Near real-time comprehension classification with artificial neural networks: decoding e-Learner non-verbal behaviour

Mike Holmes, Annabel Latham, Keeley Crockett, and James D. O'Shea

Abstract—Comprehension is an important cognitive state for learning. Human tutors recognise comprehension and non-comprehension states by interpreting learner non-verbal behaviour (NVB). Experienced tutors adapt pedagogy, materials and instruction to provide additional learning scaffold in the context of perceived learner comprehension. Near real-time assessment for e-learner comprehension of on-screen information could provide a powerful tool for both adaptation within intelligent e-learning platforms and appraisal of tutorial content for learning analytics. However, literature suggests that no existing method for automatic classification of learner comprehension by analysis of NVB can provide a practical solution in an e-learning, on-screen, context. This paper presents design, development and evaluation of COMPASS, a novel near real-time comprehension classification system for use in detecting learner comprehension of on-screen information during e-learning activities. COMPASS uses a novel descriptive analysis of learner behaviour, image processing techniques and artificial neural networks to model and classify authentic comprehension indicative non-verbal behaviour. This paper presents a study in which 44 undergraduate students answered on-screen multiple choice questions relating to computer programming. Using a front-facing USB web camera the behaviour of the learner is recorded during reading and appraisal of on-screen information. The resultant dataset of non-verbal behaviour and question-answer scores has been used to train artificial neural network (ANN) to classify comprehension and non-comprehension states in near real-time. The trained comprehension classifier achieved normalised classification accuracy of 75.8%.

Index Terms—E-learning tools, Adaptive and intelligent educational systems, Intelligent tutoring systems, Nonverbal synthesis, Neural nets, Face and gesture recognition

1 Introduction

Non-verbal behaviour (NVB) is a broad descriptor of any communicative behaviour not involving verbalisation,. NVB includes facial expressions, gestures, posture, movement and physiology. Classroom studies [1], [2], [3] establish the important role of NVB in mediating the dynamics of tuition. Human tutors use intuitive understanding of learner NVB to adapt pedagogic methods and plan interventions, developing trust and improving support [4]. Review of real-world e-learning platforms [5], [6] suggests that the naivety of pedagogy in many systems negatively affects the quality, effectiveness and satisfaction of tuition. To bridge the gap in effective pedagogy between human and digital tuition, automatic analysis of learner NVB has become an active research area.

Literature on classifying learner cognitive states from NVB presents a broad range of technical approaches and assessment contexts. Related work (section 2) discusses comprehension assessment in dyadic verbal information recall tasks [7], [8], use of specialised high-speed eye-tracking cameras [9], [10] and heat-maps [11] to classify reading comprehension, classifying learning activity states using facial expressions [12] and predicting self-reported task difficulty by coarse head movements [13]. Despite the rich literature in the field, a generalised, cost-effective, practical and non-intrusive method of objectively classifying e-learner comprehension of on-screen information is absent.

This paper presents COMPASS, a novel near real-time comprehension assessment and scoring system for automatically classifying e-learner comprehension of on-screen information as it is being read and appraised. COMPASS has been developed as part of an on-going project to equip an intelligent e-learning platform with human-like understanding of comprehension indicative learner behaviour, so as to allow the system to enact timely and appropriate interventions in the learning process. COMPASS addresses the gap in the literature for a generalised, practical, non-intrusive, near real-time classifier designed to feedback on learners' comprehension levels during mental processing for a variety of on-screen information types including discursive text, numeric and algebraic equations, programming code and diagrams.

In this paper the authors discuss the design and development of COMPASS, and present an evaluation of the effectiveness of COMPASS as a system of classification for polar *strong comprehension* and *strong non-comprehension* indicative behaviours. In designing COMPASS the authors draw on literature (section 2) to produce a novel, practical, low-cost and non-intrusive method for comprehension classification during human to computer non-verbal interactions. This paper presents a novel extended model of learner behaviour incorporating educational meta-data, a novel robust image processing algorithm for tracking NVB in uncontrolled real-world e-learning scenes, and a novel neural networks based comprehension classifier for use during non-communicative consumption and mental processing of on-screen information.

The novelties included in this paper are:-

- Design and evaluation of a novel model of learner non-verbal behaviour (section 3.4.2) incorporating state and transient behavioural attributes, and general and education specific user meta-data;
- Design and evaluation of a novel, practical and non-intrusive non-verbal behaviour modelling algorithm (section 3.4.3) for extracting detailed near real-time behavioural information from web camera image data in real-world e-learning environments;
- Design and evaluation of a novel near real-time comprehension classifier (section 3.4.4) capable of learning discriminant behavioural patterns indicative of comprehension states, without reliance on self-reporting or proxy labels such as affective states.

The COMPASS image processing algorithm and comprehension classifier have been developed and evaluated using a large dataset of comprehension class-labelled web camera footage, generated by student volunteers at Manchester Metropolitan University using a bespoke on-screen quiz system. This paper discusses how the experimental dataset was collected and labelled, the development of a robust image processing algorithm which reliably produces a novel learner behavioural model, and the training and testing of a multilayer perceptron network to classify *strong comprehension* and *strong non-comprehension* indicative patterns of behaviour.

