Vision-Based Approximate Estimation of Muscle Activation Patterns for Tele-Impedance

Hyemin Ahn, Youssef Michel, *Graduate Student Member, IEEE*, Thomas Eiband, *Graduate Student Member, IEEE*, and Dongheui Lee, *Senior Member, IEEE*

***Abstract*—It lies in human nature to properly adjust the muscle force to perform a given task successfully. While transferring this control ability to robots has been a big concern among researchers, there is no attempt to make a robot learn how to control the impedance solely based on visual observations. Rather, the research on tele-impedance usually relies on special devices such as EMG sensors, which have less accessibility as well as less generalization ability compared to simple RGB webcams. In this letter, we propose a system for a vision-based tele-impedance control of robots, based on the approximately estimated muscle activation patterns. These patterns are obtained from the proposed deep learning-based model, which uses RGB images from an affordable commercial webcam as inputs. It is remarkable that our model does not require humans to apply any visible markers to their muscles. Experimental results show that our model enables a robot to mimic how humans adjust their muscle force to perform a given task successfully. Although our experiments are focused on tele-impedance control, our system can also provide a baseline for improvement of vision-based learning from demonstration, which would also incorporate the information of variable stiffness control for successful task execution.**

***Index Terms*—Deep learning for visual perception, telerobotics and teleoperation, compliance and impedance control.**

I. Introduction

IMAGINE that you teach a robot how to open a jar - which requires you to apply the force to the jar properly, synchronized with your hand motion. You fully demonstrate to the robot how to move its end-effector to perform a task, based on the motion capture device. In this scenario, the robot would fail this task if it only adjusts its end-effector trajectory. To achieve the goal, it also needs to learn when and how to adapt its end-point stiffness. In this regard, it is possible for robots to understand how humans adjust their end-point stiffness, by using various auxiliary sensors such as a grip-force sensor [1], a joystick-like device [2], or EMG sensor [3], [4]. However, relying on a special device could be a bottleneck, since it diminishes the accessibility as well as the generalization ability. Therefore, we believe that it would be beneficial to employ other affordable and general sensors - like RGB webcams.

In this letter, we propose a framework for vision-based and tele-impedance control, which relies only on RGB images without visible markers. We show that recent improvement in visual perception based on deep neural networks [5], [6], [7] can be a highly appealing solution, for building an easy-to-use, accessible, and generalizable tele-impedance control system. Experimental results show that our framework enables a robot to mimic how a human adjusts his/her end-point stiffness to successfully perform a task, based on image inputs obtained from an affordable RGB webcam.

Our vision-based tele-impedance system employs the approximate estimation of discretized muscle activation patterns for adjusting the robot's end-point stiffness parameters to successfully perform a task. To obtain this estimation, we propose a deep neural network-based model, which can infer the discretized muscle activation pattern from a short video observation of the human limb. The model output includes (1) whether the muscle is activated or not, and (2) whether the muscle activation is increasing, stable, or decreasing. Although this would not provide detailed information such as the magnitude or orientation of the end-point stiffness, our model is able to obtain some general information of human muscle activation during the task.

The advantages of our system are as below:

- Unobtrusive: It does not require humans to put any markers on their body parts.
- High accessibility: It can be applied with any type of affordable RGB webcam.
- High generalization ability: It can be also successfully applied to human users who did not participate in the training data collection process.

These advantages are shown through our experiments, which include the real-world robot demonstration of tele-impedance control. We believe that our work would be a baseline study for future researchers of vision-based impedance understanding and control, such as vision-based learning from demonstration

Manuscript received 7 February 2023; accepted 12 June 2023. Date of publication 7 July 2023; date of current version 14 July 2023. This letter was recommended for publication by Associate Editor Z. Min and Editor C. Cadena Lerma upon evaluation of the reviewers' comments. This work was supported in part by the Helmholtz Association, and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP), grant funded by the Korea Government (MSIT) through Artificial Intelligence Graduate School Program under Grant 2020-0-01336. *(Youssef Michel and Thomas Eiband contributed equally to this work.) (Corresponding author: Dongheui Lee.)*

Hyemin Ahn was with the German Aerospace Center (DLR), 82234 Wessling, Germany. She is now with Artificial Intelligence Graduate School (AIGS), Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, South Korea (e-mail: hyemin.ahn@unist.ac.kr).

Youssef Michel is with the Chair of Human-centered Assistive Robotics, Technische Universität München (TUM), 80333 Munich, Germany (e-mail: youssef.michel.1992@gmail.com).

Thomas Eiband is with the Institute of Robotics and Mechatronics, German Aerospace Center (DLR), 82234 Wessling, Germany (e-mail: thomas.eiband@dlr.de).

Dongheui Lee is with Autonomous Systems, Technische Universität Wien (TU Wien), 1040 Vienna, Austria, and also with the Institute of Robotics and Mechatronics, German Aerospace Center (DLR), 82234 Wessling, Germany (e-mail: dongheui.lee@tuwien.ac.at).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2023.3293275, provided by the authors.

Digital Object Identifier 10.1109/LRA.2023.3293275

which incorporates the information of the end-effector's position as well as stiffness.

II. Related Works

Humans are characterized by their unique skill to perform delicate physical interactions, thanks to their ability to adapt their end-point force and impedance. Inspired by that, several works have aimed at transferring such human interaction capabilities to robots, encouraged by the emergence of a new generation of compliant, torque-controlled robots. For example, [8] proposed an adaptive control framework based on human-motor control theory. Their framework can adapt the robot's feed-forward force, impedance, and reference trajectory to reach the optimal interactive behavior while maintaining a compromise between stability and efficiency. This controller was later extended in [9] to perform contact tasks such as cutting and haptic exploration. There also exist works that focused on developing interfaces to transfer human's variable impedance control skills to robots during teleoperated task execution. This was done using a grip-force sensor in [1], and a joystick-like device in [2] to capture in real-time information about the human end-point stiffness, or by learning a model for stiffness adaptation from human demonstrations [10], with the purpose of commanding the robot's reference stiffness profiles during tele-operation.

