# Recognition of human activity and the state of an assembly task using vision and inertial sensor fusion methods

James Male and Uriel Martinez-Hernandez

***Abstract*— The development of reliable human machine interfaces is key to accomplishing the goals of Industry 4.0. This work proposes the late fusion of a visual recognition and human action recognition classifier. Vision is used to recognise the number of screws assembled into a mock part while action recognition from body worn Inertial Measurement Units (IMUs) classifies actions done to assemble the part. Convolutional Neural Network (CNN) methods are used in both modes of classification before various late fusion methods are analysed for prediction of a final state estimate. The fusion methods investigated are mean, weighted average, Support Vector Machine (SVM), Bayesian, Artificial Neural Network (ANN) and Long Short Term Memory (LSTM). The results show the LSTM fusion method to perform best, with accuracy of 93% compared to 81% for IMU and 77% for visual sensing. Development of sensor fusion methods such as these is key to reliable Human Machine Interaction (HMI).**

## I. INTRODUCTION

The development of Industry 4.0 is widely regarded as the next major milestone for the manufacturing industry [1]. This development will allow highly customizable products and services with increases in efficiency and faster delivery time [2][3][4]. There are a wide variety of technologies underpinning the development of Industry 4.0: cyber-physical systems, Human Robot Interaction, sensor fusion, artificial intelligence, Internet of Things, amongst others [1].

A key aspect to the fulfilment of achieving Industry 4.0 is the development of natural human robot collaboration (HRC) [1][2][4]. In HRC, humans must be able to work closely with robots in an efficient, safe and predictable manner. These processes require the robot to be capable of perceiving accurately the current task being achieved and its relation with the wider goal. This knowledge can then lead to a predictive nature where the robot proactively decides what actions are required next for more natural interaction. This approach can be applied to Human Machine Interaction (HMI) where a machine may be made more intelligent by understanding the users need.

For machines and humans to work closely in an efficient and safe manner, the machines should be capable of understanding and making decisions during interaction tasks. These processes have led to the research and development of computation methods for the recognition of human actions, gesture control and learning from interaction with the environment [2][5][6][7]. These computation methods need to make use of information from various sensing modalities, together with fusion strategies that allow machines or robots to make accurate decisions [8][9][10]. Examples of sensor fusion studied before include multimodal vision [11], visual-inertial [12][13], visual-audio [14].

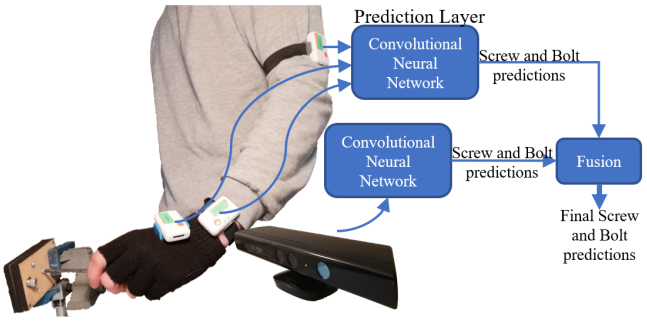


Fig. 1. Overview of the multimodal sensor fusion approach for recognition of the number of screws and bolts assembled into a wooden part.

Convolutional Neural Networks (CNNs) have been shown to have large success and versatility across sensing modalities with limited requirement for prior feature extraction and hence are used in this study. While fusion methods should build on being adaptable to a wide range of sensor modalities, this work focuses on the use of IMU and visual sensing. These sensing modalities complement each other well and are useful in an industrial setting having already shown success in various other applications [12][15][16].

The work presented here is focused around a mock industrial assembly task. A wooden part requiring three screws and a bolt to be inserted is used to simulate a part with assembly sequence requiring various tasks (Fig. 1). Visual recognition is used to determine the current number of screws and bolts inserted while action recognition from bodyworn IMU sensors tracks the tasks completed so far in the assembly process. In both vision and IMU recognition, CNN classifiers are used for recognition of the desired features while fusion of the two output classes by averaging, SVM, Bayesian, ANN and LSTM methods provides a more reliable estimate of the current part state. This work builds on that done in [17] where a weighted averaging fusion method is shown to have improved results in a 3D printer assembly task when compared to individual vision, IMU and EMG methods.

Training of the visual and IMU classifiers is initially done separately offline to optimise and evaluate performance. Evaluation of the final fusion system is done using a data set recorded in real time by completing the full assembly process and evaluating the accuracy of the individual vision and IMU methods along with the proposed fusion method.