In this paper the authors present relevant literature on existing methods for computational analysis of learner non-verbal behaviour for cognitive and comprehension classification (2). The authors present a study (3) in which a large volume of visual- and meta- data have been collected (3.4.1), modelled (3.4.2) and used to train and test an artificial neural network to classify comprehension and non-comprehension states (3.4.4). The authors present results from training and testing of the comprehension classifier (3.5), conclusions (4) and future work (5).

2 Related research

In this section the author reviews literature to highlight and discuss the advantages and disadvantages of the various technical approaches taken to classifying learner performance by analysis of non-verbal behaviour.

Research on affective state dynamics [14] provides a framework for analysing how affects (such as boredom and frustration) can be used as proxies for cognitive events, such as impasse (non-comprehension). Affect responsive systems [12], [15], [16], [17] use facial expressions [12] or elements of the Facial Action Coding System [18] to model and adapt to the emotional valence of a learner. While facial expression recognition can be reliably performed using low-cost and non-intrusive camera hardware [19], the approach does have limitations. In trying to model learner comprehension *in near real-time* affect does not provide sufficient temporal accuracy. The literature [14] establishes that potential target affects such as boredom and frustration are lagged, occurring only after sustained or repeated instances of impasse. In this way, the marker (an affect expressive facial expression) for the event (non-comprehension) is presented some unknown time after the true event occurred an unknown number of times.

More temporally accurate methods use surveys of non-verbal behaviour to model patterns indicative of underlying states. Research using both skeletal movement [7] tracking and broad survey coarse grained multi-channel NVB analysis [8] seek to map complex behavioural patterns to objective ground-truth states. Research has shown that learner NVB is a strong indicator of learning and subject comprehension when analysed in human to human verbal information recall tasks, such as post-tuition interviews.

The use of standard video camera equipment in FATHOM [8] indicates a practical real-world technical solution for objective comprehension classification, but the literature lacks demonstration that the approach or findings are transferable from human to human verbal interaction to non-communicative on-screen information processing.

Analysis of NVB shows promising results when applied to predicting self-reported task difficulty [13], [20]. The research [13] shows that even coarse NVB such as head movement can be a strong indicator of the perception of task difficulty. While the results are promising, both studies rely on subjective post-hoc self-reporting of difficulty, rather than establishing an objective measure of comprehension as a ground-truth.

Analysing objective learner comprehension of on-screen information has been approached using eye-tracking techniques and body attached sensors [9], [10], [11], [21], [22]. The literature shows positive results in providing real-time comprehension classifications for mental processing of on-screen materials in a variety of laboratory experiments. Objective measures of comprehension or learner performance are established by appraising demonstrable performance under examination (e.g. when answering questions or identifying a region of programming code). However, the technical approach is limited in real-world applications due to the high costs and impracticality of using specialist hardware, such as high speed eye tracking cameras, head mounts, chin rests, special chairs or body attached sensors, in a real classroom environment with many students. An eye-tracking and heat-map approach [11] presents a second problem. The approach depends on modelling discriminant fixation heat-maps for each state, but also for each information display. In this way, the approach lacks generality. A concern in adopting such an approach would be that information could not easily be changed without the classifier needing to learn new discriminative behavioural patterns.

Review of related work has highlighted a specific gap in the literature. While affect detection by classification of facial expressions is a low-cost and widely available technology, the classifications cannot be considered *near real-time*. Temporally accurate learning and comprehension systems, such as FATHOM [8], have demonstrated success in classifying broad survey coarse grained multi-channel NVB during dyadic human to human verbal information recall tasks, but the methods have not been evaluated for analysis during on-screen information consumption and processing tasks. Arithmetic comprehension has been explored, using practical technology, but the study models subjective post-hoc reviews of task difficulty without an objective measure of true comprehension. Where objective measures of comprehension during information reading and processing have been the context, technical solutions such as eye-tracking,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TLT.2017.2754497, IEEE Transactions on Learning Technologies

heat-maps and body attached sensors prove cumbersome and impractical.

None of the approaches highlighted in literature demonstrate the requisite criteria for a practical, low-cost, non-intrusive, near real-time classification method capable of analysing NVB to estimate and classify learner comprehension during consumption and mental processing of on-screen information. In this research the authors develop and evaluate a novel practical, low-cost, non-intrusive classifier for estimation and classification of e-learner comprehension during reading and mental processing of on-screen information.

