Long-Term Yield Prediction of Greenhouse Sweet Pepper Crops

Reem Al-Halimi, and Medhat Moussa

***Abstract*— In this paper, a new model for predicting the yield of greenhouse sweet peppers (*Capsicum annuum L.*) is presented. The model can provide long-term prediction up to 7 weeks in advance with the same accuracy it can predict yield one week in advance. It uses both past and expected environmental readings as well as physiological data as input to a specially designed artificial neural network. The model was tested using 4 years of data that was obtained from commercial pepper growers. Short-term prediction accuracy (one week) is consistent with other predictive models in the literature for sweet peppers. This validates our long-term results.**

Keywords- Bell peppers, crop models, greenhouse, long-term yield prediction, neural networks

I. INTRODUCTION

Accurate prediction of greenhouse fruit yield has been of significant interest in the last several years. Some of these efforts are based on a deeper explanation of a plant's physiology such as TOMSIM and TOMGRO [1]. These are called *explanatory* models. Others, are descriptive models which do not specify the internal mechanisms within a plant and rely instead on external parameters, such as radiation levels, temperature, or past yield in simulating yield patterns thus implicitly capturing hidden interactions between the different parameters [2]–[6]. Such parameters are readily available for growers, or can be easily obtained, making descriptive models more attractive for commercial deployment. These models generally consist of input parameters, a predictive model, and an output that represents yield for a given week or the expected overall yield for the season. They can be based on simple approaches such as linear regression analysis [5] or more complex approaches such as neural networks [7]–[9], neuro-fuzzy networks [6], and time series [4]. The selection of parameters and the type of predictive algorithm used impacts prediction accuracy and how far ahead predictions can be made.

Since different crops have different challenges and physiological properties, the crop under investigation and its environment can also affect prediction accuracy. Sweet Peppers (*Capsicum annuum L.*), in particular, are challenging because of their flushing property (Heuvelink and Marcelis, 2004). Flushing means that pepper crops alternate between high and low yields throughout the season making predictions more challenging. While most pepper yield prediction models report decent accuracy for up to two weeks in advance, predictions further ahead have proven to be difficult. Lin *et al*. experimented with yield prediction from 1- 4 weeks ahead. They used radiation levels, current and past yields, temperature and week number to predict the yield. They reported that the models' accuracy was high (30%) when predicting the weekly yield up to two weeks in advance, but that the prediction quality degraded significantly at the four week mark. Sauviller *et al*. [5] suggested a simple linear regression model relating the average 24 hour greenhouse temperature to the number of days to maturity for new fruit sets. They used this model to predict the number of fruits that are harvest ready in a given week. While they reported high accuracy (7%-14%), their model is iterative where the accuracy of prediction increases as the prediction time is shortened. The model also does not predict yield but fruit numbers. However, there is a significant variation between fruit numbers and fruit weights.

This paper presents a model that predicts weekly yield 4 and 7 weeks in advance with the same accuracy as predicting weekly yield 1 week in advance. To our knowledge, no other model can provide such accuracy on a long-term basis using commercial greenhouse settings. The paper is organized as follows: Section 2 describes the parameters used and the data preparation methods. Section 3 outlines the neural network model, as well as the training and testing methodology. Finally, in section 4 we present results and discussion.

II. METHODOLOGY

A. Data source

Data was collected from a commercial pepper greenhouse grower in the Chatham area, Ontario, Canada. The data covers four years 2007, 2008, 2010 and 2011. Sweet pepper cultivars grown varied from year to year. In 2007, the cultivars were 'Fascinato' and 'Red Glory'; in 2008 and 2011 it was 'Besalga', while in 2010 the cultivar grown was 'Viper'. Planting dates were 6 December 2006, 3 December 2007, 5

DOI: 10.5176/2345-7848_2.1.11

December 2009 and 14 and 15 December 2010 for the growing seasons 2007, 2008, 2010, and 2011, respectively. Plants were grown in a hydroponics system in rockwool slabs. The first fruits appeared in January for the growing seasons 2007, 2010, and 2011, and at the end of December for 2008. The harvest period started in March for all growing seasons and ended in November.