Along the same lines, the use of EMG also received significant attention in variable impedance transfer from humans to robots. This was motivated by a seminal concept of tele-impedance [11], which was proposed as a possible alternative to the standard bilateral tele-operation. A human operator commands a remote impedance-controlled robot with real-time motion commands captured via optical tracking, and stiffness profiles estimated from muscle activation measured with EMG. Similarly, Yang et al. [3] used EMG to compute stiffness profiles for commanding the remote robot in standard bilateral tele-operation, where the operator received haptic feedback from the environment. Peternel et al. [4] proposed a framework that used EMG together with force sensing and a motion capture system to derive suitable hybrid force-impedance control strategies from human demonstrations for various contact tasks.

However, these approaches have less accessibility as well as less generalization ability since they require costly sensors. Applying affordable vision-based systems to transfer human impedance control skills to robots would be a big game changer in this research field. Unfortunately, as mentioned before, there have been no attempts to address vision-based robot impedance control. Instead, there exist works called optical myography, which objective is to estimate finger movement. To do this, [12], [13], [14] use a fixed testbed and fasten the human arms on it. Afterward, visual observation is obtained from the 10 AprilTag markers [12], [13] or a single undifferentiated marker (i.e., plain sticker) [14] that needs to be attached to the subject's forearm. These works were also extended to human hand gesture recognition with multiple orientation-free markers on the front and back of the human arm [15]. The advantage of optical myography is the use of inexpensive camera sensors, which makes the experiment process easier than other works with force or position sensors. However, existing optical myography studies require human subjects to attach visible markers, and many of them require humans to constrain their arm position. Compared to this, our system does not require any markers to be attached, and human subjects can move their arm within the range for target muscle regions to be visible.

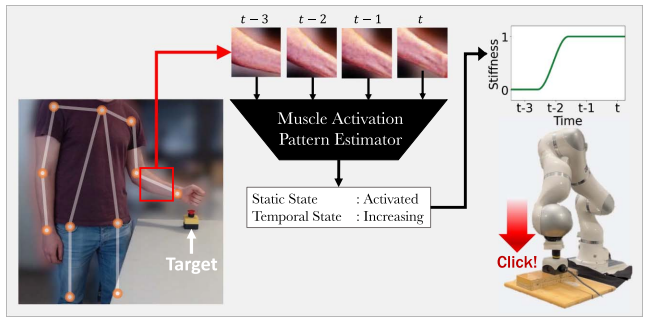

Fig. 1. Overview of how the proposed vision-based tele-impedance system works. It first estimates a human pose to obtain the image patch of the target limb (i.e., forearm). Then, the image patches are given to the proposed muscle activation pattern estimator (MAPE), resulting in a static state (activated, inactivated) and temporal state (increasing, stable, decreasing). Based on this, a proper end-point stiffness profile is obtained and the robot controls its end-point stiffness to perform a given task, by following how the human adapted one's stiffness to complete the task.

We admit that there could be infeasibility when relying only on images to obtain detailed and exact information on how humans control their impedance. However, just as we know whether an object is heavy or not by observing someone's arm lifting the object, humans are able to estimate approximate information of muscle activation by recognizing the visual pattern of muscles [16]. Even if it would be challenging to measure the exact value of human arm stiffness based on images, it would be possible for us to infer with images whether the human is stiffening up his/her arm or not. This approximate estimation of muscle activation patterns can be beneficial for a robot to control its end-point stiffness to perform a given task properly.

III. Methodology

A. Overall Structure

Fig. 1 shows how the proposed system works. On the left side, it shows how our system obtains the visual input for muscle activation pattern recognition. From the RGB image, our system first obtains the human pose information based on the lightweight 3D pose estimator from [17]. Based on that, a square image patch is automatically cropped near the region where the target limb is. Note that our system does not require the human to locate the target limb in a certain image region since the pose estimator can find the target limb position. We choose the forearm as the target since the visual appearance of its muscles would represent the wrist stiffness, which has consequently an effect on human's hand end-point stiffness [18]. After collecting a set of image patches for a short time duration, it is used as an input for the proposed muscle activation pattern estimator (MAPE). As shown in Fig. 1, MAPE results in two muscle activation states, which are static state and temporal state. The static state represents whether the muscle is activated or not, and the temporal state represents whether the degree of muscle activation is increasing, stable, or decreasing. Both states are necessary for a robot to adapt its end-point stiffness to conduct

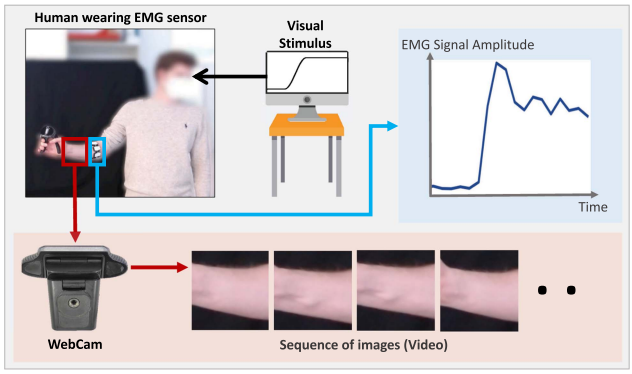

Fig. 2. Illustration of how our dataset is collected from human subjects. While a sequence of images of the forearm is collected by an affordable RGB webcam (red box), an EMG signal is also collected from the bracelet-shaped sensor (light blue box). After finishing this process, the collected raw data is properly preprocessed such that a set of triplets of images, static states, and temporal states can be obtained.

