IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 October 2022, accepted 21 November 2022, date of publication 24 November 2022,
date of current version 30 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3224584

APPLIED RESEARCH

Event Augmentation for Contact Force Measurements

FARIBORZ BAGHAEI NAEINI[1], SANKET KACHOLE[1], RAJKUMAR MUTHUSAMY[2], DIMITRIOS MAKRIS[1], AND YAHYA ZWEIRI[2], (Member, IEEE)

[1]Faculty of Science, Engineering and Computing, Kingston University, KT1 2EE London, U.K.
[2]Khalifa University Center for Autonomous Robotic Systems (KUCARS), Department of Aerospace Engineering, Khalifa University, Abu Dhabi, United Arab Emirates

Corresponding author: Fariborz Baghaei Naeini (F.Baghaeinaeini@kingston.ac.uk)

This work was supported in part by Kingston University London, and in part by the Khalifa University of Science and Technology under Award RC1-2018-KUCARS.

ABSTRACT Neuromorphic vision sensor is an attractive technology that offers high dynamic range, and low latency which are crucial in robotic applications. However, the lack of event-based data in this field, limits the sensors' performance in a real-world environments. In this paper, we propose a novel augmentation technique for neuromorphic vision sensors to improve contact force measurements from events. The proposed method shifts a proportion of events across the time domain, 'Temporal Event Shifting', to augment the dataset. A new set of grasping experiments is performed to validate and analyze the effectiveness of the proposed augmentation method for contact force measurements. The results indicate that temporal event shifting is highly effective augmentation method which improves the models' accuracy for the contact force estimation by thirty percent without performing new experiments.

INDEX TERMS Event-based augmentation, neuromorphic augmentation, vision-based tactile sensor.

I. INTRODUCTION

Vision-based tactile sensor is a category of optical sensors which aims to acquire tactile information by utilizing a camera [1], [2], [3], [4]. The camera is mounted on the robotic hand to capture images of the object's contact area, which are then processed to measure contact force, estimate force distribution, and predict object slippage. A wide range of sensors and robotic fingertips are designed to deal with various applications. Since the sensors have different physical properties, the data captured by sensors cannot be used for other sensors. Therefore, the datasets are often small and application specific in this field. On the other hand, the data collection process is a time-consuming and costly process. Therefore, alternative approaches such as simulation and synthetic data generation are studied. For example, simulation techniques have been adapted to increase the volume and diversity of datasets for training deep learning models [5], where the position and texture of the object were randomized. On the other hand, sim-to-real techniques aim to transfer learning from simulations and adapt the model to the real environment [6]. However, less attention is paid to augmentation techniques for vision-based tactile sensors.

Ordinary image sensors capture the light intensity values of each pixel at a given framerate, normally within the range of 25-120Hz. Cameras with higher framerates up to 12kHz are available but they are expensive and their dynamic range may be reduced. On the other side, neuromorphic vision sensors (event-based cameras) capture intensity changes with low latency and high dynamic range. The main advantage of neuromorphic vision sensors are a low latency (few microseconds), high dynamic range (120dB) and low power consumption (5-12mW) [7], [8], [9], [10]. The high sampling rate and dynamic range of neuromorphic cameras enable the sensor to achieve a higher sensitivity and time resolution in robotic applications. The low latency of the sensory system allows the robot to feedback control signals in real-time in order to prevent failures [11], [12], [13]. In addition, the low power consumption of the sensor may enable robotic systems to perform longer with the limitations of batteries.

Computer vision has long been a key enabler of industrial robots, where it is used to guide and control the

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Zhou.

positioning of the robot to achieve a high-level task. As supported in the literature [12], [13], [14], [15], [16], the use of conventional frame-based cameras for robotic applications introduces several limitations on the maximum speed and process robustness due to several shortcomings of frame-based cameras such as motion blur, low dynamic range, latency, exposure timing, poor perception at low-light conditions and high-power consumption. These shortcomings of frame-based cameras impose constraints on robot operational speeds, workspace volumes, and ambient lighting conditions; which affect the robustness and productivity of robotic manufacturing processes. However, the use of neuromorphic cameras introduces new challenges regarding the unavailability of enough event data for robotic applications. Hence, in this paper, a novel augmentation-based DL technique was introduced to develop predictive contact force measurement models for neuromorphic vision sensors using limited measured data. The presented results show that the developed DL models can be considered promising tools in learning measurement from limited experimental data to make high-fidelity performance predictions.

In our previous work [17], we proposed a novel vision-based tactile sensor using a neuromorphic vision sensor to estimate the contact force using deep learning techniques. A number of deep learning architectures and hyperparameters were studied whereas the deep learning model based on ConvLSTM layers achieved the highest accuracy for the contact force estimation. However, the experiments were conducted with the same object size and this approach cannot be generalized for objects with different sizes.