In designing a novel system to solve the problem, the literature has provided a suitable starting point. Although the system was designed and evaluated for use on pre-recorded video recordings of human to human verbal information recall tasks (e.g. a recorded interview), FATHOM [8], [23] provides a viable technical approach on which this research can build. The approach uses practical, low-cost and non-intrusive camera hardware to monitor multiple channels of coarse non-verbal behaviour, such as head movement, gaze direction, blink rate and skin tone change (blushing and blanching) during verbal communication of recalled information. FATHOM [8], [23] uses a trained neural network to estimate the strength of association between the input behaviour pattern, a 40 variable numeric vector, and the polar classes *comprehension* and *non-comprehension*. The neural network outputs a single real number value on the scale of -1.0 (polar non-comprehension) to +1.0 (polar comprehension). A threshold function, for example ±0.5, is applied to the output scale to allow for binary classification. Network outputs not meeting the ±0.5 threshold would be considered non-indicative of comprehension and can be disregarded. FATHOM [8] is reported to have achieved normalised classification accuracy of 76%.

FATHOM [8], [23], and its methodological predecessor [24], monitors the subtle, subconscious and uncontrolled behavioural changes which occur in response to stress during information recall. The systems leverage the effects of active problem-solving during live information recall to create distinct patterns of behaviour where information has caused low or high cognitive load.

In information recall tasks, such as those examined by Buckingham et al. [8], cognitive load is increased by attempts to recall poorly comprehended information. Originally developed for exploring the process of language comprehension, the Construction Integration (CI) framework [25] provides an explanation for the mechanistic relationship between comprehension and stress responses.

CI [25] describes comprehension as a two-step process. When an entity, a word, symbol or concept, is incorporated then a network of related entities is automatically *constructed* from memory. The network represents all the known attributes and relationships for the entity. However, not all relations in the network will be relevant, given the context. *Integration* of context triggers the removal of irrelevant or incorrect relations. This process is described by Kintsch [25] as an automatic process, with little cognitive cost. Successful CI results in a coherent model of an entity, a word, symbol or concept, based on comprehension of its' relations to other entities. However, failure to *construct* or *integrate* results in active problem-solving. The problem-solving step, where new relationships and associations are reasoned and learned, is cognitively expensive. The CI model suggests that failure to *comprehend* will result in an increase in cognitive load, which can be detected as stress response non-verbal behaviour.

As well as in recall tasks, CI can also be applied to active learning. In this research the ambition is model e-learner comprehension levels as they read and process information from the screen, in near real-time. The CI model supports the hypothesis that patterns of learner NVB may be discriminable depending on the outcome of the CI process for on-screen information. On-screen information which cannot be *integrated* will cause increased cognitive load, stress response and some degree of change in physical or physiological non-verbal behaviour.

3 Study: Real-time e-learner comprehension classification from non-verbal behaviour by non intrusive means

In this paper the authors present a study of machine classification of comprehension indicative learner non-verbal behaviour during on-screen e-learning activities. The study is broken into four phases:-

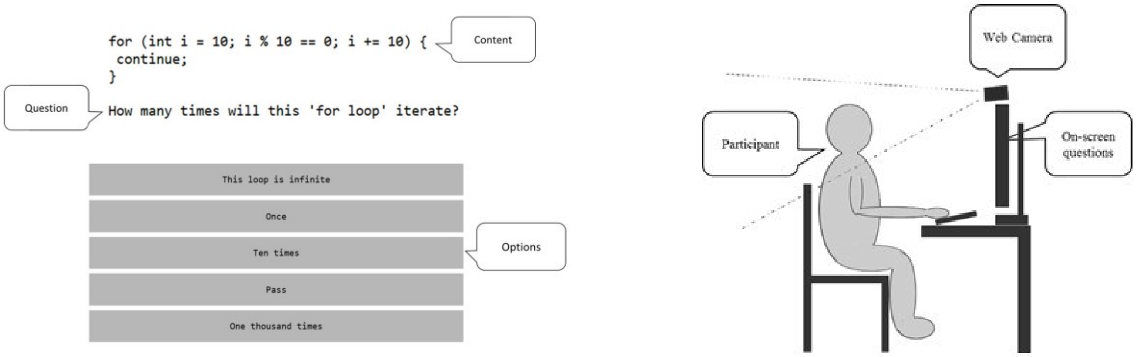
1) Data collection (3.4.1);
2) Behaviour model and extraction algorithm (3.4.2, 3.4.3);
3) Classifier selection, training and tuning (3.4.4);
4) Classifier testing and analysis of results (3.5).

In this section the authors present an outline of the study procedure and discuss each phase of the study, presenting method and results.

3.1 Study procedure outline

The study uses data collected from 44 undergraduate students at Manchester Metropolitan University (MMU) undertaking e-learning activities within a bespoke e-learning environment. Students were asked to complete a 21-question multiple choice quiz on Java programming, logic and information systems diagrams, while being recorded using a front-facing web camera attached to one of the pre-configured laptops provided for the experiment.