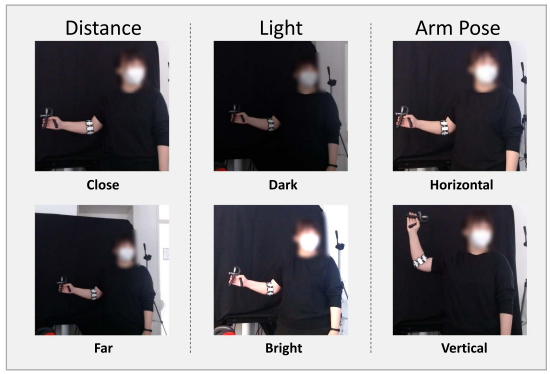

Fig. 3. Illustration of various conditions when our dataset is collected from human subjects. Conditions such as distance, light, and arm pose are varied in two ways such that the data collection process can be conducted at least eight times for each subject.

the task. A robot can decide whether to stay inactivated (low stiffness) or activated (high stiffness) using the static state, and whether to prepare the transition from the inactivated state to the activated state using the temporal state. Note that MAPE is trained in a supervised way, based on the RGB images from an affordable webcam, and static/temporal states are annotated from an EMG sensor.

For enhancing the performance of MAPE, we propose an additional approach such as *'visual data augmentation'* in the training process. Its goal is to randomly apply several transformation techniques to the given muscle image patches, such as rotation, translation, flipping, color adjustment, and background change. By doing so, it is possible for MAPE to learn how to be robust for different image conditions such as light, human pose, and background.

In addition, we also apply *'visual calibration'* in the training process. It generalizes the performance of MAPE for various human subjects, by reducing the effect of individual muscle's visual traits. To do this, it collects the muscle image patches from a human subject when one releases or applies force to the muscle. Afterward, when new image patches are given as inputs from the human subject, it generates the calibrated image features by considering the activated and inactivated muscle images of that target human subject.

B. Dataset

1) Collection: To train the muscle activation pattern estimator (MAPE), we build a dataset consisting of images observing a human's target limb (forearm) as well as corresponding discrete state labels of muscle activation patterns. The labels include the static and temporal states, where the static states denote whether the muscle is activated or not, and temporal states denote whether the degree of muscle activation is increasing, stable, or decreasing.

To collect the dataset, we prepare an environment as shown in Fig. 2. It shows a webcam (Logitech C920) and a bracelet sensor named Myo from Thalmic Labs, which arranges 8 EMG sensors around the forearm. We ask subjects to wear it on their forearm near their elbow (see the cyan box in Fig. 2). The camera observes the subject with 10 fps and a resolution of 1920×1080. Images are collected within a black background, to simplify the data augmentation process (i.e., random background). After setting up the camera, we instruct the subject to apply the force to the gripper by applying a tight grasp, synchronized with the visual stimulus shown on the screen. Here, the visual stimulus is a periodic square wave signal with a length of 120 seconds and a time period of 4 seconds. We ask the subjects to apply the grasp force of approximately 80% of their maximum voluntary contraction [19] when the stimulus value is high, otherwise release the grip force. According to [18], this instruction would also affect the human end-point stiffness.

Let us denote $s(t) \in R^8$ as an EMG signal consisting of eight channel values, which are the measurements from eight EMG sensors of our bracelet sensor Myo. Then, its amplitude is calculated as $a(t) = \|s(t)\|_2$. The rightmost graph in Fig. 2 shows an exemplary EMG signal amplitude $a(t)$ obtained from the human subject when one controls the grip force while following the visual stimulus.

To increase the generalization ability of our model, we collect the dataset from six human subjects (3 males, 3 females / 3 Caucasians, 3 Colored / Age range 28∼44 /All in normal physiques). In addition, as shown in Fig. 3, the data collection processes based on the visual stimulus are performed 8 times by combining the following settings. We captured two different distances between the camera and the human, in particular 1 m and 1.5 m. Also, we consider two different light conditions based on the intensity of the light source. Finally, two different arm poses were considered.

2) Annotation: For collected images, we crop the area of the forearm according to the estimated wrist and elbow poses from [17], as the red box shown in Fig. 2. Since our goal is to estimate muscle activation based on observed visual patterns of the muscle, we exclude the visual information of hands. When training a model, an image patch inside this red box is randomly rotated or translated for data augmentation.

For each cropped image showing the bare forearm, we need labels of whether the muscle is activated or not, and whether the muscle activation is increasing, stable, or decreasing. To annotate these discrete labels, we employ the collected EMG

signals. We first resample EMG signals to 30 fps, and apply zero-phase digital filtering to the signal [20].

To annotate a static state at time t, we define a threshold for the amplitude of the preprocessed EMG signal and label the state as 'activated' if the amplitude at time t is larger than the given threshold. The threshold is defined based on the normalized EMG signal amplitude $n(t)$, which is normalized as $n(t) = (a(t) - \min a(t))/(\max a(t) - \min a(t))$, where $a(t)$ denotes the original EMG signal amplitude at time t. Then, the static state is defined as 'activated' if $n(t)$ is larger than δ_{act}, and vice versa. Here, δ_{act} is empirically chosen uniquely for each human subject, by considering how the individual's muscle shape changes with respect to $n(t)$. If one's muscle looks activated after $n(t)$ is larger than a certain value, that value is chosen as δ_{act}.

To annotate the temporal state, we obtain the slope $s(t)$ of $n(t)$ by fitting a linear regression model to $n(t-d) \sim n(t+d)$. A temporal state is defined as 'increasing' if $s(t) > \delta_{inc}$, as 'decreasing' if $s(t) < \delta_{dec}$, and as 'stable' if $\delta_{dec} < s(t) < \delta_{inc}$. Thresholds $\delta_{inc}, \delta_{dec}$ are empirically chosen for each subject. Finally, the number of obtained $(\texttt{Images}, \texttt{Static State}, \texttt{Temporal State})$ was 6,764.