This work was supported by The Engineering and Physical Sciences Research Council (EPSRC) and the Royal Society Research Grants for the 'Touching and feeling the immersive world' project (RGS/R2/192346)

James and Uriel are with the inte-R-action lab, the Centre for Autonomous Robotics (CENTAUR) and the Department of Electronic & Electrical Engineering, University of Bath, UK (jjm53, u.martinez)@bath.ac.uk

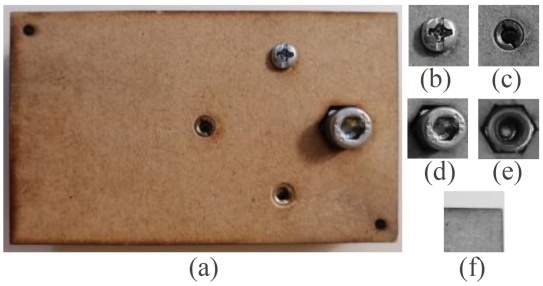


Fig. 2. (a) Part to be assembled (b-f) example images for classification: (b) Screw Full (c) Screw Hole (d) Bolt Full (e) Bolt Hole (f) Null

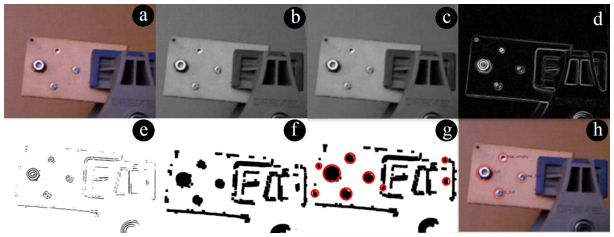


Fig. 3. Steps of image processing: (a) Original (b) Grayscale (c) Gaussian and median filtering (d) Sobel filter (e) Thresholding (f) Erosion (g) Blob detection (h) Classification and four closest points found

## II. METHODS

A semi-realistic assembly task is devised where the benefits of multimodal sensor fusion can be investigated. A 6×10 cm wooden part (Fig. 2a) is made with 3 holes requiring screws and one larger hole requiring a bolt; the screws are inserted with a screwdriver and the bolt with an Allen key. The part is used to collect visual and IMU training and testing data, as well as a data set of real time complete actions sequences for evaluation of various fusion methods. Various assembly task actions common to an industrial setting have been investigated previously [18][19][20], the screwdriver and Allen key movement actions are relatively similar so provide some challenge for a classifier to identify.

### A. Vision Sensing

Visual recognition is used to assess the current completion level of the part being assembled. A Kinect V1 RGB sensor is used to count the number of screws and bolts currently in the part. The recognition of screws being present or absent through vision has been demonstrated before [6][21] and while not challenging in ideal conditions, variation in lighting and occlusions present a cause for errors.

The RGB image is converted to greyscale before image processing techniques isolate the locations of estimated screws or screw holes (Fig. 3). The image has Gaussian and median filtering applied before a Sobel filter with thresholding to find edges. Eroding the image fills in the circular regions where a screw/bolt may be found then a circular blob detector is applied to find possible candidates.

For each of the possible screw/bolt locations a 28×28 pixel greyscale square is extracted around the centre point to feed into a classifier, see examples Fig. 2b-f. The classifier has a CNN structure (Fig. 4a) as this has been shown to have success in many image recognition tasks. The 28×28 input goes first through two convolution layers each with kernels of size 3×3 and 100 filters before a max-pooling layer with pool size of 2×2. Two more convolution layers and a max-pooling layer with the same structures are then applied before the data is flattened and fed into a fully connected layer with 128 units. A dropout layer with probability=0.5 follows before the softmax activation layer outputs to five categories: screw hole, screw, bolt hole, bolt, null.

The image classifier is trained offline using a data set of 20790 images. The data set is split into 16016 training, 4004 validation and 770 test images with an equal split between all 5 classes. For training the Adam optimiser is used with learning rate of 0.001, 12 epochs and batch size of 8.

Points classified as null are removed under the assumption they represent errors in the previous localisation techniques. Given the roughly square screw hole arrangement, the four points (or less) with most similar distance to each other are found and used as the candidates for state estimation.

To find the current state estimation an averaging approach is taken. The current 4 point classes and classification confidence values are updated at a rate of 10 Hz. The previous 1 s of estimations are used in two bin counts, one for the number of screws and one for the number of bolts. The bin counts are weighted by the confidence prediction from the classifier to give a final confidence for the number of screws and bolts with the maximum from each taken as the current state estimation. As there cannot be more than 3 screws or 1 bolt, the outputs for each class are limited to these values.