In total, there were 143 data points over all 4 years excluding rows from early and late season weeks with zero weekly yield. The data consists of automatically collected daily environmental readings both inside and outside the greenhouse as well as physiological readings related to specific zones within the greenhouse as follows:

- Environmental readings included radiation level, outside temperature, average 24 hour temperature, CO_2 levels, day and night daily humidity readings inside the greenhouse, irrigation-related readings including the number of irrigations and amount of water absorbed by the plant (in L/m^2).
- Physiological readings related to 10 sample plants from each of 4 areas in the greenhouse were also collected daily and included: plant length (in cm), plant growth per week, number of flowers, number of new fruit sets (1cm or larger), total fruit load (includes unripe fruits), and the number of fruits harvested daily from the plant. In 2010 there were 10 sample plants in each of 6 areas within the facility instead of the usual four.

- Select the min yield *Min*, max. yield *Max*, and accuracy a
- Divide yield values between *Min* and *Max* into ranges $[l, m)$ so that $\frac{(l-m)}{l} \leq a$
- Map each yield value in the data to the appropriate yield category c with range $[l_c, m_c)$ so that $\underline{l_c} \leq v < m_c$
- The yield values are used in training and testing the ANNs. For each prediction i made by the ANN,
  - select the category c_i with the highest likelihood as the ANN's yield prediction for case i.
  - represent the ANN's prediction for case i by the center of the selected category c_i.

Figure 1. A general algorithm for replacing yield values by approximate values given a desired accuracy *a*, a minimum yield value *Min*, and a maximum yield value *Max*.

B. Data representation and preprocessing

Most of the previous efforts using artificial neural networks (ANN) for yield prediction, formulated the problem as a regression problem predicting yield values as output. Yet farmers are not interested in a very precise model but rather an approximate yield prediction, and would accept a reasonable error in the predicted value. As such, the problem was formulated in this paper as a classification problem. Rather than predicting an exact yield value, the model predicts a range of values within which the yield is expected to fall. This approach also entails grouping similar values together into a single yield category thus increasing the number of training examples available to the network for each value.

A brief algorithm for creating yield categories is shown in Figure 1. The process starts by setting a minimum yield value and selecting an accuracy value $\boldsymbol{a}$ between 0 and 1. This accuracy should reflect the error that farmers find acceptable in the predicted yield value. An accuracy value of 0.3, for example, means that the predicted yield can vary ±30% from the predicted value. Given the minimum yield, the accuracy *a*, and the maximum possible yield value, yield values are then divided into sub-ranges $[l, m)$ so that each range has a minimum value *m* and a maximum value *l* where:

$$\frac{(l-m)}{l} \leq a \qquad (1)$$

For example, for an accuracy of 0.3, a minimum yield value of 0.11 and a maximum yield value of 1, the range of values between 0.11 and 1 is divided into the ranges: [0.11, 0.143), [0.143, 0.186), [0.186, 0.242), *etc.* Each range is then considered a category with its category center at (l+m)/2. Once categories are created, yield values in the data are mapped into the appropriate yield categories. The new yield categories are then used in training and testing the ANN. Instead of predicting exact yields, the output of the ANN in this case is a set of values each representing the likelihood of the expected yield to belong to each of the given categories (yield ranges). The category $\boldsymbol{c}$ with the highest likelihood is selected as the predicted yield category.

In the final step, each category c is represented with its center value: (l_c+m$_c$)/2 where l_c and m_c are the largest and smallest values within category c. In the rest of the paper we will refer to the center of the predicted category as *predicted yield* and the center of the target yield's category as simply the *target yield*.

To make the task of building effective neural networks more efficient, the environmental variables utilized were reduced to include only those commonly found in yield prediction ANN models such as those used by Lin *et al.*[7], Ehret *et al.*[9] and Sauviller *et al.*[5]. In particular, the following environmental variables were included:

- average radiation levels,
- 24 hour temperature,
- day humidity,
- average CO_2 levels.

Some physiological readings were also used including:

- the average plant absorption level,

- plant growth,
- the number of new fruit sets.
- the calendar week number as an indicator of plant age.

In addition to the raw environmental and physiological data collected, a set of environmental reading averages was also developed to be used in some of the predictive models as will be explained later.

C. Neural network modeling

The Matlab Neural Network Toolbox was used to build and test the neural networks used in this paper. Input data was first preprocessed using principle component analysis to reduce dimensionality thus enhancing generalization, especially given the limited data sets available. Training and testing were conducted using a 4-fold cross-validation process where the testing set was selected as an entire year and the remaining three years were used in training. After four training/testing cycles, the average error over all 4 years was used to report the results.