C. Muscle Activation Pattern Estimator (MAPE)

The goal of our muscle activation pattern estimator (MAPE) is to recognize the approximate label of muscle activation patterns at each time stamp $t = 1 \dots T$ based on the visual information. However, the temporal state label of whether the muscle activation is increasing or not would not be captured from a single image. Therefore, MAPE gets an input of $\boldsymbol{V}_t \in R^{N\times 3\times H\times W}$, where a short video clip consists of N frames $\{\boldsymbol{I}_{t-N+1} \dots \boldsymbol{I}_t\}$. Here, $\boldsymbol{I}_t \in R^{3\times W\times H}$ is an RGB image observing the bare forearm at timestamp t.

MAPE generates two score vectors S_t and T_t based on $\boldsymbol{V}_t$, such that: $[S_t; T_t] = \texttt{MAPE}(\boldsymbol{V}_t)$, where $S_t \in \mathbb{R}^2, T_t \in \mathbb{R}^3$. Here, S_t denotes an estimation score for the static state label at the time stamp t. It would be recognized that the muscle is activated if $S_t[1] < S_t[2]$, where $S_t[i]$ is the i-th element of S_t. Similarly, T_t denotes an estimation score for the temporal state label at the time stamp t. If $\arg\max_i T_t[i] = 1$, it would be recognized that the muscle activation is decreasing. Otherwise, it would be recognized that the trend of muscle activation is stable or increasing if $\arg\max_i T_t[i]$ is 2 or 3.

Fig. 4 visualizes the structure of the proposed neural network-based MAPE. In this figure, it is assumed that the input consists of 4 images. MAPE first extracts a set of image features from V_t, based on the ResNet [5]. Let $h_k \in R^D$ denote a feature extracted from the k-th image $\boldsymbol{I}_k$ consisting $\boldsymbol{V}_t$. After extracting $\boldsymbol{H}_t \in R^{N\times D}$ which consists of $\{h_{t-N+1} \dots h_t\}$, $\boldsymbol{H}_t$ is fed to recurrent neural networks (RNNs) [21]. After $\boldsymbol{H}_t$ passes the RNNs, the resulting vector $v_t \in R^D$ is mapped to S_t and T_t with two separate fully connected layers.

D. Visual Data Augmentation

To train MAPE, it is crucial to collect a sufficient dataset containing triplets of $D = \{\boldsymbol{V}_t, S_t, T_t\}_{t=1\dots T}$. To do this, as we mentioned above, we recruited six human subjects and collected a dataset from a different light, camera distance, and human pose

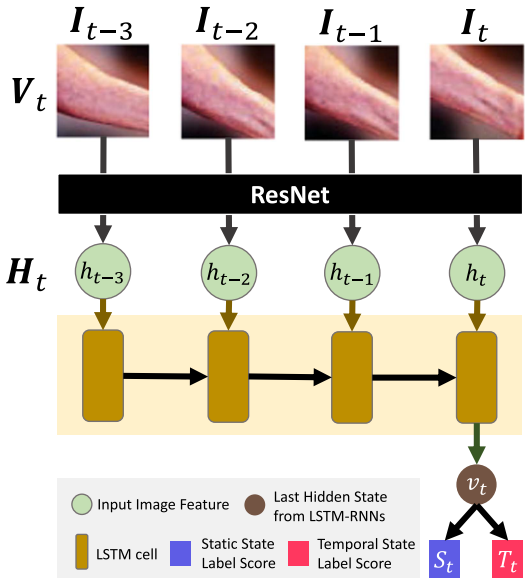

Fig. 4. Structure of the muscle activation pattern estimator (MAPE).

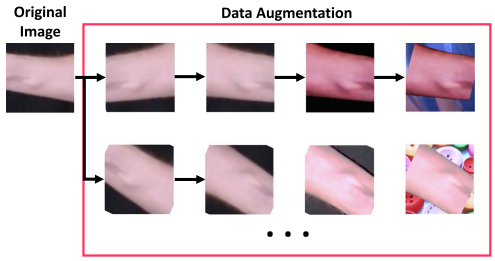

Fig. 5. Data augmentation process for improving the generalization performance of MAPE. After randomly rotating or translating the image patch cropped near the target region (i.e., forearm), it is randomly flipped in the horizontal or vertical direction. Then, the brightness, contrast, and saturation of the image are randomly adjusted. Finally, the background is replaced with various random images from https://picsum.photos/.

conditions. However, even if we considered various conditions in the data collection process, we still cannot guarantee that the collected dataset is perfect since it is quite challenging to satisfy diversity among all possible different conditions.

Therefore, to increase the generalization performance of the model, we augment our input data $\boldsymbol{V}_t$ based on the below process shown in Fig. 5.

1) Obtain the image patch of the forearm near the wrist (see red box in Fig. 2).
2) Crop the image inside the red box, and randomly rotate and translate that cropped image.
3) Flip the cropped image horizontally or vertically with a probability of 0.5.
4) Randomly adjust the brightness, contrast, and saturation of the cropped image.
5) Replace the black background with various random images from https://picsum.photos/.

The second and third procedures are to compensate for the disadvantage of our data that is only collected from two types of human postures. The fourth and fifth procedures are to increase the generalization performance of the model with various brightness conditions and backgrounds. The entire data augmentation

Fig. 6. Comparison between subjects with different muscle shapes and deformation patterns. Compared to Subject 6, Subject 1 has a smaller muscle size and the variation of its deformation pattern is also less than Subject 6. In addition, it is shown that the activated muscle image of Subject 1 looks similar to the inactivated muscle image of Subject 6.

process is performed per each training iteration, such that more randomly augmented data can be used for training as the number of iterations increases.