### B. Body Worn Sensors

Bodyworn IMUs are used to identify the current action being performed by the worker. The IMU devices used are Shimmer3's with three axis gyroscope and accelerometer signals streamed over Bluetooth to a laptop at 51.2 Hz. The three sensors, each with an accelerometer and gyroscope measuring over three axes, give a total of 18 channels of data. Given the dominant hand focused nature of the screwing tasks, three sensors are used on the persons dominant side placed on the top of the hand, wrist and upper arm (Fig. 1).

Classification is done using a sliding window and a 1D CNN classifier. The window length is 3 s, equating to 154 samples, with a step giving a new activity estimate every 0.5 s. The window length is chosen to provide a long enough period to have enough data to get an accurate classification while providing minimal overlap on adjacent classes. Each of the 18 data channels is initially preprocessed with scaling to give unit standard deviation and mean of zero.

The IMU classifier, Fig. 4b, takes the 154×1 window with 18 channels as its input. First there are two 1D convolutional layers, each with 100 filters and 5×1 kernel, followed by a 1D max-pooling layer with pool of 5. Two more convolution layers and another max-pooling of the same structures as before follow, then the data is flattened. A fully connected layer with 64 units and dropout with probability=0.5 follows before the final softmax output layer to 5 categories: screwing in, screwing out, Allen key in, Allen key out, null.

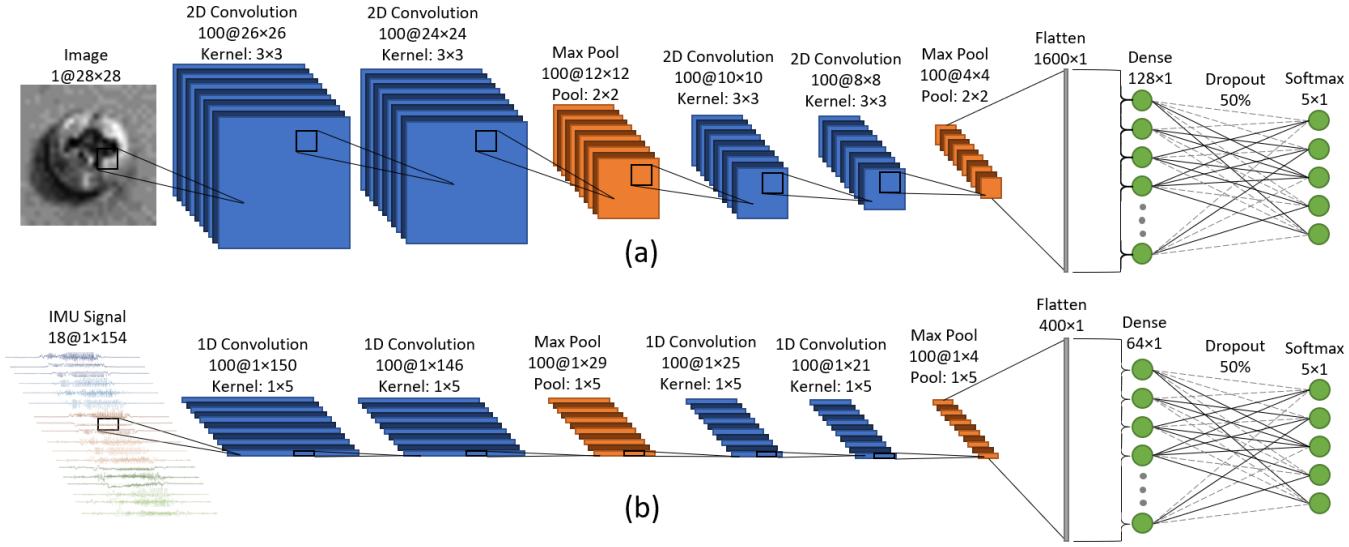


Fig. 4. (a) Image classifier CNN structure (b) IMU classifier CNN structure

Training the IMU classifier is performed offline in two phases. In the first phase of training 9 subjects perform 3 screwing in/out actions and 1 Allen key in/out action with data surrounding each action taken as null. The time taken to perform 3 screwing actions roughly equals the time to perform one Allen key action hence the different number of times each is performed. Overall 1453 s of data is used for training with an even split of 291 s for each category to avoid bias. The data for each category is split into 3 s windows with a 0.5 s step giving 567 windows for each category, a total of 2835 windows which is split into 2268 for training and 567 for validation. This training is run for 10 epochs with batch size of 8, learning rate of 0.001 and Adam optimiser.

