# A Flexible Knowledge-Vision-Integration Platform for Personal Protective Equipment Detection and Classification using Hierarchical Convolutional Neural Networks and Active Leaning

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)

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

(edward.szczerbicki@newcastle.edu.au)

# A Flexible Knowledge-Vision-Integration Platform for Personal Protective Equipment Detection and Classification using Hierarchical Convolutional Neural Networks and Active Leaning

This work is part of an effort to develop of a Knowledge-Vision Integration Platform for Hazard Control (KVIP-HC) in industrial workplaces, adaptable to a wide range of industrial environments. The paper focuses on hazards resulted from the non-use of personal protective equipment (PPE). The objective is to test the capability of the platform to adapt to different industrial environments by simulating the process of randomly selecting experiences from a new scenario, querying the user, and using their feedback to re-train the system through a hierarchical recognition structure using Convolutional Neural Network (CNN). Thereafter, in contrast to the random sampling, the concept of active learning based on pruning of redundant points is tested. Results obtained from both random sampling and active learning are compared with a rigid systems that is not capable to aggregate new experiences as it runs. From the results obtained, it can be concluded that the classification accuracy improves greatly by adding new experiences, which makes it possible to customize the service according to each scenario and application as it functions. In addition, the active learning approach was able to reduce the user query and slightly improve the overall classification performance, when compared with random sampling.

Keywords: Decisional DNA (DDNA), Set of Experience Knowledge Structure (SOEKS), Industrial Hazard Control, Hierarchical Convolutional Neural Networks, Semi-supervised Learning, Active Learning.

## Introduction and Background

Hazards are present in all workplaces and can result in serious injuries, short and long-term illnesses, and death. According to the Safe Work Australia (2017), from the period of 2003 to 2015, over three thousand workers have died in work-related incidents. Additionally, Safe Work Australia (2017) also claims that in 2014–2015, there were over a hundred thousand serious worker' compensation claims related to disease and injuries. These incidents result in productivity loss, human suffering and the cost of billions of dollars to the Australian Economy.

Aiming at reducing incidents related to occupational health and safety, the Australian Work Health and Safety Strategy 2012–2022 has identified the manufacturing industry as a priority to reduce the work-related injuries, illnesses and fatalities. From 2003 to 2015 the manufacturing sector had the third highest incidence rate of serious claims and the fourth highest proportion of deaths (Safe Work Australia, 2012).

Hazard control is essential to ensure the safety of workers and occupational health (Safetycare Australia, 2015). In this context, monitoring of workers activities and identifying any risk present emerged as a need. The use of sensors data and computer vision techniques support the fast and automated detection of potentially dangerous conditions. This information can be used to provide feedback, and manage workers behaviour to perform the work in a safe manner (Han & Lee, 2013). However, as stated by Little at al. (2013) there is no such flexible system at the moment which is able to function in a broad industrial environments without the necessity of rewriting most of existing application codes each time the circumstances or conditions change. In addition, as the system grows the scalability becomes an issue, and the quantity of considerations required for each application in particular can easily lead to a rigid and inflexible structure (Zambrano, Toro, Nieto, Sotaquirá, Sanín & Szczerbicki, 2015).

This work is part of a research effort to develop a flexible Knowledge-Vision Integration Platform for Hazard Control (KVIP-HC) in industrial workplaces, attending a wide range of industrial environments (de Oliveira, Sanin & Szczerbicki, 2017). In this system, classifiers, feature images, and any context information collected through the platform, are represented explicitly by the Set of Experience Knowledge Structure (SOEKS), grouped and stored as Decisional DNA (DDNA) (Sanín & Szczerbicki, 2005; Sanín & Szczerbicki, 2007). The collected knowledge is used for reasoning and retraining of the system from time to time, customizing the service according to each scenario and application.

This paper focuses on hazards resulted from the non-use of personal protective equipment (PPE). It aims to test the capability of the platform to adapt to different industrial environments by simulating the process of randomly selecting experiences from a new scenario, querying the user, and using their feedback to re-train the system over five iterations. In addition, we apply the concept of active learning based on pruning of redundant points to reduce the user query. We compare the results obtained from random sampling and active learning with a rigid system that is not capable to aggregate new experiences as it runs. To train the system, a hierarchical recognition structure using Convolutional Neural Network (CNN) is proposed, where in the first layer a CNN binary classification is performed to detect the existence or not of PPEs. If a PPE is detected a multi-class CNN is used to classify the PPEs into Boots, Gloves and Hats in the second layer. Other market PPEs will be included in the future work, as well as a broader range of hazardous activities.