The bespoke quiz system developed for the experiment (Figure 1a) displayed a series of multiple choice questions, along with supporting information to be comprehended, in a random order. For each question the learner was presented with a piece of information to comprehend. This information included text, mathematics, diagrams and programming code. The learner then answered a multiple choice question in appraisal of the information. The learner was recorded using a front-facing web camera (Figure 1b) during each question period. Image streams for each question period were combined with the correctness of the answer given to produce a comprehension ground-truth labelled set of image streams. Image stream data for each question period was broken into 1-second sliding windows and analysed to extract a model of observed behaviour containing 42 numeric values. The model contains 37 behaviour observations and 5 learner meta-data constants. A multi-layer perceptron artificial neural network was trained to classify



```
for (int i = 10; i % 10 == 0; i += 10) {
 continue;
}
```

How many times will this 'for loop' iterate?

This loop is infinite

Once

Ten times

Pass

One thousand times

(a) Screen shot of quiz system

(b) Physical layout of experiment

Fig. 1. Data collection method

comprehension by back-propagation of errors [26] using 10-fold cross validation.

3.2 Study participants

44 student volunteers were randomly selected from the Science and Engineering Faculty at Manchester Metropolitan University (MMU). The participant group was diverse, a mix of ages, ethnicities and programming experience, to reflect the student body in computing related subjects. Participants’ ages ranged from 18 and 38, with a mean average age of 21 years’ old. 40 were enrolled on computing or computer science related courses, while 4 were enrolled on other science, mathematics or engineering courses. 39 participants identified that they had prior programming experience, while 5 identified no prior experience. The group is ethnically and culturally diverse, representing the inherent diversity of the student body. Ethnic demographic make-up of the group is shown in table 1. Ethnicity may be of particular importance in this experiment, as literature [23], [24], [27] suggests that ethnicity and culture play important roles in mediating subconscious non-verbal behaviour.

TABLE 1
Ethnic demographic groups

Group	Label	Count	Percentage
1	Asian or British Asian	15	34
2	Black or Black British	1	2
3	Mixed or multiple ethnic groups	1	2
4	Other ethnic group	0	0
5	White or White British	26	59
6	Undisclosed	1	2

3.3 Ethical considerations

Participants were required to sign a consent form prior to participation. The consent agreement detailed the information collected, its research use, and how the information would be securely stored and distributed. Personally identifiable information about participants, including images, will not be made public. Participation was paid by means of a retail voucher.

3.4 Methods

3.4.1 Data collection

COMPASS aims to classify comprehension of on-screen information during e-learning interactions. A novel set of recordings were made which show learners reading on-screen information (question text, diagrams, mathematics and programming code) before answering a related multiple-choice question (MCQ). The resultant dataset of behaviourally descriptive image data and objective ground-truth MCQ answers will be used to train and test the comprehension classifier (sections 3.4.3 and 3.4.4).

Participants completed a 21-question multiple-choice quiz on topics of programming, logic and information systems diagramming. Participants undertook the experiment individually and without interruption but were situated in a semi-public space within the university foyer. Learners undertook the quiz using a bespoke quiz application designed for the experiment (figure 1a). The bespoke quiz application captured the participants demographic data - age, gender, ethnicity and academic level, as well as whether they were wearing glasses, level of prior programming experience and if they were enrolled on a computing related course. The quiz application presented a baseline question, followed by 20 randomly ordered *on-topic* questions. The baseline question asked participants how old they would be in four years time and was included to ensure some comprehension behaviour for each participant. During each *on-topic* question the participant was presented with a piece of information to comprehend. Depending on the topic this may be a question text or statement, a numeric or algebraic equation or a diagram. The screen (figure 1a) also presented a set of answers to appraise the on-screen information. Answer were displayed in a random order to prevent the position of answers biasing observed behaviour. A ’pass’ option was always included to discourage guessing behaviour. A 3-second countdown was displayed between each question, to prevent behavioural overlap in different answer periods. During each question period the participant is recorded using a front-facing web camera (figure 1b). For each question period the image stream and answer score, +1 for correct or -1 for incorrect or pass, were stored.

The novel dataset provides a record of learner non-verbal behaviour during authentic comprehension or non-comprehension of on-screen information, followed by an immediate objective ground truth.

Questions within the quiz were based on the first year undergraduate computer science syllabus and were evaluated by programming course lecturers at MMU. Questions fell into three categories on Bloom's revised taxonomy [28] - *Remembering*, *Understanding* and *Analysing*. Questions were designed to provide more or less challenge, in line with Bloom's cognitive domain. Examples of quiz questions and taxonomic categories are shown in table 2.

TABLE 2
Example of questions in Bloom's revised taxonomy of the cognitive domain

Question	Category
In relation to Java programming, what does the acronym JVM stand for?	Remembering
$\forall x(P(x) \wedge Q(x)) \equiv \forall xP(x) \wedge \forall xQ(x)$ Is this statement true or false?	Understanding
Given the constructor $for(inti = 10; i\%10 == 0; i+ = 10)$ how many times will this loop iterate?	Analysing

The 44 participants completed 869 questions and generated 185,075 web camera video stream images for analysis. Not all participants answered the full 21 questions. In two cases volunteers' other commitments prevented completion and in one case the application crashed. Incomplete quiz data was included in the dataset as the random ordering of questions prevented per-question bias. Table 3 shows a breakdown of the data collected in both correct and incorrect (or pass) classes.