In this paper, we conduct a new set of experiments by considering three different object sizes. In addition, new augmentation methods are proposed to synthesize experiments for an unseen object size without performing real experiments. We demonstrate that the augmentation methods improve the neural networks' accuracy without performing further experiments. Our approach significantly reduces the cost and time for the data collection process by creating new synthetic datasets for unseen object sizes. In the proposed augmentation techniques, both 2D (image-based) and temporal (time-domain) are investigated and a novel technique is proposed that shifts events along the time dimension to generate further synthetic samples. For evaluation purposes, all the augmentation methods are validated on the best deep learning architecture (ConvLSTM) from [17].

The main contributions of this paper are:

- Developing time-domain and image-based augmentation for the neuromorphic tactile sensor for objects with different sizes.
- Proposing a novel event-based augmentation technique, ''Temporal Event Shifting'', to synthesize sequences and increase the model's accuracy.
- Performing new experiments with various object sizes to validate the effectiveness of augmentation methods considering ConvLSTM architecture proposed in [17].

A. RELATED WORK

Data augmentation techniques aim to generate synthetic data for training to improve model generalization. Augmentation methods can be divided into two main categories [18]: Model-based augmentation, and data manipulation. Model-based augmentation methods focus on training models to generate synthetic data from the real data such as Generative Adversarial Networks (GAN) introduced in [19].

Algorithmic data manipulation techniques apply fundamental operations to the data to generate realistic samples. For images, the geometric transformation of the training data such as rotation, translation, and shear has shown an improvement for classification tasks [20]. In [21], geometric translations and dropout layers are utilized to improve traffic sign recognition. The results indicate that the validation accuracy was improved by more than 5% considering rotation, translation, and shearing augmentation methods. In addition to spatial methods, other image-based augmentation techniques such as image distortion, morphological, and noise injection techniques have increased the networks' accuracy for image classification [22].

In the augmentation process, many variables are involved that can be tuned based on application and system characteristics. Some of the studies such as [23] proposed an automatic framework for data augmentation. The proposed approaches consider both feature-space and data-space augmentation methods to generate synthetic data. To validate the augmentation methods, the models are trained multiple times to account for random initialization of weights. From another point of view, the effectiveness of refining the labels for augmentation is investigated in [24]. The authors demonstrate that algorithmic augmentation methods including the cropping technique may result in inaccurate labels for specific classes. Therefore, rules and conditions must be applied in the augmentation process by considering samples of each class independently.

Time-series augmentation methods consider time and frequency domain features to generate synthetic data. One of the common approaches in the time-domain is shifting inputs in regard to the ground truth to introduce a random delay in the sequence. In [25], signals are shifted randomly to make the model robust against unseen signals. Moreover, the authors considered a combination of pitch shifting in the frequency domain and time warping to improve the accuracy of the model for classifying environmental sounds. Window slicing is another popular approach in time-series classification which considers a sequence of the original signal during both the training and testing process of the model [26].

GANs are a class of machine learning models that includes two networks jointly trained to synthesize data. The first network (known as generative) learns to generate samples from a latent feature space while the second network (discriminator) identifies the realism of the produced samples. Although GANs achieved impressive results in [19],

there is a lack of stability for training in practice [27]. Several studies have modified the GAN structure to improve the generated samples. For instance, a cascade CNN with pyramid (multi-scale) features is proposed in [28] which has produced high-quality realistic samples. In [29], a novel class of architecture, Deep Convolutional Generative Adversarial Networks (DCGAN), is presented to generate samples in an unsupervised manner. In addition to image generation, time-dependent GANs are designed to capture temporal features and produce time-series samples. In [30], recurrent neural networks are employed in both generator and discriminator to produce continuous time-series samples. Similarly, recurrent conditional GAN is proposed in [31] and [32] with conditions in the time dimension to generate multi-dimensional time-series samples. A comprehensive review of recent time-series GANs is provided in [33]. as

However, training a GAN model requires a considerable number of existing data to achieve acceptable results. The training process for GAN is time-consuming and the results are required to be confirmed by human. Furthermore, interpretation of learning representations using GANs and deep learning models are difficult compared to the algorithmic augmentation methods [18]. Due to the lack of event-based data for grasping applications, we propose algorithmic augmentation methods to enrich data for the contact force estimation models. Algorithmic augmentation methods target handcrafted features to produce synthetic data based on logic and observations tailored for a particular application.

Evaluation of the augmentation methods is often performed on a validation set by using the augmented data in the training process. Since deep neural networks can easily overfit the training data, performance on the validation set provides a more intuitive evaluation. For instance, algorithmic and GAN augmentation methods are used in [34] to evaluate the effectiveness of each method for a classification task on the validation set. Similarly, various augmentation methods are proposed in [35] to classify medical images. The networks' accuracy is evaluated on the validation set to analyze the effectiveness of augmentation techniques.