The rest of the paper is organized as follows. In the "Knowledge-Vision Based Approach" section, a background about vision and knowledge-based systems is introduced. The section "Convolutional Neural Network" presents the concept of Deep

Neural Networks, including the Convolutional Neural Network proposed. The semi-supervised approach for collection of experiences from user feedback, including random sampling and active learning, is described in “Experience Selection” section. The “Methodology” section explains the Hierarchical Recognition Structure (HRS), the process of creating datasets, and describes the training and test process. The section “Experimental Results” presents the classification performance for the three approaches analysed in the paper: (a) when no experiences are added to the system, (b) when a randomly selected experiences are used to query the user and their feedback is applied to re-training of the system, and (c) when a randomly selected set of experiences are collected and only the least redundant are used to query the user. Conclusions and future work are presented in the last section.

## Knowledge-Vision Based Approach

The use of sensors data and computer vision techniques support automatic detection and tracking of workers indicating potential dangerous situations. However, the current sensor-based technologies used for this purpose require the devices or markers to be attached on the human body, which can disturb some actions (Han & Lee, 2013). On the other hand, visual sensing facilities, such as video cameras, can monitor workers behaviour and environment conditions without any disturbances. Furthermore, the generated data, such as video sequences or digitized visual data can be processed in powerful computers in real time (Chen, Hoey, Nugent, Cook & Yu, 2012). For the reasons described above, computer vision approach has been a research focus for a long time in surveillance systems, human detection, and tracking. However, the accuracy of current systems when operating in real life scenarios, subject to change in illumination, variation in backgrounds, noise and different camera resolutions, still remains a challenge

(Mosberger, Andreasson & Lilienthal, 2013). In addition, sensor data and computer vision technologies are, mostly, not scalable and lack adaptability to broad industrial environments and existing situations. As a result, the existing technologies create case-based applications that work only for specific circumstances and any change in conditions would result in rewriting most of the application code.

In this context, methods incorporating prior knowledge and context information are gaining interest. The visual and context knowledge can be used for reasoning and customizing the service along with each workplace and application. Also, with additional contextual information it is possible to enhance the speed and accuracy of the detection algorithm and reduce scalability issues (Davis, Shrobe & Szolovits, 1993). For instance, an automatic semantic and flexible annotation service able to work in a variety of video analysis with little modification to the code using Set of Experience Knowledge Structure (SOEKS) was proposed in Zambrano, Toro, Nieto, Sotaquirá, Sanín & Szczerbicki, 2015. Though, a lot of information contained in the classifiers was lost by separating the classification of the humans and objects from the event recognition.

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

SOEKS is a knowledge representation structure designed to obtain and store formal decision events in an explicit way. It is based on four key basic elements of decision-making actions (variables, functions, constraints and rules). Variables are generally used to represent knowledge in an attribute-value form, following the traditional approach for knowledge representation. Functions, constraints, and rules of SOEKS are different ways of relating knowledge variables. Functions define relations between a dependent variable and a set of input variables; therefore, SOEKS uses functions as a way to create links among variables and to build multi-objective goals. Likewise, constraints are functions that act as a way to limit possibilities, restrict the set of possible solutions, and control the performance of the system in relation to its goals. Finally, rules are relationships that operate in the universe of variables and express the condition-consequence connection as “if-then-else” and are used to represent inferences and associate actions with the conditions under which they should be implemented (Sanín & Szczerbicki, 2004, 2005, 2007, 2008).

The repository is composed of experienced decision events, which are represented as SOEKS in our platform, grouped according to the areas of decision categories – the Decisional DNA. A SOEKS from a formal decision event represents a portion of an organization’s DDNA (as a gene that guides decision-making). This gene belongs to a decisional chromosome from a certain type or category. A group of chromosomes from different kinds (e.g. hazard control decisions, production decisions, marketing decisions) comprise the DDNA, a decisional genetic code of an organization. The DDNA can be used to solve the scalability issues found in current vision-based approaches by introducing an experience-based approximation that aims to recognize events defined by the user using production rules, adaptable for different conditions, clients and situations (Sanín & Szczerbicki, 2007).