A second stage of training is then performed with more realistic data sequences to improve the real time performance. Two subjects perform the entire assembly sequence of screwing the 3 screws and 1 bolt in during one trial (sequence shown in Fig. 5) with all data recorded and manually labelled. The first subject is to be used as training data and performs two runs with a reorientation of the part between each run providing slight variation to the actions, the second subject performs the run once over. The same windowing and preprocessing technique is applied with the truth label taken to be the most common label within the window. The training set has a total of 340 windows, 68 for each category split with 272 for training and 68 for validation; the test set has a total of 240 with 48 for each category. Training is done for 20 epochs with a batch size of 8, Adam optimiser and learning rate of 0.001.

The current action estimation is updated every 0.5 s and initially a basic filter is applied to remove outlying estimations. Denoting the action predicted at time $t$ as $A^t$, filtering is done as follows:

$$A^{t-1} = \begin{cases} A^t & A^t = A^{t-2} \\ A^{t-1} & otherwise \end{cases} \quad (1)$$

Adjacent actions of the same type are then grouped together with the confidence of each group found as the mean of the confidence values for each time step in the group. The final prediction for the total number of each actions completed is found from a weighted bin count of the list of grouped actions with the confidence of each group as the weight. As with the image prediction, the total number of screw predictions is limited to 3 and bolt predictions to 1.

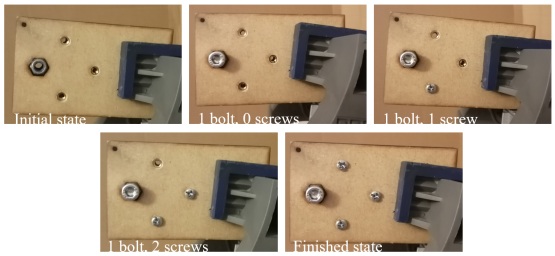


Fig. 5. Sequence of assembly states

## C. Fusion

By fusing vision and IMU predictions a more reliable estimate of the current part state should be achieved. As discussed by other sources there are three types of fusion: data, feature and decision level [10][22]. Decision level fusion methods are investigated by taking the estimate from each mode and fusing to find a final state estimation. The methods used are mean, weighted average, SVM, Bayesian, ANN and LSTM. The general fusion process is shown in the flowchart in Fig. 6.

To evaluate the different fusion methods, a data set is recorded of vision and IMU classifier outputs during completion of the assembly task. Two participants record the data set, each performing ten assembly sequences where a single assembly comprises screwing in three screws and one bolt, Fig. 5. For each participant five runs are done with the screws first and 5 with the bolt first in case of variation between the options. The test part is held in a vice on a table and the camera aligned to observe the screw/bolt holes, Fig. 1. The operator stands for the tasks and though their hand obstructs the camera view while performing an action, a clear view is

TABLE I

FUSION OUTPUT CLASSES

| Output Class | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| No. Screws | 0 | 1 | 2 | 3 | 0 | 1 | 2 | 3 |
| No. Bolts | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

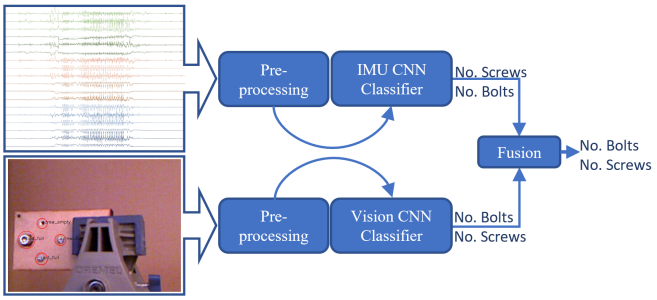


Fig. 6. Overall fusion system structure

available in between each action as the operator reaches for a new screw/bolt. After each action the vision/IMU classifier outputs are recorded, with the first frame where the operators hands are not occluding the part taken as the measurement point. This gives five analysis points in each run (including the starting state) for a total of twenty runs giving 100 analysis points.

Various fusion methods are implemented and analysed. Where applicable, 5 fold cross validation is used to train/fit a model to 16 of the 20 trials before testing on the remaining 4. Each method has an input size of 4 corresponding to the IMU and vision prediction of how many bolts and screws are present with the output to one of 8 possible classes (Table I). The mean method takes the mean number of bolts/screws from each classifier. The weighted average uses the accuracy, $Ac$, of each sensor mode for each fastener type, $f$, as the weight for each state prediction, $S$, i.e.:

$$S_f^{wa} = \frac{S_f^{vision} \cdot Ac_f^{vision} + S_f^{imu} \cdot Ac_f^{imu}}{Ac_f^{vision} + Ac_f^{imu}} \tag{2}$$

The Bayesian method implements a categorical Naive Bayes model while the SVM method used a one-vs-one multicategorical method. For the ANN fusion, a classifier with 2 hidden layers, each with 128 units followed by 0.5 dropout and trained for 40 epochs is used.