TABLE 3
Overview of learners' question answer data

	Learner correct	Learner incorrect	Total
Answers	500	369	869
Footage (seconds)	6,836	5,503	12,336
Images	102,535	82,540	185,075
Class %	55.42	44.58	100.00

3.4.2 Behaviour modelling

TABLE 4
Behavioural data model

Type	Channels	Examples
Learner	5	Gender, Ethnicity, Academic level, Specialism, Experience
Eyes	17	Openness, Gaze, Blink
Geometries	18	Position, Rotation, Movement
Physiological	2	Blush, Blanche

A novel data model (table 4) has been designed to capture a broad range of non-verbal behaviour channels. The model is populated by surveying the state of individual behavioural channels within each image of the web camera image stream and combining this with meta-data about the learner. A behavioural channel is a single observed behaviour, such as 'left eye gaze right' or 'head rotated left'. Each behavioural channel is a true or false question, which is represented as either +1.0 or -1.0.

Over each question period the image data is segmented into one second chunks. Each one second chunk contains 15 time-sequential still images, as captured from the web camera video stream. The behavioural data model is populated for each static image in turn, tracking not only state (e.g. left eye closed) but also change over time (e.g. blink). Each one second chunk produces a matrix of dimensions 42 by n-1, where n is the number of images in the time period. Finally, the matrix is summarised to produce a single 42 variable cumulative behavioural feature vector (CBFV), representative of the average behaviour expressed in the time period. If a full behavioural feature set cannot be extracted for the time period then processing stops and awaits the next time window.

3.4.3 Behaviour extraction

To reduce processing overheads, images are reduced to 360 x 240px and grey-scaled before being decomposed into regions of interest (ROI) containing features (for example, left and right eyes) using Haar cascades [29]. The scale invariant properties of Haar cascades allowed for effective feature location when the position of the learner relative to the camera is not strictly controlled. The ROI pixel data is then further reduced using principal component analysis, before being classified using a specific behavioural channel classifier. An example of the process for a single behavioural channel is shown in figure 2. Variables for state-change, geometries and pixel data are also extracted from each still image.

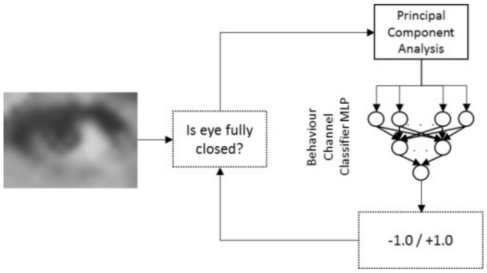

Principal Component Analysis
Is eye fully closed?
Behaviour Channel Classifier MLP
-1.0 / +1.0

Fig. 2. Channel classification method for 'eye fully closed'

TABLE 5
Breakdown of cumulative behavioural feature vectors extracted from question response periods

Comprehension	CBFV extracted	% of data
+1.0	3,551	53.86
-1.0	3,041	46.13
Total	6,592	100.00

The extraction process is repeated for each image in the given time window, producing a matrix of feature vectors comprehension of on-screen information, followed by an immediate objective ground truth.

behavioural feature vector

	Ch 1	Ch 2	..	Ch 41	Ch 42
Image 1	1	-1	..	1	1
...					
Image 14	1	-1	..	-1	1
CBFV	1	-1	..	0.5	1

which is the summarised to a Cumulative Behavioural Feature Vector (CBFV), as described in section 3.4.2. Table 5 shows the number of CBFV successfully created in each comprehension class.

The +1/-1 class distribution for CBFV (table 5) is close to that of the source (table 3), indicating consistent behaviour extraction across comprehension classes. The CBFV produced for each second will be the input vectors for comprehension classifier training and testing (sections 3.4.4, 3.5).

3.4.4 *Classifier training and evaluation*

A feed-forward multilayer perceptron artificial neural network was coded, tuned, trained, and tested. The network was configured with 42 input nodes, a single hidden layer containing 20 fully connected nodes (equation 1), and a single output node (equation 2). Network weights were initialised to 0±1/fan-in where fan-in is the number of inputs to the neuron. Binary comprehension classification is performed by application of a threshold function to the network output (equation 3, where *t* is a threshold value in the range 0.0 to 1.0).

$$h_i = \sigma \sum_{j=1}^{n=42} w_{ij} x_j + b_i \tag{1}$$

$$out = \sigma \sum_{i=1}^{n=20} w_i h_i + b \tag{2}$$

$$class = \begin{cases} 1.0 & tanh(out) \geq +t \\ -1.0, & tanh(out) \leq -t \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The dataset of comprehension labelled CBFV was split into two sets, training (Tr = 90%) and testing (Te = 10%). The classifier was initially trained using 10-fold cross validation, with each CBFV appearing once in the test set. For each fold the classifier was trained by back-propagation of errors for a maximum of 2000 epochs [26]. Training was halted early if training set mean square error (MSE) did not reduce over 200 epochs. The MSE was checked at 100 epoch intervals. If early stopping was triggered, the last best configuration of weights was saved to disk. The performance of the classifier has been evaluated by the average classification accuracy and precision over the 10 folds.