A multi-layer perceptron ANN (MLP) was used in all learning. The network was training with Matlab implementation of the scaled conjugate gradient back-propagation method (MLP-BP). Both topology and threshold for stopping the training were varied during experimentation.

The performance of a MLP-BP is impacted significantly by its topology. The topology should not have a large number of weights relative to the number of training patterns, otherwise over-fitting will occur and the network will have poor results on the testing patterns. On the other hand, a small topology may not have the computational complexity to learn the target function, which leads to poor results. Since the actual complexity of the learning task is unknown, a large number of topologies were tested to find those that provide the best results. Networks with two to eight hidden neurons were built forming seven different topologies. With each topology, we experimented with six different thresholds to stop the training automatically. The threshold was based on Mean Squared Error over each training iteration (which included the entire training dataset). For each threshold we also ran 10 different models, each with a different random selection of initial weights. In total 1680 different neural network models were tested for each experiment described below.

D. ANN Model Selection

As explained above, the neural networks predict yield as centers of yield categories. Yields that were recorded by farmers were also converted into centers of yield categories. The error in prediction was then measured by comparing the predicted yield, i.e. the center of the predicted yield's category, to the target yield, i.e. the category center of the recorded yield's category.

For each run two values were used to measure the model's effectiveness: the correlation coefficient *r* and the weekly prediction error *WPE*. The *WPE* was defined as the ratio of the difference between the predicted yield and the target yield to the target yield:

$$WPE = \frac{predicted_{yield} - target_{yield}}{target_{yield}} \quad (2)$$

To select the best ANN models, the two measures, *r* and *WPE* were combined in one function called the *combined error* (*ce*):

$$ce_m = \left(1 - AVG(WPE_m)\right) + r_m \quad (3)$$

where $Avg(WPE_m)$ is the average weekly prediction error over the whole growing season for the model built in run *m*, and r_m is the correlation coefficient for the same model *m*. The best model was defined as the model with the highest *ce* value.

Each of the four cross-validations was represented by the best model from among all its threshold-run combinations. Given the four best ANNs for each topology *t*, one per cross validation c_i, the relative prediction error $AvgPE_t$ per topology *t* was defined as the average *WPE* over the four selected models. Similarly, the correlation coefficient r_t per topology was calculated as the average r_m over all four models.

III. Results and Discussion

A. Experiment Overview

Several models were tested. Each model represented one case of input parameters used or a different prediction period. Table 1 shows all the cases modeled and the parameters used in each case.

We explored using new input parameters not used before such as the expected (future) environmental readings for a given time of the year, and how prediction accuracy would be affected by the length of time between the target week for which the yield was being predicted and the current week from which the prediction was being made. Table 1 lists all the cases considered in our experiments. The base case (case 1) predicted next week's yield given the weekly average 24 hour radiation level, 24 hour temperature, plant water absorption rate, CO_2 levels, average day humidity, and the average plant growth per sample plant with and without the calendar week number. Next, the number of fruit sets that formed six weeks earlier was added to test the effect of this knowledge on short-term yield prediction (case 2). Following that the average environmental data from the past six weeks was added to test the effect on yield prediction (case 3). Cases 4, 5, and 6 looked at the change in yield prediction errors when long-term prediction came into play as shown Table 1. These three cases examined the effectiveness of long-term prediction under two varying conditions: the length of time between the current week and the target week for which prediction was sought, and the use of expected environmental data given the time of the year.

B. Results

Table 1 shows the best prediction errors $AvgPE_t$ and those with the highest correlation coefficient values for each case in the experiments, as well as the topologies that achieved these values. In what follows we look more closely at the result of each case as it compares to the other cases.

TABLE I. THE DIFFERENT CASES EXAMINED FOR YIELD PREDICTION[a]