## Convolutional Neural Network

Neural networks are a very powerful Machine Learning (ML) classifier applied to both binary and multi-class problems. The first computational model for neural networks was based on mathematics and algorithms called threshold logic (Warren & Pitts, 1943). In 1965, the first functional networks with many layers (Deep Neural Network DNN) was proposed. However, at that time the computers didn't have enough processing power to effectively handle the work required by large neural networks (Ivakhnenko &

Grigor′evich, 1967). A key trigger for the renewed interest in neural networks and learning was the backpropagation algorithm that accelerated the training of multi-layer networks (Werbos & Beyond, 1975). Another technique called dropout used to reduce overfitting (a very common problem in deep networks) gave major improvements over other regularization methods (Srivastava, 2014).

To learn all the PPEs existing in the market and extend the model to an immense complexity of possible risky activities, a model with a large learning capacity is needed. Convolutional Neural Networks (CNNs) constitute one such class of models. CNN is a class of deep, feed-forward artificial neural networks (Ciresan, Meier, Masci, Gambardella & Schmidhuber, 2011). Supervised deep learning method was the first artificial pattern recognizer to achieve human-competitive performance on certain tasks (Ciresan, Meier & Schmidhuber, 2012). In addition, when compared to standard feedforward neural networks with similarly-sized layers, CNNs have much fewer connections and parameters and so they are easier to train (Krizhevsky, Sutskever & Hinton, 2012).

**Selection of Experiences**

Finding a way to choose informative images to ask the user to label them can be challenging. Such request for the label of an image is called a pool-query (Tong & Chang, 2001). Most machine learning algorithms are passive in the sense that they use the most fundamental and simple way to select a data subset for manual labelling: through sampling a small number of data samples from the pool of unlabelled data randomly. Random sampling is simple and fast to implement, but the points selected by random sampling are not guaranteed to be optimal. Therefore, for the performance of a classifier to reach a certain level, a significant number of data samples should be sampled (Woo &

Park, 2012). It can result in the oracle (typically a human) being required to spend extensive time to label the classes of the instances, increasing in memory required to store the data and rising the computational costs of training.

### *Active Learning*

Active learning (also known as “query learning” or “optimal experimental design” in statistics) is a subfield of semi-supervised machine learning. Its key hypothesis is that if the learning algorithm is allowed to choose the data from which it learns, it will perform better with less training (Settles, 2010). An active learning scenario involves evaluating the informativeness of unlabelled data, and many ways of formulating these query strategies have been proposed in the past few years such as: uncertainty sampling, query by committee, expected model change, expected error reduction, variance reduction, balance exploration, and Exponentiated Gradient Exploration (EGE) for Active Learning (Settles, 2010).

In contrast, when only unlabelled or very little labelled data is given for training an initial classifier, most of the existing active learning methods are difficult to apply. In this context, Woo and Park (2012) suggest an optimal experimental design in that the goal is to select a subset of samples on which an optimal classification model is trained. In this article a similar approach is proposed. However, differently from Woo and Park (2012) method, a dataset containing a small set of samples are initially accompanied with class labels.

## Methodology

In this section, (i) the HRS used for detection and classification of the PPEs are presented; (ii) the process of creating the dataset is explained; and (iii) the training and test process is described. The codes to perform the data-preparation, training, tests and analysis of the methodologies were developed in Python 3 and Jupyter Notebook web application platform (Van Rossum, 2009; Kluyver, Ragan-Kelley, Pérez, Granger, Bussonnier, Frederic & Ivanov, 2016).

### *Hierarchical Recognition Structure*

There is no trivial solution existing for a multi-class problem when one of the classes represents the absence of all the other classes. The proposed system must be able not only to classify a PPE, but also indicate if no PPE is present. In this case, the *"every else"* feature space is unbounded, and it cannot be described by a limited dataset in a multi-classification approach. To overcome this problem and avoid the training of binary classifier for each desired PPE (which could lead to scalability issues as the system grows), a hierarchical recognition structure is proposed. In this structure, shown in Figure 1, the hierarchical division of the output space, i.e. the classes, are arranged into a tree. The tree created is such that the classes at each parent node are divided into a number of clusters, one for each child node. The process continues until the leaf nodes contain only a single class. At each node of the tree, a simple classifier, usually a binary classifier, makes the discrimination between the different child class clusters (Aly, 2005). In the proposed platform, the binary classifier to detect the presence or not of PPE is combined with a multi-class classifier for classification in case of PPE detected.

Each layer of the recognition system is composed by a CNN implemented with random regularization to prevent overfitting (CNN Dropout). To build the networks and train the classifiers the TensorFlow library, an open source library developed to conduct machine learning and deep neural networks research, was utilized (Abadi et al., 2016).