The LSTM method is trained using each trial as a data window, i.e. with a training input of 5 timesteps. The model has a single layer with 16 units and the full sequence returned followed by 0.3 dropout and the final softmax output. This is trained for 60 epochs.

## III. EXPERIMENTS

### A. Vision Recognition

Offline training and testing of the vision classifier is done as detailed in Section II-A. Using the 16016 training, 4004 validation and 770 testing images gives a final accuracy of 0.987 and loss of 0.059, see Fig. 7. The confusion matrix in Fig. 8a shows the even split in errors across all classes.

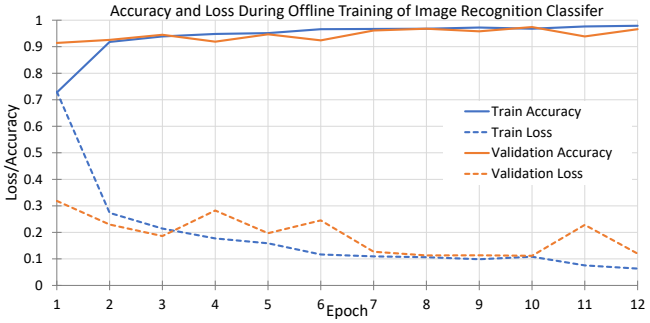


Fig. 7. Training plot of vision recognition CNN classifier

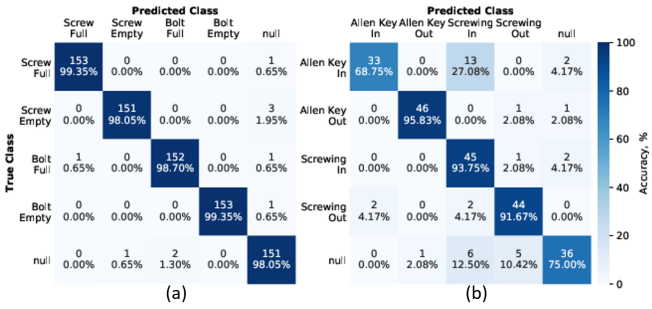


Fig. 8. Offline testing confusion matrices (a) vision (b) IMU classifiers

### B. Body Worn Sensor Classification

As detailed in Section II-B, the offline training of the IMU data classifier is split into two phases. For the first stage the 2835 windows recorded from 9 participants are split into 2268 training and 567 validation windows. For this stage of training the windows all contain readings from the same action type and each action is collected in a relatively structured way. This allowed faster collection of more data for the pretraining stage at the expense of not having completely natural actions. Training on this data set gives a validation accuracy of 0.977 and loss of 0.087.

The second stage of training is performed on realistic data. Two participants record complete assembly processes with the full sequences used for training and testing, the sequential part states are shown in Fig. 5. The first participant records 340 windows of data, split into 272 for training and 68 for validation; the second participant records 240 windows used for testing. This process yields an accuracy of 0.850 and loss of 0.771, Fig. 9. The confusion matrix from the two stage training approach, Fig. 8b, shows the main error as misclassifying Allen key in as screwing in, understandable given the similar nature of the two tasks.

The two stage training approach helps increase accuracy by first finding the distinguishing features for each class before training on more real data to help distinguish the appropriate features from background noise. The results of the different training possibilities are shown in Table II where the same test sequence was used on the classifier after only training on the first constrained data set, only training on the second realistic data set, training on both data sets shuffled together and the proposed two stage approach.

### C. Fusion Methods

Each fusion method is assessed on the same data set of 20 trials recorded by 2 participants. The data set contains the

TABLE II

IMU CLASSIFIER RESULTS AFTER TRAINING ON DIFFERENT DATA SETS

| | 1st data set | 2nd data set | 1st & 2nd shuffled | 1st then 2nd |
|---|---|---|---|---|
| Accuracy | 0.671 | 0.608 | 0.712 | 0.850 |
| Loss | 1.176 | 1.793 | 2.144 | 0.771 |