	Experiment	Variables used	Lowest $AvgPE_t$	Topology	Highest r_t	Topology
1a	Predicting next week's yield	Current week's environmental readings.	0.43 (0.06)	8	0.41 (0.1)	5
1b		Same variables as case 1a. For case 1b we also use the current calendar week number.	0.4 (0.07)	3	0.47 (0.12)	8
2a	Effect of using new fruit set numbers in predicting next week's yield	Same as in case 1a with the number of new fruit sets that appeared six weeks ago.	0.45 (0.04)	6	0.46 (0.17)	4
2b		Same variables as case 2a. For case 2b we also use the current calendar week number.	0.44 (0.03)	5 and 6	0.5 (0.2)	7
3a	Effect of using past average environmental readings on yield prediction	Average readings for the current week and the past six weeks. Number of new fruit sets 6 weeks ago	0.43 (0.03)	6	0.48 (0.09)	8
3b		Same variables as case 3a. For case 3b we also use the current calendar week number.	0.4 (0.05)	7	0.4 (0.08)	6 and 7
4a	Effectiveness of yield prediction seven weeks early with expected environmental readings	Target week number. Number of new fruit sets for this week. Expected average environmental readings for the next seven weeks.	0.28 (0.03)	2	0.53 (0.12)	7
4b		Current calendar week number since transplanting. Number of new fruit sets for this week. Expected average environmental readings for the next seven weeks.	0.42 (0.04)	2 and 7	0.56 (0.09)	7
5a	Yield four weeks from now	Target calendar week number. Number of new fruit sets that formed three weeks ago. Average environmental readings for the past four weeks.	0.32 (0.04)	2	0.47 (0.04)	8
6a	Yield four weeks from now with expected environmental readings	Same as in case 5a above. Expected environmental readings for the next four weeks.	0.38 (0.07)	6	0.46 (0.05)	6

1) *One to two week prediction accuracy*

Case 1a was the base case and it predicted next week's yield using the current week's environmental readings as shown in Table 1. The best prediction error $AvgPE_t$ in this case was 0.43 which was among the highest (worst) for all cases. Varying the parameters used for short-term prediction did not make significant improvements in the average prediction accuracy as can be seen in cases 1a, 1b, 2a, 2b, 3a and 3b which added a variety of parameter combinations including the calendar week number, the number of new fruit sets, and the average past environmental readings. The prediction errors for these cases ranged between 0.4 and 0.45 which is very close to case 1a's.

2) *Four week prediction accuracy*

As the prediction term increased and other variables were utilized, prediction errors decreased. Cases 5a and 6a predicted yield four weeks in advance. Both cases had better prediction errors than those for cases 1a, 1b, 2a, 2b, 3a and 3b whose models predicted next week's yield. In particular, cases 3a and 5a both used average past environmental readings and the number of new fruit sets observed. But case 3a predicted next week's yield while case 5a predicted yield four weeks in advance. Yet, at 0.32, case 5a's prediction error was 25% lower than that of case 3a. Interestingly, increasing the prediction term in this case from one week for case 3a to four weeks in case 5a and decreasing the number of weeks used for environmental readings from 7 weeks for case 3a to only four weeks in case 5a did not jeopardize prediction quality. This might indicate that not all past weeks are equal in their strength

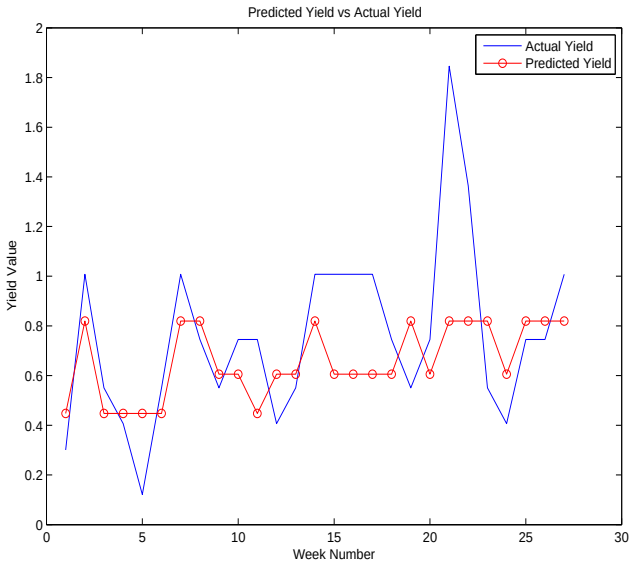

Figure 1: The predicted yield values versus the actual recorded yield for a case 1b model which predicts next week's yield.

as indicators of a plant's response in terms of yield production. This is consistent with Lin et al. [7] observation that some environmental factors have more influence when they are closer to the prediction week and that this influence lessens as we go further back from the prediction week.

Adding the expected environmental readings slightly worsened the best prediction value in case 6a when compared to case 5a but this value remained better than those observed for shorter term predictions.