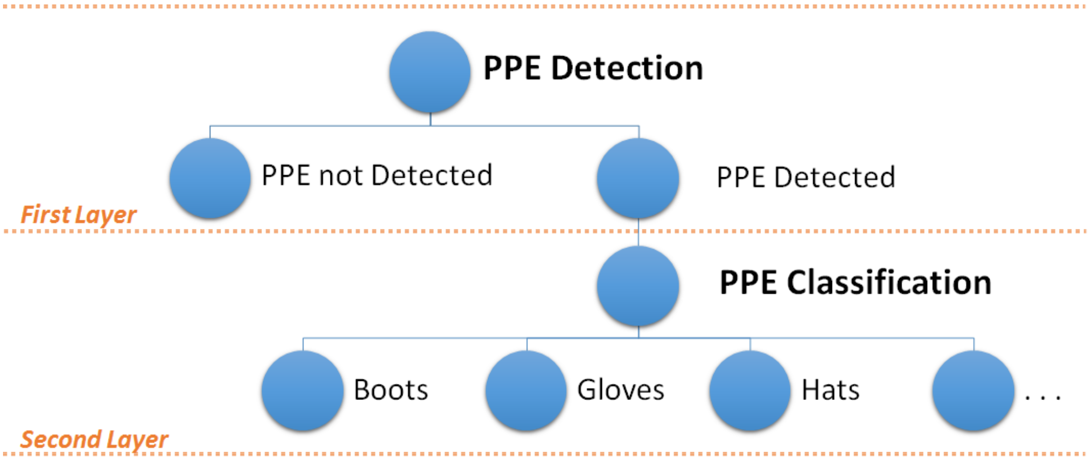


Figure 1. Hierarchical Recognition Structure.

***Dataset***

For the creation of the samples that compose the positive dataset of the first layer and the Gloves, Boots and Hats dataset of the second layer, 18 videos with different sizes and resolutions were used (containing diverse images of boots, gloves and hard hats). These videos were separated into frames, and a set of frames containing the objects of interest was selected randomly. From each selected frame, a Region of Interest (ROI) comprising a PPE was extracted, resized (28x28 pixels), and saved as a positive sample. The video processing flow is present in Figure 2.

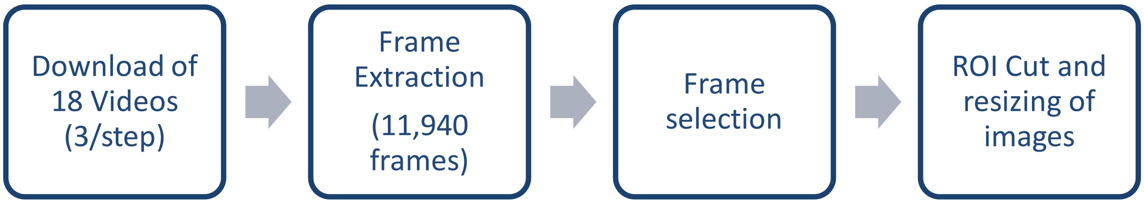


Figure 2: Video processing flow.

For the first layer, samples from every three videos resulted in a different positive dataset which were distributed among the initial step and the following five iterations. Therefore we simulate a full change of scenario and PPE characteristics in each step, and a low accuracy is expected for the system when it starts every new iteration. Images of random objects and scenes were also resized to 28x28 pixels to compose the negative dataset of each step. These datasets were split into training and test sub-sets.

For the second layer, a set of samples from each video resulted in a different dataset which was labelled as Boots, Gloves or Hats and distributed among the initial step and the five consecutive iterations. These datasets were also subdivided into training and test. Figure 3 shows the distribution of samples along each iteration for each layer of the recognition system.