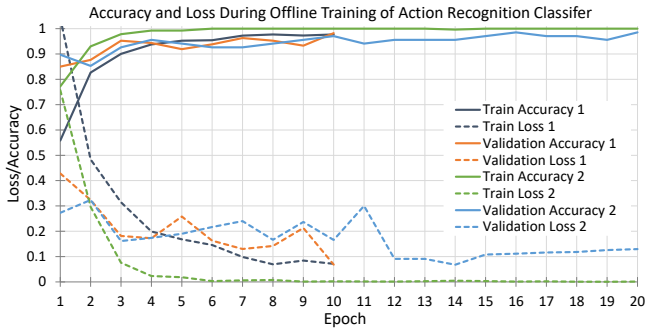


Fig. 9. Training plot of IMU recognition CNN classifier

screw and bolt output from the vision and IMU classifiers at the 5 measurement points during the assembly sequence. A final state prediction is made for each measurement point in each trial by the fusion method. If the fusion method requires a model to be fit/trained then 5 fold cross validation is performed by iterating through training on 16 and predicting on 4 trials.

Analysing the change in state from one measurement point to the next avoids counting single errors multiple times, for instance if the first action is misclassified then all future predictions will be wrong however the change in state could still be correct. The results from all trials combined gives the precision, recall and F1 scores shown in Table III.

Each state can be represented as underestimating, overestimating or correctly estimating the number of screws and bolts. At each state the prediction of over, under or correct for both screws and bolts is done for each method and the results shown in Table III.

The accuracy score gives an indication as to how far each method is from the correct prediction at each point as opposed to a binary correct/incorrect score. For each trial, the maximum number of errors, $E_{max}$, for screws is 13 and bolt is 5 as the maximum prediction is limited to 3 screws and 1 bolt. Taking the predicted part state for participant $p$ during trial $t$ at measurement point $n$ as $S^{p,t,n}_{pred}$, the average accuracy for each of screws or bolts can be found as follows:

$$Accuracy = 1 - \sum_{p=1}^{P}\sum_{t=1}^{T}\frac{\sum_{n=1}^{N}\left|S^{p,t,n}_{true} - S^{p,t,n}_{pred}\right|}{E_{max}} / P \times T \qquad (3)$$

Using the same error data, the root mean squared (RMS) error, $E_{rms}$, for each fusion method is found as follows:

$$E_{rms} = \sqrt{\frac{\sum_{p=1}^{P}\sum_{t=1}^{T}\sum_{n=1}^{N}(S^{p,t,n}_{true} - S^{p,t,n}_{pred})^2}{P \times T \times N}} \qquad (4)$$

## IV. DISCUSSION AND FUTURE WORK

The F1 analysis indicates the IMU method tends to over predict results, shown by the high recall and also seen in the confusion matrix. This is likely due to the relatively simplistic method of grouping and filtering IMU predictions where a single action could easily be classed as multiple actions if misclassification events occur, splitting the prediction into multiple events. This was observed during the experiment as operators fumbled the screws or paused during an action.

TABLE III

RESULTS ON FINAL DATA SET, BEST OF EACH RESULT IN **BOLD**

| | | IMU | Vision | Mean | W. Ave. | SVM | Bayes | ANN | LSTM |
|---|---|---|---|---|---|---|---|---|---|
| Precision | | 0.73 | 0.66 | 0.61 | 0.75 | 0.76 | 0.76 | 0.73 | **0.82** |
| Recall | | **0.95** | 0.68 | 0.80 | 0.72 | 0.68 | 0.69 | 0.71 | 0.86 |
| F1 Score | | 0.83 | 0.67 | 0.69 | 0.73 | 0.72 | 0.73 | 0.72 | **0.84** |
| State Estimation | Over | 0.24 | 0.27 | 0.39 | 0.11 | 0.10 | 0.09 | 0.10 | **0.07** |
| | Under | **0.03** | 0.14 | 0.04 | 0.14 | 0.14 | 0.14 | 0.14 | 0.05 |
| | Correct | 0.73 | 0.60 | 0.58 | 0.75 | 0.76 | 0.78 | 0.77 | **0.88** |
| Accuracy | Screw | 0.78 | 0.85 | 0.83 | 0.84 | 0.85 | 0.85 | 0.86 | **0.96** |
| | Bolt | 0.87 | 0.56 | 0.55 | 0.87 | 0.87 | **0.90** | 0.86 | 0.87 |
| | Overall | 0.81 | 0.77 | 0.75 | 0.85 | 0.86 | 0.87 | 0.86 | **0.93** |
| Overall RMS Error | | 0.73 | 0.66 | 0.70 | 0.57 | 0.52 | 0.53 | 0.54 | **0.35** |

The confusion matrices and accuracy scores show that the IMU method is generally better at predicting the number of bolts while the vision method is better at screws. While taking the basic mean of the two methods tended to result in worse accuracy, the weighted average method successfully gives improved correct guesses and better accuracy.