Literature [24], [27] suggests that ethnicity plays an important role in mediating subconscious non-verbal behaviour. To establish the effect ethnicity has on classifier behavioural channels within each image of the web camera image stream and combining this with meta-data about (table 1).

3.5 Results and discussion

TABLE 7
10-fold cross validation training

	All data			Ethnic group 5 only		
Fold	Epoch	Tr MSE	Te MSE	Epoch	Tr MSE	Te MSE
0	1100	0.74	0.95	800	0.77	1.00
1	600	0.75	0.93	1300	0.72	0.91
2	400	0.80	0.97	800	0.73	1.03
3	2000	0.65	1.07	1900	0.64	1.06
4	800	0.76	1.01	1600	0.70	1.11
5	1000	0.74	0.94	200	0.84	0.92
6	400	0.77	0.93	1200	0.70	1.03
7	800	0.74	0.99	700	0.74	0.96
8	400	0.82	0.98	1500	0.66	0.97
9	1500	0.67	1.04	1500	0.70	0.94

Table 7 shows the training (Tr), validation (va) and test set (Te) mean square error (MSE) for each of the 10 folds, along with the epoch at which the minima was found. The high average MSE suggests that there is a degree of noise within the model. The authors anticipated the model would be noisy, as the training data set would contain many weak examples of the label. The threshold function (equation 3) allows for noise, weakly indicative patterns, to be excluded.

Table 8 shows performance statistics (equations in 4) for true positive (TP), true negative (TN) and normalised classification accuracy (CA), as percentages, averaged over 10 folds with selection thresholds ranging ±0.6 − ±0.9. The results in table 8 show that accuracy increases as the threshold is raised.

$$\begin{aligned} TP &= \frac{tp}{tp + fn} \\ TN &= \frac{tn}{tn + fp} \\ CA &= \frac{tp + tn}{tp + fp + tn + fn} \end{aligned} \tag{4}$$

The results in table 8 show that when the classifier is trained using data from different ethnic groups the test set classification accuracy for non-comprehension (TN) behaviour is weaker than when trained on a single group. The results support the suggestion from literature [24], [27] that stress response NVB differs by demographic grouping. Literature also suggests that gender should be considered; however there were an insufficient number of female participants in this experiment to isolate the variable.

The trade-off in application therefore relates to accuracy versus frequency of classification. With a lower threshold, the network classifies a greater number of the CBFV but at the expense of accuracy, or visa versa. This effect is evident when comparing the count and accuracy of classifications shown in the confusion matrices in table 9.

Figures 3, 4 and 5 show COMPASS near real-time comprehension time-series data for the group 5 only data subset. While the time-series for figures 3 and 4 show consistent

TABLE 8
Classifier performance

	Training			Testing		
Threshold	TP (%)	TN (%)	CA (%)	TP (%)	TN (%)	CA (%)
All data						
±0.6	88.8	78.1	83.5	79.4	62.2	71.7
±0.8	91.3	81.5	86.4	81.3	64.7	73.9
±0.9	92.3	81.2	86.7	81.0	65.3	74.1
Only data from ethnic group 5						
±0.6	88.2	81.3	84.7	74.2	64.7	69.9
±0.8	90.5	84.8	87.6	77.2	68.0	73.1
±0.9	91.7	85.6	88.6	78.8	72.2	75.8

TABLE 9
Test set confusion matrices for group 5 classifier

Threshold ±0.8		Prediction		
		Positive	Negative	
Observation	Positive	602	178	77.2%
	Negative	199	422	68.0%
		75.2%	70.3%	
Threshold ±0.9		Prediction		
		Positive	Negative	
Observation	Positive	304	82	78.8%
	Negative	86	223	72.2%
		77.9%	73.1%	

comprehension indicative behaviour, matching the outcome of the interaction, figure 5 shows how COMPASS can track as a learner's comprehension indicative behaviour changes as they process information from the screen. Figure 5 highlights the value of the system. The strong change in behaviour identified in second 10 can be used as a trigger for appropriate intervention in the learning process.

