A Multi-sensor Fusion Approach for Intention Detection

Rahul Kumar Singh[1], Rejin John Varghese[1], Jindong Liu[1], Zhiqiang Zhang[2], Benny Lo[1]

***Abstract*— For assistive devices to seamlessly and promptly assist users with activities of daily living (ADL), it is important to understand the user's intention. Current assistive systems are mostly driven by unimodal sensory input which hinders their accuracy and responses. In this paper, we propose a context-aware sensor fusion framework to detect intention for assistive robotic devices which fuses information from a wearable video camera and wearable inertial measurement unit (IMU) sensors. A Naive Bayes classifier is used to predict the intent to move from IMU data and the object classification results from the video data. The proposed approach can achieve an accuracy of 85.2% in detecting movement intention.**

I. INTRODUCTION

The process of translating intention into action is an intuitive, natural and seamless phenomenon for a healthy individual. However, in cases of people suffering from neuromuscular or cerebrovascular diseases, *e.g.* stroke, cerebral palsy, paraplegia, limb amputation [1], and also in the case of neuromuscular weakening as seen in the elderly population, are not often able to translate the intention into action. In order to effective control exoskeletons [2], soft robotics gloves [3], and prosthetic hands [4], it is imperative to detect intention accurately and translate that into a control signal. The estimated user's intention information could be used to generate a high-level abstract control signal (for reaching, grasping and manipulating an object) simplifying the control mechanism and enabling instant responses. The proposed system uses contextual information from a vision-based sensor (a monocular camera) which captures user's field of view and inertial sensors worn on the upper and lower arm which detect proprioceptive information to detect the user's intention.

II. METHODOLOGY

Our vision plays a key role in motor control providing direction, guidance and feedback for upper and lower limbs movements. During any intended action, we generally try to bring the object of interest into our field of view. Thus, hand-eye co-ordination plays a crucial role in ADL. We used YOLO (You Look Only Once) [5], a CNN (Convolutional Neural Network) based architecture for the object recognition. It was trained on the Coco dataset containing 80 classes. The network predicts the object class probabilities directly from the image in a single evaluation, and also location of the

[1]Rahul K Singh, Rejin J Varghese, Jindong Liu and Benny Lo are with the Hamlyn Centre, Imperial College London, SW7 2AZ, UK {r.singh17, r.varghese15, benny.lo}@imperial.ac.uk

[2]Zhiqiang Zhang is with School of Electronic and Electrical Engineering, School of Mechanical Engineering, University of Leeds, Leeds, LS2 9JT, United Kingdom {z.zhang3}@leeds.ac.uk

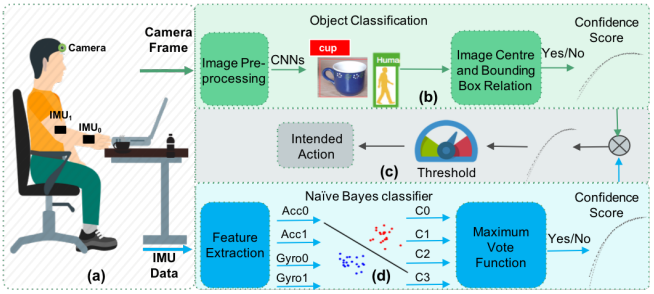

Fig. 1. System architecture showing (a) Sensor placement on the user's body (b)&(d) Confidence score calculation from camera and IMU data (c) Fusing two confidence scores to predict the intended action.

object in the image. Once the objects are detected from image frame, we then convert this information into the likelihood of interested object based on temporal information. The likelihood or visual intent score for the k^{th} object denoted by $P_k[i]$ which increases with time if the object stays in the field of view ($obj_k[i] = 1$) and decreases exponentially if it is not in the field of view ($obj_k[i] = 0$). α_{cam} and β_{cam} are the rate of increase or decay constants of visual score. $obj_k[i]$ denotes output of CNN for k^{th} object in image frame at i^{th} time stamp. P_{k0} is the prior probability value or the biased probability term representing likelihood score obtained from prior knowledge of the k^{th} object.