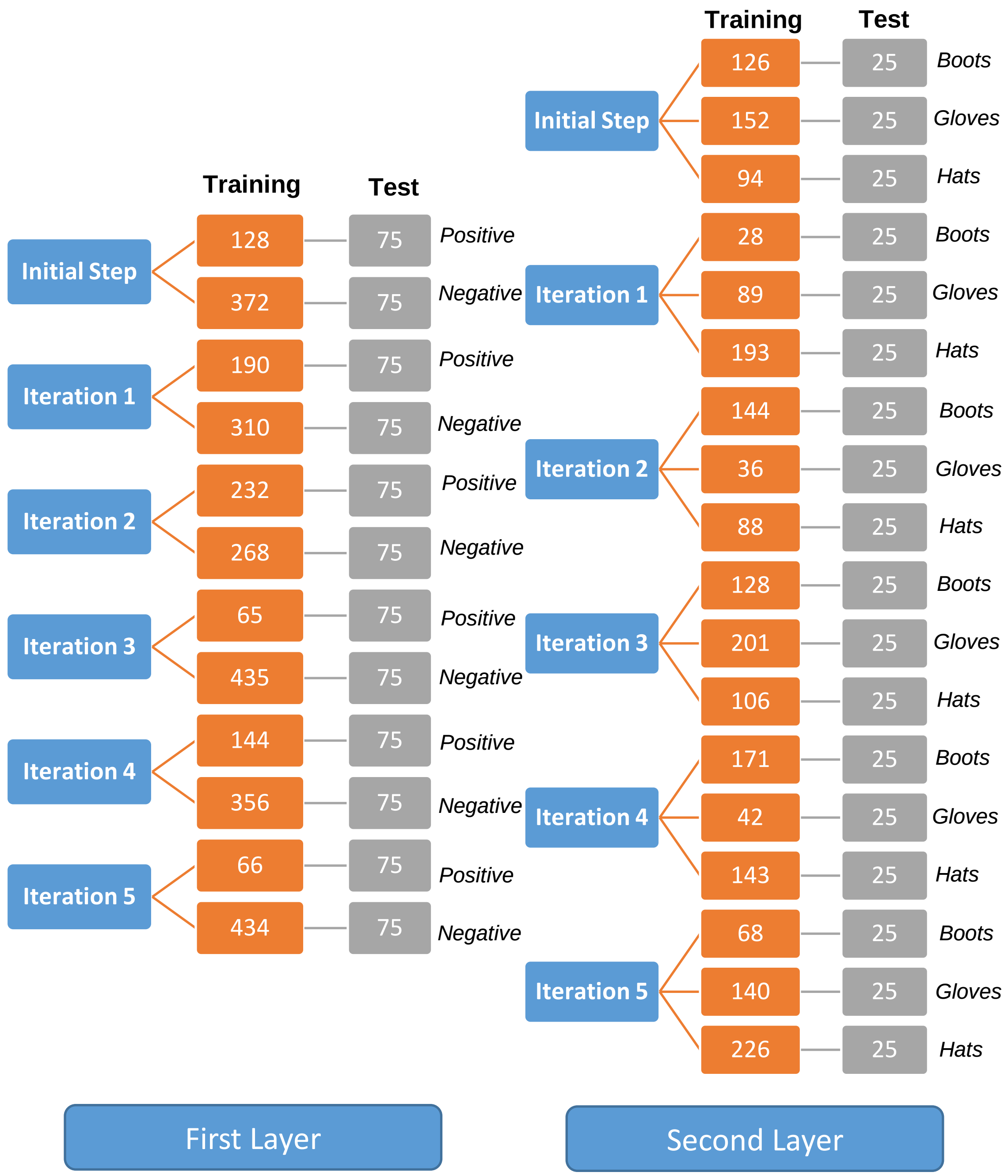


Figure 3: Distribution of samples along each iteration for each layer of the recognition system.

***Training and Test***

The experiment is performed over five iterations starting from an initial dataset as shown in Figure 4. Three approaches are tested:

- No Learning: The classifier trained in the initial dataset is tested on the next iterations without aggregation of experiences and retraining of the system in each iteration;
- Random Selection: For each iteration, the randomly selected experiences are used to query the user and retrain the system. The test is performed on the same and next iteration after retraining the system;
- Active Learning by pruning: For each iteration, the randomly selected experiences are compared among each other. The most redundant samples are deleted by searching for a pair of data samples having the smallest distance and removing one of them from the subset. Cosine distance is used as a distance measure (Qian, Sural, Gu & Pramanik, 2004). The remaining samples are used to query the user and retrain the system. The test is performed on the same and next iteration after retraining the system.