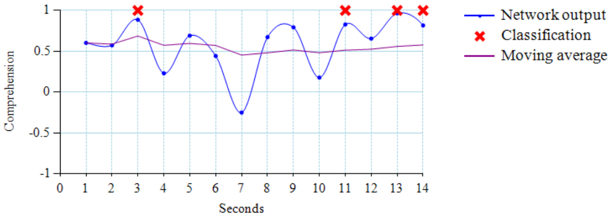


Fig. 3. COMPASS time-series for a correct answer period

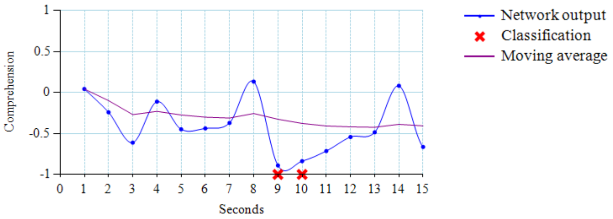


Fig. 4. COMPASS time-series for an incorrect answer period

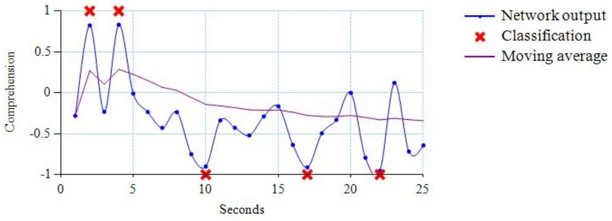


Fig. 5. COMPASS time-series for an incorrect answer period

4 Conclusions

This paper has presented the design, development and evaluation of COMPASS, a novel computer vision and machine learning based approach to near real-time comprehension classification for on-screen information. The contribution to literature is to demonstrate that non-verbal behaviour can be used as an effective indicator of comprehension and non-comprehension of on-screen information in an e-learning context, without the need for impractical and intrusive hardware or lagged proxies such as affect.

The paper has presented methods for extracting a behavioural data model from web camera image streams using a combination of Haar cascades, artificial neural networks, and geometries. The paper has presented methods for training and evaluating a multilayer perceptron network (MLP) to classify behavioural patterns as indicative of comprehension or non-comprehension.

The results presented in this paper show that the methods are effective in extracting a data model of non-verbal behaviour from web camera image streams, and that a MLP is an effective tool for classification of the behavioural model. The results show that the application of a logistic function to the MLP output allows for tuning of classification accuracy, by exclusion of patterns which are weakly indicative. The results identify a threshold of ±0.9 as optimal for classification accuracy where the classifier is trained on individual demographic groupings, achieving a test set normalised classification accuracy of 75.8% and average precision of 75.5%. Analysis of the COMPASS time-series' for question answer periods demonstrates how the classification system could be used as a near real-time feedback channel for an adaptive e-learning platform, enabling timely and appropriate interventions in the learning process.

However, there are a number of limitations to this study. The results confirms that comprehension indicative NVB differs between demographic groups. The demographic make-up of the participant group made it difficult to fully assess the effect that demographic variables have on indicative patterns of behaviour. There were too few female participants to evaluate the effect of gender on behaviour. The study also does not address how question type 2 might promote differing behaviour. In future work these two questions should be addressed.

The results suggest that COMPASS can identify timely intervention points at which an intelligent e-learning platform could scaffold the learning experience for learners who fail to comprehend tutorial information. Additionally, COMPASS could be used to analyse tutorial content as part of a learning analytics application.

5 Future work

The authors will evaluate the accuracy of the trained classifier in more complex e-learning platforms, by integrating COMPASS with a conversational intelligent tutoring system called Hendrix [30]. The effect of question type (table 2) on behaviour should also be investigated. A study to evaluate the effectiveness of comprehension based interventions will be conducted. Future work should explore whether NVB modelling and behaviour labelling can be improved using deep learning techniques. The approach may also be useful in analysing other behaviours and cognitive states relating to e-learning.

References

[1] S. Machida, "Teacher accuracy in decoding nonverbal indicants of comprehension and noncomprehension in Anglo-and Mexican-American children." *Journal of Educational Psychology*, vol. 78, no. 6, p. 454, 1986.

[2] M. W. Alibali, L. M. Flevares, and S. Goldin-Meadow, "Assessing knowledge conveyed in gesture: Do teachers have the upper hand?" *Journal of Educational Psychology*, vol. 89, no. 1, p. 183, 1997.

[3] J. M. Webb, E. M. Diana, P. Luft, E. W. Brooks, and E. L. Brennan, "Influence of pedagogical expertise and feedback on assessing student comprehension from nonverbal behavior," *The Journal of Educational Research*, vol. 91, no. 2, pp. 89–97, 1997.

[4] M. DePaulo and S. Friedman, "Nonverbal communication," in *The handbook of social psychology*, S. Fiske, D. Gilbert, and G. Lindzey, Eds. Boston : McGraw-Hill New York : Distributed exclusively by Oxford University Press, 1998, vol. 2.

[5] G. Salmon, "Flying not flapping: a strategic framework for elearning and pedagogical innovation in higher education institutions," *ALT-J*, vol. 13, no. 3, pp. 201–218, Oct. 2005.

[6] L. A. Daz and F. B. Entonado, "Are the functions of teachers in e-learning and face-to-face learning environments really different?" *Educational Technology & Society*, vol. 12, no. 4, pp. 331–343, 2009.

