a different datasets for Faster R-CNN (faster_rcnn_resnet101_coco_11_06_ 2017) and SSD (ssd_mobilenet_v1_coco_11_06_2017) is reused to initialize our new model. This is a common process known as transfer learning. The Figure 4 shows the loss decreased very fast due to the pre-trained models for SSD and for R-CNN.

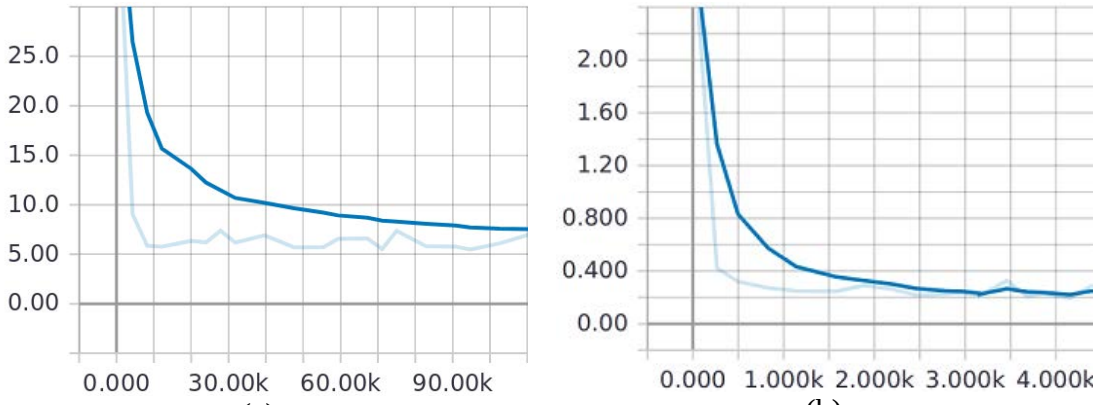

Figure 4: Total loss curve for (a) SSD and (b) Faster R-CNN when using pre-trained models.

MOST WIEDZY Downloaded from mostwiedzy.pl

The detection quality map is presented in Figure 5 for both methods. As can be observed they present good accuracy given the limited number samples and characteristics of them (low resolution of the objects, noise, occlusions etc.). Faster R-CNN has been proven to give better overall accuracy when compared to SSD, but at costs of much higher training time. The same is applied to our analysis. The total training time for SSD was 5h 32m 52s whilst for Faster R-CNN was 13h 22m 36s. SSD converges after 90K steps (84.51% accuracy), which is equivalent to 3h 56m 50s and Faster R-CNN after around 4h of training is still at step 5K and accuracy at lower than 77%.

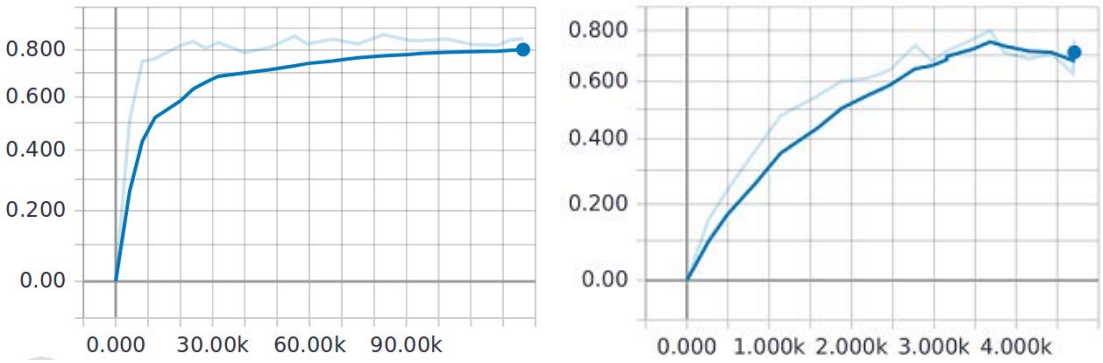

Figure 5: Detection quality (map) for (a) SSD and (b) Faster R-CNN.

Finally, detection examples resulting from the trained model applied to the test dataset is shown in Figure 6.

MOST WIEDZY Downloaded from mostwiedzy.pl

Figure 6: Detection examples on PPE dataset, each colour corresponding to an object class.