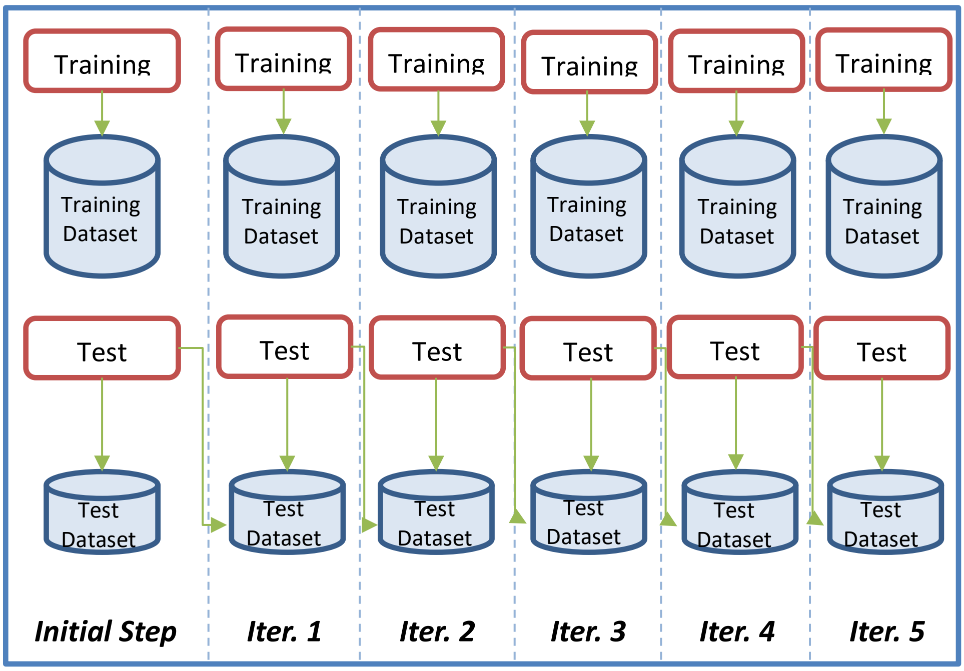


Figure 4: Iterations of the experiment

**Experimental Results**

The CNN was built using *Softmax* regression and cross-entropy to compute the loss of the model that is used. To minimize the loss, *ADAM optimizer* is applied in a learning rate of *0.01*. In addition, weights and biases are created. The pooling is a max pooling over 2x2 blocks. The convolutional layer has patch size of 5x5 with 3 input channels and 32 output channels. In the fully-connected layer, there are *1024 neurons* (to allow processing on the entire image, which is reduced to 14x14), that feed the readout layer together with the number of classes to predict. Finally, the regularization function is applied before the readout layer to prevent overfitting. The training is performed over the whole dataset allocated for each iteration and the best accuracy for test dataset out of 10 epochs is chosen. For each iteration, the neural network is used for detection of the PPEs in the first layer and classification of them in case of detection in the second layer.

Figure 5 shows the results for accuracy for the first and second layer when no experiences are added along the iterations.

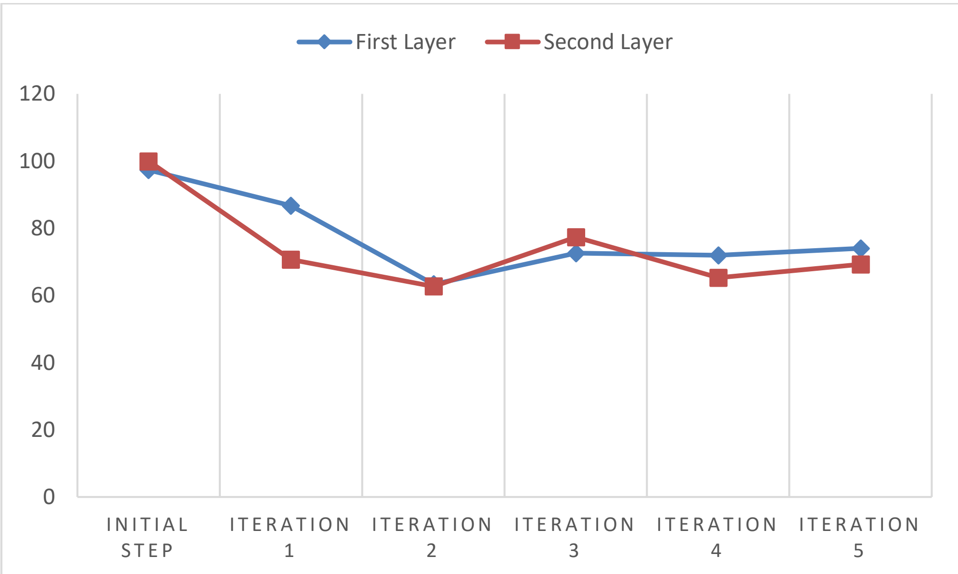


Figure 5: Results for accuracy for the first and second layer when no experiences are added along the iterations.