E. Visual Calibration

However, even if we train MAPE with the augmented dataset, we find out that the vanilla MAPE sometimes fails to obtain the generalized result. Fig. 6 shows one of the examples that provokes the cases of failure in generalization. It shows that the visual pattern change between inactivated and activated muscles of Subject 1 is less distinguishable compared to Subject 6. Also, the image of the activated muscle of Subject 1 looks similar to the inactivated muscle of Subject 6, rather than the activated muscle of Subject 6. In this case, if the proposed vanilla MAPE is trained with images from Subject 1, the 'inactivated' muscle image of Subject 6 can be recognized as 'activated' due to this visual similarity. On the contrary, if MAPE is trained only based on the images of Subject 6, it would be challenging for the model to discriminate the muscle activation status of Subject 1 since his/her muscle activation change is less visible.

Based on this observation, we conclude that an additional approach that can reduce the impact of an individual muscle's visual characteristics is necessary. Therefore, we introduce the approach named 'Visual Calibration', which can generalize the performance of MAPE for various human subjects. To do this, it first requires human subjects to release or apply the gripping force and collects images from inactivated and activated muscles. Let $\boldsymbol{I}_{inact}$ and $\boldsymbol{I}_{act}$ denote the collected images of inactivated and activated muscles from this process. Then, MAPE with Visual Calibration extracts image features from $\boldsymbol{I}_{inact}$ and $\boldsymbol{I}_{act}$ based on ResNet. Let h_{inact} and h_{act} denote the obtained image features. Then, the set of image features $\boldsymbol{H}_t = \{h_{t-N+1} \dots h_t\}$ which was obtained from the original input $\boldsymbol{V}_t$ is calibrated based on h_{inact} and h_{act}, such that $\boldsymbol{H}^c_t = [h_k - h_{inact}; h_k - h_{act}]_{k=t-N+1,\dots,t}$ can be obtained. Finally, the calibrated features $\boldsymbol{H}^c_t$ are given as an input to the LSTM-RNNs, and final results (S_t, T_t) are obtained. Fig. 7 summarizes how MAPE with Visual Calibration works. Note that the same ResNet is used to extract features from $\boldsymbol{V}_t$, $\boldsymbol{I}_{inact}$ and $\boldsymbol{I}_{act}$, so that the efficiency can be improved in terms of the number of parameters.

F. Implementation Details

When training MAPE, we use $\boldsymbol{V}_t \in R^{4\times3\times112\times112}$, where four images $\boldsymbol{I}_{t-3}, \boldsymbol{I}_{t-2}, \boldsymbol{I}_{t-1}, \boldsymbol{I}_t$ of size $3 \times 112 \times 112$ are used to construct $\boldsymbol{V}_t$. In experiments, our video clip is 10 fps, such

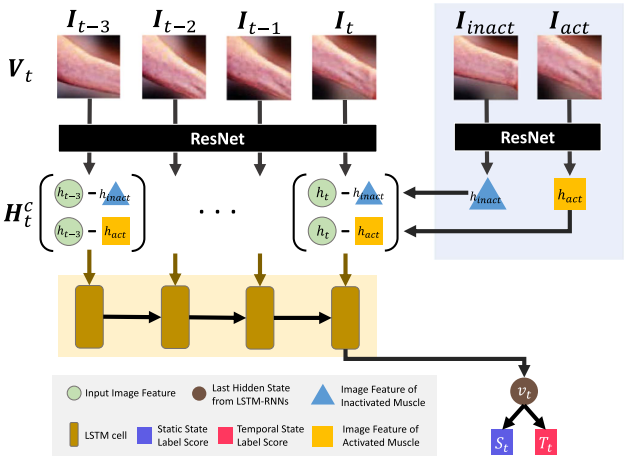

Fig. 7. Network architecture of the proposed machine activation pattern estimator (MAPE) when visual calibration is applied.

that the video clip $\boldsymbol{V}_t$ contains information during 0.4 s. After extracting a set of image features $\boldsymbol{H}_t \in R^{4\times2048}$ from $\boldsymbol{V}_t$ based on ResNet-50, it is normalized based on its L2-norm magnitude and is fed to LSTM-based RNNs which dimension of the hidden state is 2048. When Visual Calibration is added to MAPE, input for the LSTM-RNNs is $\boldsymbol{H}^c_t = [h_t - h_{inact}; h_t - h_{act}]$, where h_t, h_{inact}, and h_{act} are image features after L2-norm based normalization. After LSTM-RNNs process the input $\boldsymbol{H}_t$, its last hidden state vector becomes $v_t \in R^{2048}$, and it is mapped into S_t and T_t based on two separate fully connected layers. To train vanilla MAPE or MAPE with Visual Calibration, a cross-entropy loss function is used for each S_t and T_t, and Adam optimizer [22] with a learning rate of 0.0001 is employed. At each iteration, a batch size of 64 is sampled from the training dataset after the data augmentation. The training procedure of MAPE takes less than one day with a single GPU.

IV. Experiments

A. Qualitative Results

Our dataset for training MAPE consists of RGB images as well as static/temporal state annotations collected from six different human subjects. After training MAPE with a dataset from five human subjects, we obtain its qualitative results as shown in Fig. 8 by giving data from the unseen human subject as inputs. Compared to the ground truth of static/temporal state labels, the estimation result shows that the proposed system combining visual calibration as well as data augmentation to MAPE generates the most reliable result (see the orange block in Fig. 8). In addition, it is shown that the proposed system without visual calibration or data augmentation generates less reliable results (see gray blocks in Fig. 8). Note that relevant quantitative results will be also shown in the next paragraph. Based on this, we argue that the proposed visual calibration as well as data augmentation are crucial for enabling MAPE to understand the changes in muscle deformation patterns of the unseen human subject. Therefore, it can be concluded that the proposed two auxiliary processes are also important for increasing the generalization ability of our vision-based tele-impedance control system.