3) Seven week prediction accuracy

When past environmental readings were removed altogether in case 4a and only expected environmental readings and the number of new fruit sets were used to make yield predictions seven weeks in advance, the prediction error reached its best value of 0.28, a 35% improvement over the base case.

Figures 2 and 3 plot the predicted yield versus the actual yield recorded by the farmers for one of the ANN models built for cases 4a and 6a, respectively. The models correctly follow the yield fluctuation most of the weeks.

4) Impact of parameters: use of week number and new fruit set numbers

In almost all cases, using the current calendar week number did not improve prediction accuracy significantly (cases 1b, 2b, 3b, and 4b compared to cases 1a, 2a, 3a, and 4a, respectively), and, in case 4b, reduced the best prediction error value to 0.42 from 0.28. Using new fruit set numbers by itself also did not improve prediction accuracy (cases 1a and 1b compared to cases 2a and 2b).

C. Comparison to Previous Work

As Lin et al. [7] notes, comparing the effectiveness of yield prediction models to other work is difficult due to different approaches in measuring yield and estimating errors. For example, the prediction error, as measured in this work, reached 28% for yield predictions seven weeks early when using the expected environmental readings. This is lower than the best reported error of 30% for predicting next week's yield in (Lin et al. [7]). However, while we select a best topology as represented by the average over four best models, one per cross validation, Lin et al. select the best model based on the ANN architecture, the correlation coefficient values, and the root mean square error. We also categorize yields prior to creating the ANNs while Lin et al. [7] and other researchers use exact yield values in their ANN models.

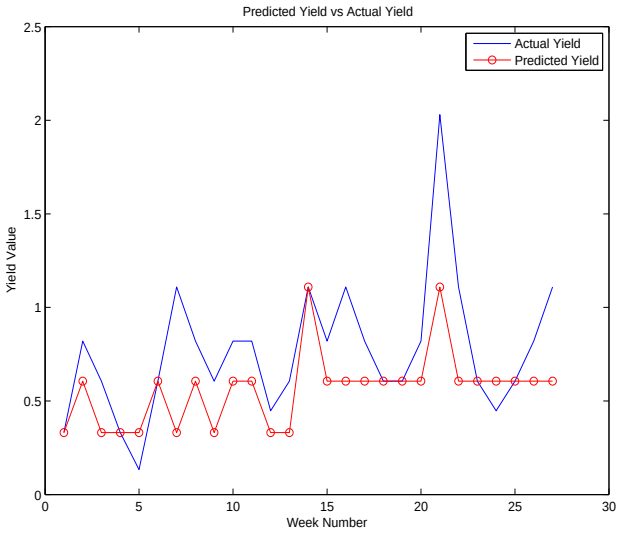

Figure 2: The predicted yield values versus the actual recorded yield for a case 4a model which predicts yield seven weeks in advance.

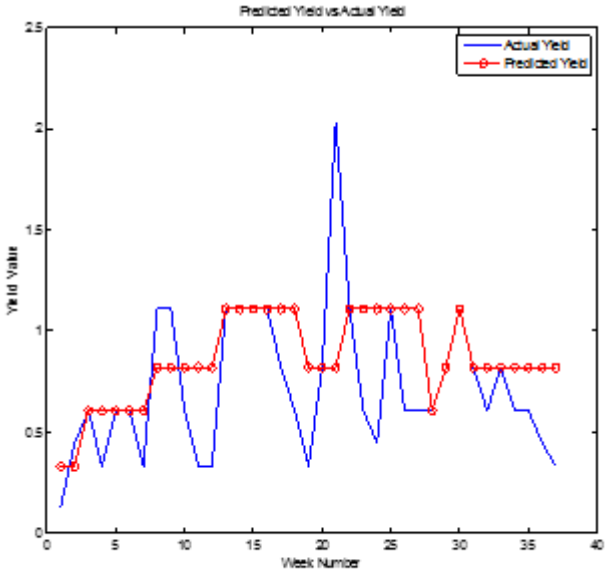

Figure 1: The predicted yield values versus the actual recorded yield for a case 6a model which predicts yield four weeks in advance.

Sauviller et al. [5] reported an error of 7-13% on external data. However, they measured their yield by the number of fruits per m^2 based on thirty-two sample plants. It is unclear how this approach would generalize to commercial greenhouses with thousands of plants and how prediction accuracy is affected when we are further away from the prediction week. Also, the number of fruits does not necessarily account for the weight of the total yield. Fruit weight and size has been shown to be affected by environmental and physiological factors other than temperature such as light and the plants' fruit load [11]. This introduces potential variation in the actual yield weight for the same fruit number. Furthermore, the degree to which a specific factor affects yield may also be influenced by other factors such as the effect of light in higher temperatures [11]. This adds new variables affecting the total yield but not accounted for by the temperature-based model.