Comparing the weighted average, SVM, Bayesian and ANN methods it can be seen the majority of metrics show similar results. All of these methods show better accuracies and RMS error metrics when compared with the individual sensor estimates, showing that while the IMU method in particular achieves good abilities to recognise a change in state, as measured by the F1 analysis, the other methods are better at aligning this to the true state.

In almost all metrics the LSTM method is shown to have significantly improved results. This would be expected given the LSTM method has the ability to take into account previous state readings while the other methods only take inputs of the current state estimation. This can be seen in the confusion matrices where the IMU, SVM, Bayesian and ANN methods in particular show much worse predictions with higher numbers of screws, i.e. as time progresses. The LSTM method shows much improved consistency in accuracy over the duration of the assembly task.

## V. CONCLUSION

Various sensor fusion methods were analysed and shown to give improved accuracy on state prediction of an industrial style assembly task when compared to each mode of classification individually. CNN based classifiers for both visual recognition and human action recognition from IMU data were developed and showed promising offline performance. Testing was done in a mock assembly task where each individual classifier method was compared along with various fusion methods. Of the fusion methods tested, an LSTM based network architecture performed the best. Development of accurate sensing and perception methods such as this is key to successful HMI, a core development to Industry 4.0.

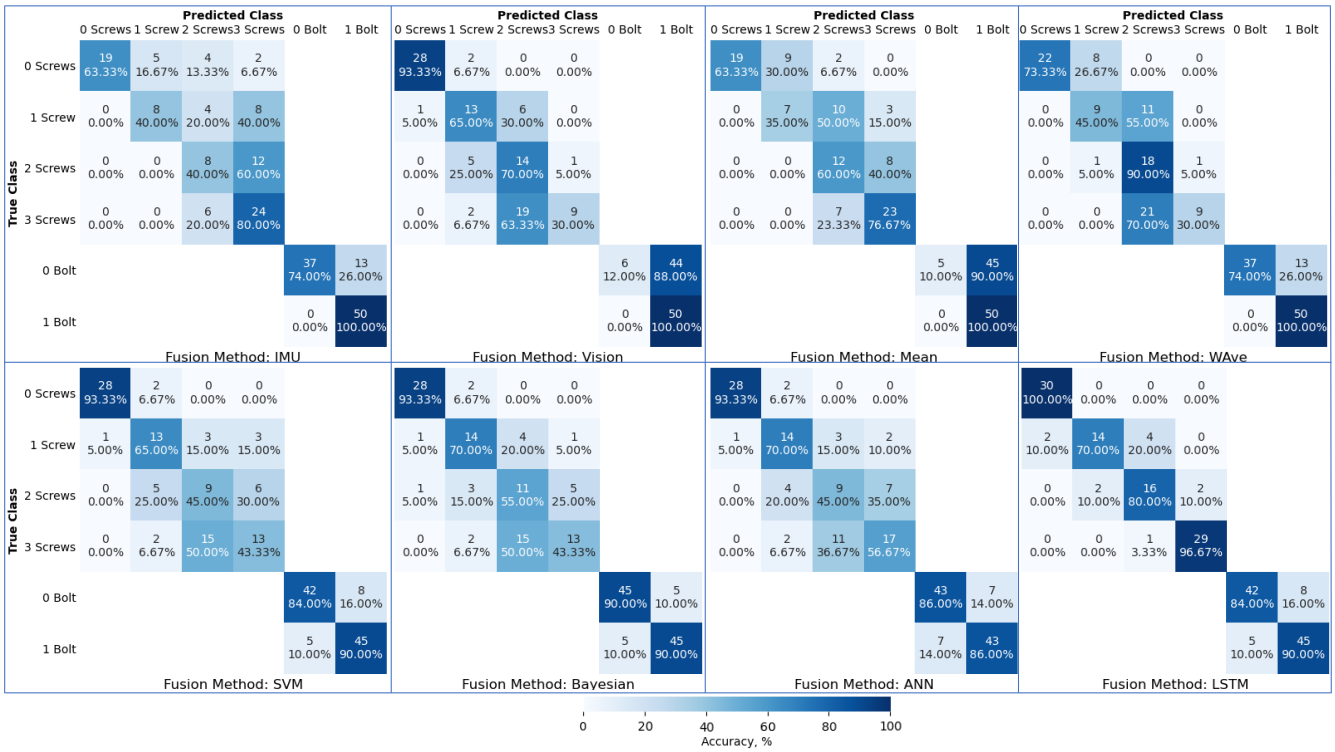


Fig. 10. Confusion matrices for fusion methods analysed

## REFERENCES


[1] M. A. K. Bahrin, M. F. Othman, N. N. Azli, and M. F. Talib, "Industry 4.0: A review on industrial automation and robotic," *Jurnal Teknologi*, vol. 78, no. 6-13, pp. 137–143, 2016.