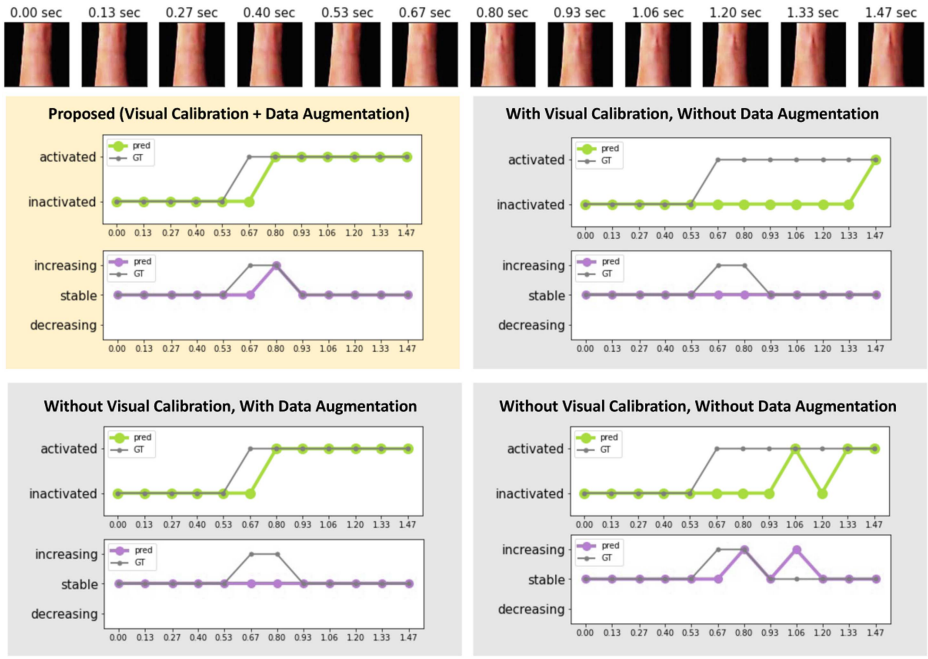

Fig. 8. Qualitative result from MAPE when it is tested to the dataset of the human subject which is unseen during the training phase. Note that input images are modified for better visualization. The result shows that the proposed system with visual calibration as well as data augmentation results in the most reliable estimation compared to other cases.

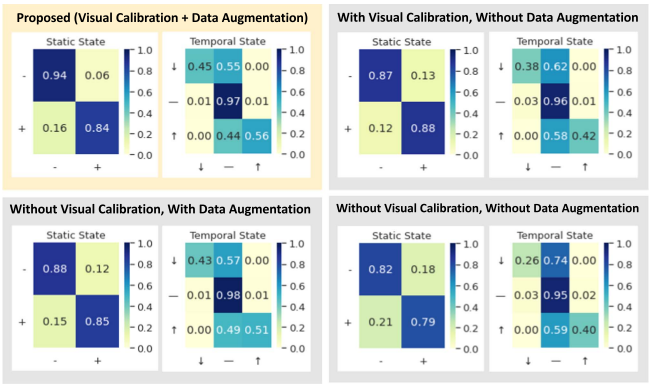

Fig. 9. Confusion matrices obtained from the quantitative ablation study. On matrices for the static state, - denotes inactivated and + denotes activated state. On matrices for the temporal state, ↓ denotes decreasing, — denotes stable, and ↑ denotes increasing state. The result shows that the proposed system with visual calibration and data augmentation is the best when considering both static and temporal state estimation.

TABLE I
SCORES OF STATIC STATE ESTIMATION WITH AND WITHOUT DATA AUGMENTATION (DA) AND VISUAL CALIBRATION (VC)

	Proposed		w/o DA		w/o VC		w/o DA&VC	
	-	+	-	+	-	+	-	+
Precision	**0.937**	0.844	0.872	**0.882**	0.878	0.849	0.817	0.790
Recall	0.886	**0.912**	**0.905**	0.842	0.883	0.843	0.835	0.769
F1 Score	**0.911**	**0.877**	0.888	0.861	0.881	0.846	0.826	0.779

- Denotes inactivated, and + denotes activated. The bold fonts denote the highest score when comparing proposed, w/o DA, w/o VC, w/o DA & VC.

B. Quantitative Results

To measure the performance of MAPE, we perform 6-fold cross-validation, by dividing the dataset with respect to the human subject identity. Fig. 9 presents the confusion matrices from the proposed system with and without visual calibration or data augmentation. As expected from the qualitative results, it is shown that the system with both visual calibration and data augmentation produces the best result in both static and temporal state estimation.

Compared to the high performance of static state estimation, the accuracy of temporal state estimation is not as high. However, we want to emphasize that the most threatening error occurs if MAPE confuses 'increasing' with 'decreasing'. And note that this does not happen in all cases of this ablation study using MAPE. When the ground truth label is 'increasing' or 'decreasing', only the confusion with 'stable' occurs, which can be easily corrected by post-processing techniques such as filtering or Finite State Machine (FSM). To check how we used FSM for post-processing, please check our supplementary material.

To summarize the performance in a better way, we present the precision, recall, and F1 score of state estimation as shown in Tables I and II. Even the systems without visual calibration or data augmentation sometimes perform better than the proposed system in terms of precision or recall, it is shown that the F1 score of the proposed system is higher in all cases. Based on this, we claim auxiliary approaches such as visual calibration and data augmentation are crucial for improving the generalization ability of the MAPE.