[7] A. S. Won, J. N. Bailenson, and J. H. Janssen, "Automatic Detection of Nonverbal Behavior Predicts Learning in Dyadic Interactions," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 112–125, Apr. 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6827904/

[8] F. J. Buckingham, K. A. Crockett, Z. A. Bandar, and J. D. O'Shea, "FATHOM: A neural network-based non-verbal human comprehension detection system for learning environments," in *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on*. IEEE, 2014, pp. 403–409.

[9] R. Bednarik and M. Tukiainen, "An eye-tracking methodology for characterizing program comprehension processes," in *Proceedings of the 2006 symposium on Eye tracking research & applications*. ACM, 2006, pp. 125–132.

[10] L. Copeland, T. Gedeon, and S. Mendis, "Predicting reading comprehension scores from eye movements using artificial neural networks and fuzzy output error," *Artificial Intelligence Research*, vol. 3, no. 3, 2014.

[11] S.-C. Chen, H.-C. She, M.-H. Chuang, J.-Y. Wu, J.-L. Tsai, and T.-P. Jung, "Eye movements predict students' computer-based assessment performance of physics concepts in different presentation modalities," *Computers & Education*, vol. 74, pp. 61–72, 2014.

[12] J. Whitehill, Z. Serpell, A. Foster, Y.-C. Lin, B. Pearson, M. Bartlett, and J. Movellan, "Towards an optimal affect-sensitive instructional system of cognitive skills," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*. IEEE, 2011, pp. 20–25.

[13] M. van Amelsvoort, B. Joosten, E. Krahmer, and E. Postma, "Using non-verbal cues to (automatically) assess childrens performance difficulties with arithmetic problems," *Computers in Human Behavior*, vol. 29, no. 3, pp. 654–664, May 2013. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0747563212002968

[14] S. DMello and A. Graesser, "Dynamics of affective states during complex learning," *Learning and Instruction*, vol. 22, no. 2, pp. 145–157, 2012.

[15] R. Rajendran, S. Iyer, S. Murthy, C. Wilson, and J. Sheard, "A Theory-Driven Approach to Predict Frustration in an ITS," *IEEE Transactions on Learning Technologies*, vol. 6, no. 4, pp. 378–388, 2013.

[16] H.-C. K. Lin, C.-H. Wu, and Y.-P. Hsueh, "The influence of using affective tutoring system in accounting remedial instruction on learning performance and usability," *Computers in Human Behavior*, 2014.

[17] R. A. Calvo and S. D'Mello, "Frontiers of affect-aware learning technologies," *Intelligent Systems, IEEE*, vol. 27, no. 6, pp. 86–89, 2012.

[18] P. Ekman and W. V. Friesen, *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*. Cambridge, MA: Malor Books, 2003.

[19] M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Measuring facial expressions by computer image analysis," *Psychophysiology*, vol. 36, no. 02, pp. 253–263, 1999.

[20] D. Hrubes and R. S. Feldman, "Nonverbal Displays as Indicants of Task Difficulty," *Contemporary Educational Psychology*, vol. 26, no. 2, pp. 267–276, 2001.

[21] S. K. DMello and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Modeling and User-Adapted Interaction*, vol. 20, no. 2, pp. 147–187, 2010.

[22] S. Yusuf, H. Kagdi, and J. I. Maletic, "Assessing the comprehension of UML class diagrams via eye tracking," in *Program Comprehension, 2007. ICPC'07. 15th IEEE International Conference on*. IEEE, 2007, pp. 113–122.

[23] F. J. Buckingham, K. A. Crockett, Z. A. Bandar, J. D. O'Shea, K. M. MacQueen, and M. Chen, "Measuring human comprehension from nonverbal behaviour using Artificial Neural Networks," in *Neural Networks (IJCNN), The 2012 International Joint Conference on*. IEEE, 2012, pp. 1–8.

[24] J. Rothwell, Z. Bandar, J. O'Shea, and D. McLean, "Silent talker: a new computer-based system for the analysis of facial cues to deception," *Applied Cognitive Psychology*, vol. 20, no. 6, pp. 757–777, 2006.

[25] W. Kintsch, *Comprehension: A paradigm for cognition*. Cambridge, MA: Cambridge University Press, 1998.

[26] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Macmillan, 1994.

[27] C. F. Bond, A. Omar, A. Mahmoud, and R. N. Bonser, "Lie detection across cultures," *Journal of Nonverbal Behavior*, vol. 14, no. 3, pp. 189–204, 1990. [Online]. Available: http://dx.doi.org/10.1007/BF00996226

[28] L. W. Anderson, D. R. Krathwohl, and B. S. Bloom, *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives*. Longman, 2001.

[29] M. Castrilló n-Santana, O. Déniz Surez, L. Antó n Canalís, L. Navarro, and J. Javier, "Face and facial feauture detection evaluation: Performance evaluation of public domain haar detectors for face and facial feature detection," 2008.

[30] M. Holmes, A. Latham, K. Crockett, J. D. O'Shea, and C. Lewin, "Hendrix: A conversational intelligent tutoring system for Java programming." Presented at the 15th annual UK Workshops on Computational Intelligence (UKCI), Exeter, UK, 2015.