IV. Conclusion

Yield prediction is a challenging problem due to the variety of possible variables affecting yield and the irregular yield pattern of greenhouse peppers. In this paper, the problem was approached as a classification problem rather than a regression problem and used expected future values to predict yield up to seven weeks in advance. Our results indicate that long-term (+7 weeks) prediction can be achieved with the same or better accuracy than short-term (+1 week) prediction. These results

were achieved using commercial data from four years. We plan in the future to field test this model on a large scale inviting multiple farmers to submit data to a specially designed website.

REFERENCES

[1] E. Heuvelink, *Tomatoes*. Wallingford: CABI Publishing, 2005.

[2] L. F. . Marcelis, E. Heuvelink, and J. Goudriaan, "Modelling biomass production and yield of horticultural crops: a review," *Sci. Hortic. (Amsterdam).*, vol. 74, no. 1–2, pp. 83–111, Apr. 1998.

[3] S. Adams, "Predicting the weekly fluctuations in glasshouse tomato yields," *Acta Hortic.*, pp. 19–23, 2002.

[4] P. Verroens, B. E. Verlinden, J. Lammertyn, B. De Ketelaere, and B. M. Nicolaï, "Time Series Analysis of Capsicum annuum Fruit Production Cultivated in Greenhouse," *Acta Hortic.*, vol. 718, pp. 97–104, 2006.

[5] C. Sauviller, W. Baets, B. Verlinden, and B. M. Nicolai, "Predicting the Weekly Yield Fluctuations of Greenhouse Bell Pepper," *Acta Hortic.*, vol. 817, pp. 261–268, 2009.

[6] K. Qaddoum and E. Hines, "Adaptive Neuro-Fuzzy Modeling For Crop Yield Predictio," in *Recent Researches in Artificial Intelligence, Knowledge Engineering and Data Bases*, 2011, pp. 199–204.

[7] W. C. Lin and B. D. Hill, "Neural network modelling to predict weekly yields of sweet peppers in a commercial greenhouse," *Can. J. Plant Sci.*, vol. 88, no. 3, pp. 531–536, May 2008.

[8] M. Kaul, R. L. Hill, and C. Walthall, "Artificial neural networks for corn and soybean yield prediction," *Agric. Syst.*, vol. 85, no. 1, pp. 1–18, Jul. 2005.

[9] D. L. Ehret, B. D. Hill, T. Helmer, and D. R. Edwards, "Neural network modeling of greenhouse tomato yield, growth and water use from automated crop monitoring data," *Comput. Electron. Agric.*, vol. 79, no. 1, pp. 82–89, Oct. 2011.

[10] E. Heuvelink, L. Marcelis, and O. Korner, "How to reduce yield fluctuations in sweet pepper?," *ACTA Hortic.*, pp. 349–355, 2004.

[11] B. Marcelis, L.F.M and Hoffman-Eijer, "growth analysis of sweet pepper.pdf," *Acta Hortic.*, vol. 412, pp. 470–478, 1995.

AUTHORS' PROFILE

Dr. Reem Al-Halimi recieved her B.Sc. in Computer Science from Rensselaer Polytechnic Institute, New York, U.S. in 1993, and her Ph.D. in Math from the University of Waterloo, On, Canada in 2004. She is currently the CTO at iKlyk Inc. in Waterloo, On, Canada. Her research interests include information retrieval, artificial intelligence, machine learning, and neural networks.

Medhat A. Moussa received the B.A.Sc. degree from the American University, Cairo, Egypt and the M.A.Sc. degree from the Universite de Moncton, Moncton, NB, Canada, both in mechanical engineering, and the Ph.D. degree in systems design engineering from the University of Waterloo, Waterloo, ON, Canada, in 1987, 1991, and 1996, respectively. He is currently a Professor with the School of Engineering, University of Guelph, Ontario, Canada. His research interests include user-adaptive robots, machine vision, machine learning, neural networks, and human–robot interaction. Dr. Moussa is a member of the IEEE and ACM associations.

©The Author(s) 2015. This article is published with open access by the GSTF