$$P_k[i] = \begin{cases} (1-e^{-x}) + P_{k0}e^{-x}, & \text{if } obj_k[i] = 1 \\ P_k[i-1]e^{-y}, & \text{if } obj_k[i] = 0 \end{cases} \quad (1)$$

where, x = t[i-1] + α_{cam}(t[i] - t[i-1])

$y = \beta_{cam}$(t[i] - t[i-1])

$P_{k0} = P_k[i]$ & $t[i] = 0$, if, $obj_k[i] = 0$

Movement intention transforms into action through the movement of the upper or lower limbs. In order to counter the 'Midas touch' [6] problem and capture the motor intention, two IMUs - one on forearm (near wrist joint), and the other on upper arm (near elbow joint) were placed on the user. The 3-axis data from accelerometer and gyroscope data from both IMU sensors are used for classifying motion intentions. A 0.5 sec signal length with a stride of 1 on all 4 sets of data was used for feature calculation (mean and variance). The classification was separated into 4 parallel pipelines where each pipeline gives a class output (as shown in Fig.1(d)). The features from these 4 sets of data are then fed into 4 different NB classifiers (namely $NB_0[i]$, $NB_1[i]$, $NB_2[i]$ and $NB_3[i]$) denoting the same intentional action. A voting function is used to obtain the intention output from 4 NB classifiers. In the voting function (given by Eq. 2), each classifier ($NB_k[i]$, where k=0,1,2,3) votes for a class j based on the output of

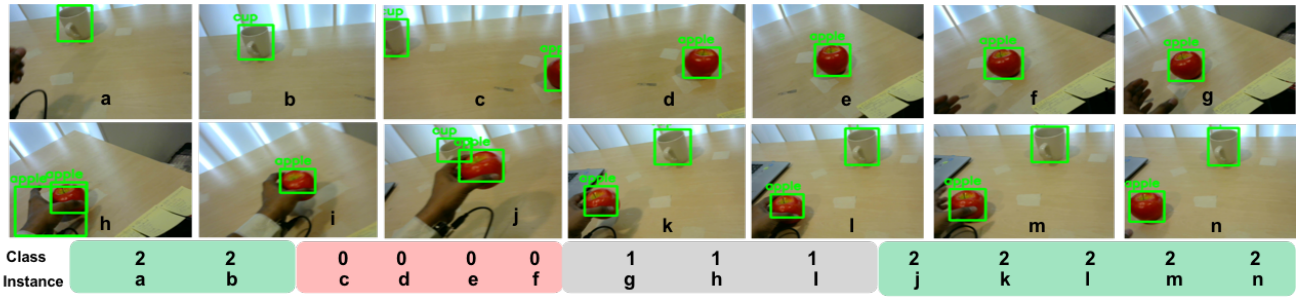

Fig. 2. User's hand (a)-(b)Moves to initial position (c)-(d) Rests on armrest. (g)-(i) Moves towards apple (j)-(n)Places apple to a different location.

classifier denoted by $V_j(k)$. $V_j(k)$ could be 0 or 1 for a class j depending on classifier output. Each vote is summed up to form a total vote $Vote_j[i]$ for each class j. The total vote is then thresholded to obtain the movement intention given by $NB_{out}[i]$ in Eq. 3. V_{max} denotes the voting threshold. Class 1 indicates that the person is reaching out for the intended object while Class 2 means the user placing the object from one place to another and then retracting the hand back to its normal position (*i.e.* on chair armrest). The output of classifier was converted into a confidence score (using Eq. 1) where camera constants (α_{cam} and β_{cam}) are replaced by IMU constants (α_{imu} and β_{imu}) and $k = 0$.