[2] S. El Zaatari, M. Marei, W. Li, and Z. Usman, "Cobot programming for collaborative industrial tasks: an overview," *Robotics and Autonomous Systems*, vol. 116, pp. 162–180, Jun 2019.

[3] A. Khalid, P. Kirisci, Z. Ghrairi, K.-D. Thoben, and J. Pannek, "A methodology to develop collaborative robotic cyber physical systems for production environments," *Logistics Research*, vol. 9, p. 23, Dec 2016.

[4] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," *Mechatronics*, vol. 55, pp. 248–266, Nov 2018.

[5] B. Gleeson, K. MacLean, A. Haddadi, E. Croft, and J. Alcazar, "Gestures for industry intuitive human-robot communication from human observation," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 349–356, IEEE, 2013.

[6] I. El Makrini, K. Merckaert, D. Lefeber, and B. Vanderborght, "Design of a collaborative architecture for human-robot assembly tasks," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1624–1629, IEEE, Sep 2017.

[7] G. Maeda, A. Maloo, M. Ewerton, R. Lioutikov, and J. Peters, "Anticipative interaction primitives for human-robot collaboration," in *2016 AAAI Fall Symposium Series*, pp. 325–330, 2016.

[8] Y. Liu, Y. Zhou, C. Hu, and Q. Wu, "A review of multisensor information fusion technology," in *2018 37th Chinese Control Conference (CCC)*, pp. 4455–4460, IEEE, 2018.

[9] M. L. Fung, M. Z. Chen, and Y. H. Chen, "Sensor fusion: A review of methods and applications," in *2017 29th Chinese Control And Decision Conference (CCDC)*, pp. 3853–3860, IEEE, Jul 2017.

[10] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[11] Y.-R. Cho, S. Shin, S.-H. Yim, H.-W. Cho, and W.-J. Song, "Multistage fusion and dissimilarity regularization for deep learning," in *2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 586–591, IEEE, Nov 2017.

[12] I. Hwang, G. Cha, and S. Oh, "Multi-modal human action recognition using deep neural networks fusing image and inertial sensor data," in *2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 278–283, IEEE, 2017.

[13] F. Santoso, M. A. Garratt, and S. G. Anavatti, "Visual–inertial navigation systems for aerial robotics: Sensor fusion and technology," *IEEE Transactions on Automation Science and Engineering*, vol. 14, pp. 260–275, Jan 2017.

[14] M. M. Loper, N. P. Koenig, S. H. Chernova, C. V. Jones, and O. C. Jenkins, "Mobile human-robot teaming with environmental tolerance," in *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 157–163, IEEE, 2009.

[15] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pp. 729–738, 2013.

[16] W. Jiang and Z. Yin, "Combining passive visual cameras and active imu sensors for persistent pedestrian tracking," *Journal of Visual Communication and Image Representation*, vol. 48, pp. 419–431, Oct 2017.

[17] M. Al-Amin, W. Tao, D. Doell, R. Lingard, Z. Yin, M. C. Leu, and R. Qin, "Action recognition in manufacturing assembly using multimodal sensor fusion," in *Procedia Manufacturing*, vol. 39, pp. 158–167, Elsevier B.V., 2019.

[18] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, p. 115, Jan 2016.

[19] H. Koskimäki, V. Huikari, P. Siirtola, P. Laurinen, and J. Röning, "Activity recognition using a wrist-worn inertial measurement unit: A case study for industrial assembly lines," in *2009 17th Mediterranean Conference on Control and Automation*, pp. 401–405, IEEE, 2009.

[20] M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H.-P. Tan, "Deep activity recognition models with triaxial accelerometers," in *The Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[21] A. Cherubini, R. Passama, A. Meline, A. Crosnier, and P. Fraisse, "Multimodal control for human-robot cooperation," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2202–2207, IEEE, Nov 2013.

[22] R. Gravina, P. Alinia, H. Ghasemzadeh, and G. Fortino, "Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges," *Information Fusion*, vol. 35, pp. 68–80, 2017.