C. Real-World Demonstrations

We also validate the scalability of our system by applying the MAPE to the tele-impedance control of a real robot. Our task is to enable a robot to push the emergency button that needs high enough stiffness to be successfully pushed. To this end, our robot uses variable stiffness based on our vision-based

TABLE II
SCORES OF TEMPORAL STATE ESTIMATION WITH AND WITHOUT DATA AUGMENTATION(DA) AND VISUAL CALIBRATION(VC)

	Proposed			w/o DA		
	↓	—	↑	↓	—	↑
Precision	**0.448**	0.974	**0.563**	0.379	0.956	0.416
Recall	**0.753**	**0.921**	0.770	0.542	0.905	0.680
F1 Score	**0.561**	**0.947**	**0.650**	0.446	0.930	0.516
	w/o VC			w/o DA&VC		
	↓	—	↑	↓	—	↑
Precision	0.425	**0.979**	0.508	0.262	0.952	0.404
Recall	0.729	0.916	**0.867**	0.459	0.895	0.604
F1 Score	0.535	0.947	0.641	0.334	0.923	0.484

↓ Denotes decreasing, — denotes stable, and ↑ denotes increasing. The bold fonts denote the highest score when comparing proposed, w/o DA, w/o VC, w/o DA & VC.

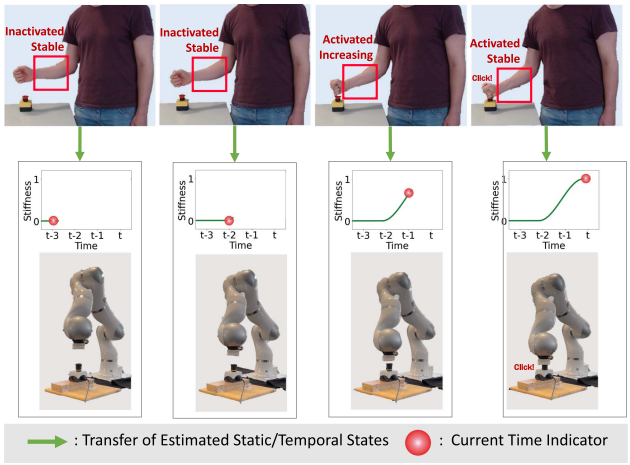

Fig. 10. Our real demonstration for the task of push-button. While a human subject is lowering one's arm by synchronizing to a pre-defined robot's vertical downwards motion, our system captures the forearm region of the subject and estimates its static/temporal states. The estimation result is transferred to the robot, and the robot generates a proper stiffness profile. Based on this, the robot successfully executes the task by adapting its end-point stiffness as a human does to complete the task.

tele-impedance framework. Its end-point stiffness will follow the human strategy, which increases the stiffness only when necessary, and otherwise being compliant during a free motion to reduce the metabolic cost [23]. Fig. 10 shows how our system works with a real robot. We move the robot's end-effector to be located around 20 cm above the button. In the meantime, the human subject makes a fist with one hand and locates it to be around 20 cm above another button. Then, we instruct the subject to successfully push the button by vertically moving down one's hand toward the button. As we focus on the muscle activation estimation part, we simplified the experimental setup by commanding the robot end-effector with a pre-programmed vertical downwards motion towards the button, with a velocity of 5 cm/sec. The subject synchronizes one's motion by following the robot's motion, so that the human can press the button at the same time as the robot does. In addition, it is required for the subject to increase one's grip force properly when the robot's end-effector is about to push the button.

The estimated static/temporal state information is transferred to the robot control loop whenever the MAPE outputs a prediction. To command the robot's stiffness, we devise a Finite State Machine (FSM) that acts on the incoming static/temporal state estimation while also filtering out possible noisy states predicted by MAPE. Based on the internal state of FSM, the cartesian robot stiffness in the direction of motion (z axis relative to the robot base frame) is set. For instance, in Fig. 10, when MAPE estimates the static/temporal states as 'activated' and 'increasing', our FSM transits to the 'increasing' state, and the robot starts smoothly increasing its stiffness profile from a low value to a high value. When the stiffness reaches its maximum, the FSM transitions to and stays in the 'high' state while the static state is 'activated' and the temporal state is 'stable'. Thereby, the robot adapts its end-point stiffness to mimic the same human behavior, allowing it to complete the task, by applying the necessary force needed to push the button.

Our supplementary video shows more examples of real-world demonstrations with various subjects, including ones who did not participate in the data collection process. Our system can be applied to various humans, thereby we claim the generalization ability of our system and the feasibility to be deployed in a real-world scenario is demonstrated.

V. DISCUSSION AND FUTURE WORKS

In this letter, our goal is to transfer the human ability of end-point stiffness adaptation to the robot solely based on RGB images. To do this, we introduce a model named muscle activation pattern estimator (MAPE). Based on the given image frames, MAPE can infer (1) whether the muscle is activated or inactivated (static states), as well as (2) whether the degree of muscle activation is increasing, stable, or decreasing (temporal states). To train MAPE, we collect triplets of images, static states, and temporal states, from six human subjects in various conditions.

However, since the collected dataset can be biased, we also suggest visual data augmentation as well as visual calibration. Visual data augmentation is to randomly apply various image transform techniques to the input image so that MAPE can adapt to images with various conditions. The visual calibration is to reduce the effect of personal factors such as the size or visibility of individual muscles on the forearm. Our experiment results show how MAPE works, and how much the visual data augmentation and visual calibration processes can enhance the performance of MAPE in a qualitative and quantitative way. Finally, we show a real-world demonstration that solves the push-button task, by applying the estimated muscle activation patterns to adapt the robot's end-point stiffness.

Our current system has several open challenges to address. First, it is not able to obtain an exact measure of human stiffness. But especially for tele-impedance applications, this might not be of major importance as long the robot is commanded with adequate impedance. This is the case for instance in EMG-based tele-impedance where scaling factors are commonly used to map the human arm stiffness to the cartesian robot stiffness [11]. Second, our estimation is limited to only a one-dimensional stiffness estimation. For our simplified experiment scenarios, this was sufficient. However, more complex tasks such as screwing would require more complex stiffness settings where the magnitude and orientation of the stiffness ellipsoid are specified.