Conclusion and Future Work

This study demonstrated the power of Convolutional Neural Networks to solve classification problems, and the comparison of two algorithms (SSD and Faster R-CNN) for the detection of 11 classes of personal protective equipment (PPE) extracted from frame-videos of low resolution cameras, which were taken in real life industrial environments (subject to noise, occlusion, change in illumination etc.). From the preliminary results obtained for small portion of annotated images, a good accuracy has been achieved using both SSD and Faster R-CNN. In spite of being faster, SSD usually performs worse for small objects comparing to others methods. However, for our dataset, the accuracy gap between SSD and Faster R-CNN is very small whilst the training time is 3 times longer for Faster R-CNN. For this reason SSD is shown as a good option to be used for learning purposes in our platform.

In future work, the annotation process of the entire dataset will be completed to enrich the system with more experiences, reduce overfitting and improving accuracy when tested in real time settings. Additionally, other objects that might represent risks

MOST WIEDZY Downloaded from mostwiedzy.pl

will be considered as well as activities (or interactions of the workers with these objects) and their representation in the universe of SOEKS. Finally, the reuse of these experiences will be examined when reasoning about existing risks while the system is running in real time, as well as for retraining process (and improvement of specificity) for a given number of different scenarios.

References

Chen, L., Hoey, J., Nugent, C. D., Cook, D. J., & Yu, Z. (2012). Sensor-based activity recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(6), 790-808.

Ciresan, D. C., Meier, U., Masci, J., Maria Gambardella, L., & Schmidhuber, J. (2011, July). Flexible, high performance convolutional neural networks for image classification. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence* (Vol. 22, No. 1, p. 1237).

Davis, R., Shrobe, H., & Szolovits, P. (1993). What is a knowledge representation?. *AI magazine*, *14*(1), 17.

de Oliveira, C. S., Sanin, C., & Szczerbicki, E. (2017). Hazard Control in Industrial Environments: A Knowledge-Vision-Based Approach. In *International Conference on Information Systems Architecture and Technology* (pp. 243-252). Springer, Cham.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. Region-based convolutional networks for accurate object detection and segmentation. TPAMI, 2015.

Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

MOST WIEDZY Downloaded from mostwiedzy.pl

Han, S., & Lee, S. (2013). A vision-based motion capture and recognition framework for behavior-based safety management. *Automation in Construction*, *35*, 131-141.

Huang, J., Rathod, V., Chow, D., Sun, C., Zhu, M., Fathi, A., & Lu, Z. (2017). Tensorflow object detection api. *Code:github.com/tensorflow/models/tree/master/object detection.*

Ivakhnenko, A. G., & Lapa, V. G. (1967). Cybernetics and forecasting techniques.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Little, S., Jargalsaikhan, I., Clawson, K., Nieto, M., Li, H., Direkoglu, C., & Liu, J. (2013, April). An information retrieval approach to identifying infrequent events in surveillance video. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval* (pp. 223-230). ACM.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

Mosberger, R., Andreasson, H., & Lilienthal, A. J. (2013, November). Multi-human tracking using high-visibility clothing for industrial safety. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on* (pp. 638-644). IEEE.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91-99).

Safetycare Australia. (2015). *Recognition, evaluation and control of hazards*. Victoria.

MOST WIEDZY Downloaded from mostwiedzy.pl

Sanín, C., Szczerbicki, E. (2005). Set of experience: a knowledge structure for formal decision-events. *Found. Control Manag. Sci*., 3, 95–113

Sanín, C., Szczerbicki, E. (2007). Towards the construction of decisional DNA: a set of experience knowledge structure java class within an ontology system. *Cybern. Syst.*, 38

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In ICLR, 2014.

Shafiq, S. I., Sanin, C., Szczerbicki, E. 2014. Set of Experience Knowledge Structure (SOEKS) and Decisional DNA (DDNA): Past, Present and Future, Cybernetics and Systems, vol. 45, pp. 200-215.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, *15*(1), 1929-1958.

Werbos, P. J. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences. *Doctoral Dissertation, Applied Mathematics, Harvard University, MA*.

Zambrano A., C Toro, M Nieto, R Sotaquira, C Sanin, E Szczerbicki, Video Semantic Analysis Framework based on Run-time Production Rules - Towards Cognitive Vision, *Journal of Universal Computer Science*, Vol. 21, No 6, 2015, pp. 856-870.

MOST WIEDZY Downloaded from mostwiedzy.pl