As can be observed in Figure 5, the accuracy drops significantly from the initial step and as no new experiences are added to the system it has no capacity to learning and improve its performance for next iterations. The results for the first and second layer of random selection and active learning, and test on the same and next iteration are present in Figure 6 and Figure 7, respectively.

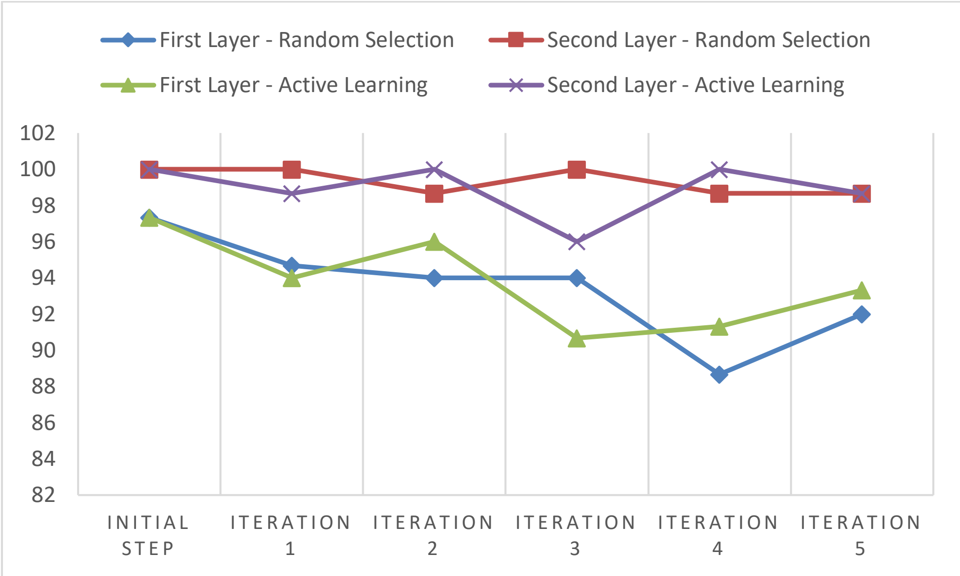


Figure 5: Results for the first and second layer of random selection and active learning and test in the same iteration.

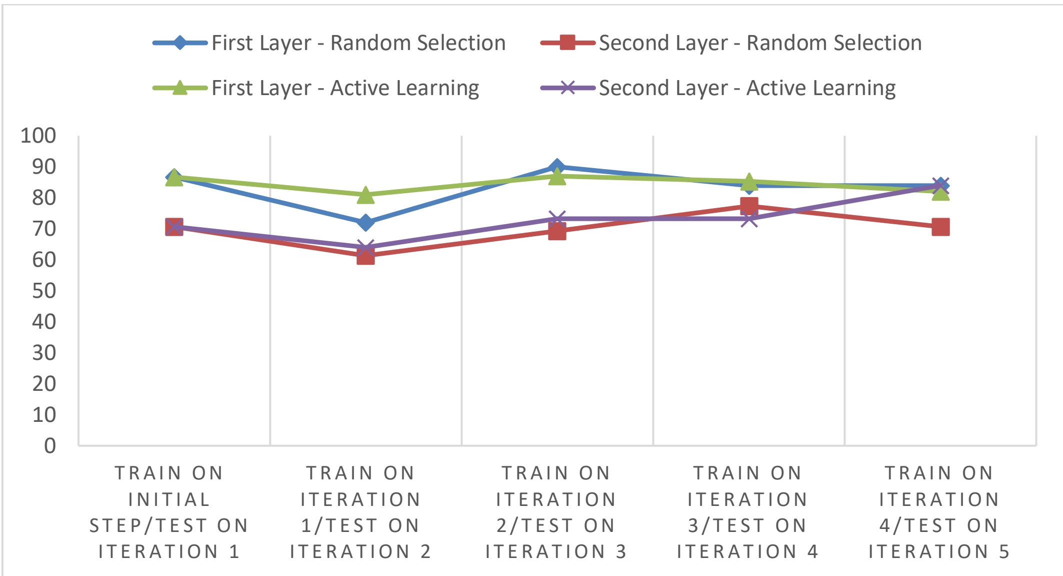


Figure 6: Results for the first and second layer of random selection and active learning and test in the next iteration.