Third, various and distinct individual characteristics such as limb size, fat, and body hair, may impact the visibility of the

muscle and subsequently affect the accuracy of our MAPE. We observed these effects during our experiment, and believe that including a greater variety of individuals in our training data could increase the generalization capacity of our network, making it more robust to these factors. Fourth, the current system cannot differentiate force from stiffness, while an increase in muscle activation can be due to an increase in force or stiffness. To make this differentiation between force and stiffness, [11] stated that the computation of two decoupled subspaces, namely, the force- and stiffness-generating subspaces, are required. However, this differentiation was not the primary focus of our study. Instead, we simplified the assumption that muscle activation is highly related to stiffness. Our experimental results show that this assumption is empirically shown to be feasible for our purpose, which is to provide a practical and usable method for vision-based tele-impedance control. We believe the current limitation would be supplemented by our future work, and would like to highlight more on how we validate the feasibility of vision-based muscle activation pattern estimation, as well as the possibility of using the estimation results for successful tele-impedance control.

Finally, we would like to emphasize that our system has the potential to aid existing vision-based learning from demonstration (LfD) or imitation learning studies, which only considered the end-effector trajectory in their learning process. It would be possible for our system to also extract the end-point stiffness information from the demonstration videos of human, so that advanced LfD methods can be realized. We believe that our study would be a baseline for improved research on vision-based skill transfer to robots, which can also consider how humans adapt the stiffness when executing a task.

Ethics: In Bavaria, Germany, where the study was conducted, ethical review and approval from an ethical committee are waived due to the anonymized data collection process (https://ethikkommission.blaek.de/studien/sonstige-studien/antragsunterlagen-ek-primarberatend-15-bo) (Accessed on 20th June, 2023).

REFERENCES

[1] D. S. Walker, R. P. Wilson, and G. Niemeyer, "User-controlled variable impedance teleoperation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2010, pp. 5352–5357.

[2] L. Peternel, T. Petrić, and J. Babić, "Robotic assembly solution by human-in-the-loop teaching method based on real-time stiffness modulation," *Auton. Robots*, vol. 42, pp. 1–17, 2018.

[3] C. Yang, C. Zeng, P. Liang, Z. Li, R. Li, and C. Su, "Interface design of a physical human–robot interaction system for human impedance adaptive skill transfer," *IEEE Trans. Automat. Sci. Eng.*, vol. 15, no. 1, pp. 329–340, Jan. 2018.

[4] L. Peternel, L. Rozo, D. Caldwell, and A. Ajoudani, "A method for derivation of robot task-frame control authority from repeated sensory observations," *IEEE Robot. Automat. Lett.*, vol. 2, no. 2, pp. 719–726, Apr. 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[6] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[7] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1492–1500.

[8] C. Yang, G. Ganesh, S. Haddadin, S. Parusel, A. Albu-Schaeffer, and E. Burdet, "Human-like adaptation of force and impedance in stable and unstable interactions," *IEEE Trans. Robot.*, vol. 27, no. 5, pp. 918–930, Oct. 2011.

[9] G. Ganesh, N. Jarrassé, S. Haddadin, A. Albu-Schaeffer, and E. Burdet, "A versatile biomimetic controller for contact tooling and haptic exploration," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2012, pp. 3329–3334.

[10] Y. Michel, R. Rahal, C. Pacchierotti, P. R. Giordano, and D. Lee, "Bilateral teleoperation with adaptive impedance control for contact tasks," *IEEE Robot. Automat. Lett.*, vol. 6, no. 3, pp. 5429–5436, Jul. 2021.

[11] A. Ajoudani, N. Tsagarakis, and A. Bicchi, "Tele-impedance: Teleoperation with impedance regulation using a body–machine interface," *Int. J. Robot. Res.*, vol. 31, no. 13, pp. 1642–1656, 2012.

[12] C. Nissler, N. Mouriki, C. Castellini, V. Belagiannis, and N. Navab, "OMG: Introducing optical myography as a new human machine interface for hand amputees," in *Proc. IEEE Int. Conf. Rehabil. Robot.*, 2015, pp. 937–942.

[13] C. Nissler, N. Mouriki, and C. Castellini, "Optical myography: Detecting finger movements by looking at the forearm," *Front. Neurorobot.*, vol. 10, 2016. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnbot.2016.00003/full

[14] C. Nissler, I. M. Badshah, C. Castellini, W. Kehl, and N. Navab, "Improving optical myography via convolutional neural networks," in *Proc. Myoelectric Controls Symp.*, 2017, pp. 1–4.

[15] Y. T. Wu, E. Fujiwara, and C. K. Suzuki, "Evaluation of optical myography sensor as predictor of hand postures," *IEEE Sensors J.*, vol. 19, no. 13, pp. 5299–5306, Jul. 2019.

[16] M. E. Huber, C. Folinus, and N. Hogan, "Visual perception of limb stiffness," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 3049–3055.

[17] D. Osokin, "Real-time 2D multi-person pose estimation on CPU: Lightweight openpose," 2018, *arXiv:1811.12004*. [Online]. Available: https://github.com/Daniil-Osokin/lightweight-human-pose-estimation.pytorch

[18] A. Takagi, G. Xiong, H. Kambara, and Y. Koike, "Endpoint stiffness magnitude increases linearly with a stronger power grasp," *Sci. Rep.*, vol. 10, no. 1, pp. 1–9, 2020.

[19] B. Peacock, T. Westers, S. Walsh, and K. Nicholson, "Feedback and maximum voluntary contraction," *Ergonomics*, vol. 24, no. 3, pp. 223–228, 1981.

[20] F. Gustafsson, "Determining the initial states in forward-backward filtering," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 988–992, Apr. 1996.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proc. 19th Int. Conf. Learn. Representations*, 2015, pp. 1051–1060.

[23] Y. Li, N. Jarrassé, and E. Burdet, "Versatile interaction control and haptic identification in humans and robots," in *Geometric and Numer. Foundations of Movements*. Berlin, Germany: Springer, 2017, pp. 187–206.