$$Vote_j[i] = \sum_{k=0}^{3} V_j(k) \qquad where,\ j = 0,1,2 \tag{2}$$

$$NB_{out}[i] = \begin{cases} \arg\max(Vote_j[i]), & \text{if } max(Vote_j[i]) > V_{max} \\ NB_{out}[i-1], & \text{otherwise} \end{cases} \tag{3}$$

$$P_{overall}[i]_{Kx1} = (P_{vision}[i]_{Kx1})(P_{imu}[i]_{1x1}) \tag{4}$$

The visual and motion (IMU) scores are updated at 25 Hz and thus, overall score is obtained at the same rate as well. The overall confidence score is obtained from the joint probability (given by Eq. 4), and is used to predict the intention. The parameters α's (α_{cam} and α_{imu}) and β's (β_{cam} and β_{imu}) are person specific and determines the system's responsiveness. $P_{vision}[i] = \{P_1[i], P_2[i], ..., P_k[i]\}$ is a $Kx1$ vector containing visual intent scores of k objects while $P_{imu}[i]$ denotes motor intent score calculated from IMU data.

III. EXPERIMENTAL RESULTS

The CNN model implementation took around 5 sec to process each image frame. Due to this processing time for object recognition, we recorded the data from the camera and two IMUs, along with the time stamp and then processed the data offline. In the experimental setup 5 objects, namely a cup, apple, keyboard, bottle and phone, were used in the experiment and laid on the table. The objects are positioned with sufficient distance apart on the table to minimize overlaps in the field of view. The participant of our study picked up any intended object and placed it on to the designated space. The results of the proposed system (*i.e.* 10 instances of pick up and place of intended objects) are shown in Table I where accuracy for each object for each individual is shown. The constants value (α's and β's), largely depends on the average time between the start and end of the intention. As vision triggers first for intention and takes longer time during the completion of intended action while motor intention are generated after visual intention. Thus, in our experiment, α_{imu} for IMU was set as 2.0 while α_{cam} for camera it was set to 1.0. While the decay constant (β) of confidence score was set to 2.0 for camera (β_{cam}) and 3.0 for IMU (β_{imu}). The reason for keeping β higher than α is to let the system reach to an initial condition quickly, in order to avoid false positives. The threshold for the joint score was set to 0.7.

TABLE I

Accuracy of the proposed intention detection algorithm tested on 5 volunteers with 5 different objects

User/ Object	Apple	Bottle	Cup	Keyboard	Phone
A	90	70	100	90	80
B	80	80	90	100	90
C	90	80	90	100	90
D	90	70	90	90	80
E	90	80	90	100	90
Average	88 ± 4.5	76 ± 5.5	92 ± 4.5	94 ± 5.5	86 ± 5.5

IV. DISCUSSION & CONCLUSION

Our proposed work shows preliminary results of an intention detection system its accuracy in predicting user's intention by using sensors (camera and a pair of IMUs). There are certain limitations to the current implementation such as detecting intention when multiple objects are in the field of view or in some cases when one object might occlude the another object. The current system can be improved by using sensors such as 3D camera which captures depth information, and relative distance between the object from hand. Furthermore, the spatial information of object's location in the image can also potentially increase the accuracy.

REFERENCES

[1] D. Novak and R. Riener, "A survey of sensor fusion methods in wearable robotics," *Robotics and Auto. Sys.*, vol. 73, pp. 155–170, 2015.
[2] P. Heo, G. Gu, S. Lee, K. Rhee, and J. Kim, "Current hand exoskeleton technologies for rehabilitation and assistive engineering," *Int. J. of Precision Eng. and Manufacturing*, vol. 13, no. 5, pp. 807–824, 2012.
[3] C. Y. Chu and R. M. Patterson, "Soft robotic devices for hand rehabilitation and assistance: A narrative review," *Journal Neuroeng. Rehabil.*, vol. 15, no. 1, p. 114, 2018.
[4] R. Clement, K. Bugler, and C. Oliver, "Bionic prosthetic hands: A review of present technology and future aspirations," *The surgeon*, vol. 9, no. 6, pp. 336–340, 2011.
[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.
[6] R. Jacob, "The use of eye movements in human-computer interaction techniques: what you look at is what you get," *ACM Transactions on Information Systems*, vol. 9, no. 2, pp. 152–169, 1991.