Figure 5 and 6 show that, both random selection and active learning, for testing on the same and the next iteration, the second layer classifiers perform better in terms of accuracy when compared with the multi-class classifiers of the second layer. In addition, the pruning of samples in the active learning approach did not result in overall decrease of the classifiers accuracy as can be seen in Table 1.

**Table 1:** Average accuracy of classifiers over the iterations.

| | **Test on same Iteration** | | **Test on next Iteration** | |
|---|---|---|---|---|
| | First Layer | Second Layer | First Layer | Second Layer |
| **Random Selection** | *93%* | *99%* | *83%* | *70%* |
| **Active Learning** | *94%* | *99%* | *84%* | *73%* |

In fact, the average of accuracy of the classifier when using active learning is slightly better for both layers.

In the end of all iterations, 2,175 samples composed the training dataset for random sample and 1,697 for the active learning. It represents 22% decreasing in samples, saving memory consumption to store the experiences. The average of training time over the 10 epochs drops when active learning is reduced by 20% which shows that active learning also reduces the computation costs to re-train the system for reducing the samples. Finally, the classifiers size are reduced by 30% when active learning is used.

## Conclusion and Future Work

This study demonstrated the enhanced ability of CNN to solve classification problems, and the effective use of Active Learning to select the experiences by pruning the most redundant samples. The use of Active Learning reduces the user query and the memory required to store the data. It can also reduce computational costs of training procedure as the system grows, without reducing the classification performance. In fact, on average the system performs better when tested for current and next iteration when using Active Learning. In addition, poor results obtained with no new experiences added to the system along the iterations, demonstrates the need of re-training the platform to ensure its flexibility to function in diverse environments and its capability to adapt to different scenarios and circumstances as it runs.

In future work, the PPE classification will be expanded, and hazardous activities in industrial settings considered. In addition, in the further studies the concept of proactive learning will be explored and the assumption of indefatigable (always answers), infallible (never wrong), and insensitive to costs oracle will be relaxed.

**References**


Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016, November). TensorFlow: A System for Large-Scale Machine Learning. In *OSDI* (Vol. 16, pp. 265-283).

Aly, M. (2005). Survey on multiclass classification methods. *Neural Netw, 19*.

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.

Ciregan, D., Meier, U., & Schmidhuber, J. (2012, June). Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 3642-3649). IEEE.

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, September). Hazard Control in Industrial Environments: A Knowledge-Vision-Based Approach. In *International Conference on Information Systems Architecture and Technology* (pp. 243-252). Springer, Cham.

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.

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

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

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., ... & Ivanov, P. (2016, May). Jupyter Notebooks-a publishing format for reproducible computational workflows. In *ELPUB* (pp. 87-90).

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).

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.

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.

Qian, G., Sural, S., Gu, Y., & Pramanik, S. (2004, March). Similarity between Euclidean and cosine angle distance for nearest neighbor queries. In *Proceedings of the 2004 ACM symposium on Applied computing* (pp. 1232-1237). ACM.

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

Safe Work Australia. (2012). *Australian Work Health and Safety Strategy 2012–2022*. Creative Commons.

Safe Work Australia SWA. (2017). Disease and injury statistics. Retrieved from: https://www.safeworkaustralia.gov.au/collection/australian-workers-compensation-statistics

Sanin, C., and E Szczerbicki, (2004), Knowledge Supply Chain System: a conceptual model, in *Knowledge Management: Selected Issues*, A Szuwarzynski (Ed), Gdansk University Press, Gdansk pp. 79-97, 2004

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

Sanin, C. and E. Szczerbicki (2008). "Decisional DNA and the smart knowledge management system: a process of transforming information into knowledge." , in *Techniques and Tools for the Design and Implementation of Enterprise Information Systems*, Gunasekaran A., (Ed.), IGI: New York, 2008, pp. 149-175.

Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison*, *52*(55-66), 11.

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.

Tong, S., & Chang, E. (2001, October). Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia* (pp. 107-118). ACM.

Van Rossum, G., & Drake, F. L. (2009). *Python 3: Reference Manual*. SohoBooks.

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

Woo, H., & Park, C. H. (2012, October). An efficient active learning method based on random sampling and backward deletion. In *International Conference on Intelligent Science and Intelligent Data Engineering* (pp. 683-691). Springer, Berlin, Heidelberg.

Zhu, X. (2011). Semi-supervised learning. In *Encyclopedia of machine learning* (pp. 892-897). Springer US.