Even though image augmentation techniques have been studied widely in the literature, few studies have been conducted to investigate event-based augmentation for any applications. In [36], dropout event augmentation techniques are proposed to drop events randomly, based on time and area of events. Authors demonstrate that such a technique leads to improved network accuracy using various event representations and datasets. Another study [37] proposed a mix of geometric augmentation including rotation, flipping, rolling, cutout, mix-up, and shear methods to augment events. This method shows a significant improvement in network accuracy for SNN and ANN networks. The event augmentation techniques can be applied directly on event streams, event-frames and other common event representations reviewed in [7]. This investigates image-based and time-series augmentation methods applied to sequences of event-frames in tactile sensing applications.

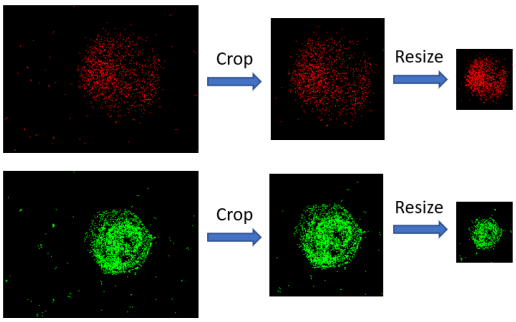


FIGURE 1. Frames are constructed by accumulation events considering two channels for positive and negative polarities. The left images show the constructed frames while the middle and right images illustrate the constructed frames after cropping and resizing respectively.

II. EVENT FRAME SEQUENCE AUGMENTATION

Events captured by neuromorphic vision sensors are characterized by location (x,y), timestamp (t), and polarity (p). Similar to the frame construction in [17], event frames are constructed by the accumulation of events over a time window while preserving spatial information. The accumulation of events is performed on positive and negative polarity events separately to construct two channels of the frame. This technique has been widely used to compress event data [38], apply image-based deep learning methods [39] and be compatible with standard hardware accelerations for images and sequence of frames.

The sensor has a dimension of 240×180 which covers the contact area and the background. To reduce the memory requirements of the system and the effect of the background noise, each frame is cropped to 140×150 pixels by considering the largest contact area size. Afterwards, the frames are downscaled to half (70×75) by adding the closest neighborhoods to a single pixel to reduce the frame size. Furthermore, the two channels are resized and then combined into one matrix to create the event frames. For visualization purposes, the image is populated with the created matrix considering red and green channels.

There is a trade-off between resizing the frame and maintaining spatial information of the events. In this application, pixel-wise information is not critical for the accuracy of the overall contact force estimation. Reduction of image size decreases the model inference and training time which is important for real-time applications. Figure 1 presents the cropping and resizing process over the two channels. After constructing of the frames, the augmentation methods are applied to generate further synthetic sequences for training the networks.

A. IMAGE-BASED AUGMENTATION

2D or image-based augmentation techniques aim to enrich the dataset to achieve a better generalization and eliminate biases in the dataset. For example, if experiments are captured within a specific range of object orientation, the rotation augmentation adds experiments with other object orientations

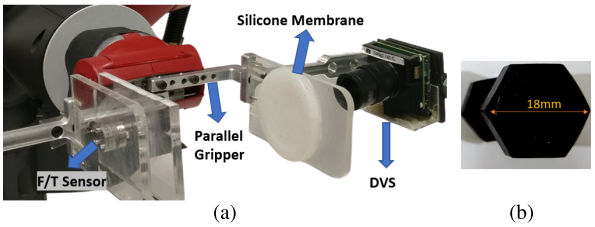


FIGURE 2. (a) A DVS is mounted on the left plane to observe the intensity changes in the contact area through the silicone membrane. A F/T sensor is located on the right plane to record force values through the grasp. (b) A bolt with an 18mm diameter painted in black.

to the dataset. Parallel grippers apply force on the object from both sides simultaneously, as shown in Figure 2. The object orientation remains the same through the grasp after the object stabilization. Assuming that objects have the same shape, two main features are varied between different objects: (i) Size; (ii) Contact area orientation. Both of the features can be augmented by affine transformations on the contact area images (event-frames).

1) ROTATION

The contact area orientation may vary across experiments. On the other hand, the object orientation remains the same in a stable grasp using a parallel gripper. Therefore, we considered the same rotation transformation for all frames of each sequence, instead of varying the transformation along the sequence. $X_t(x, y, p)$ represents the sequence of the original frames with spatial coordinates (x, y) with polarity p at timeframe t. For each experiment, the newly generated frames $X'_t(x', y', p)$ are formulated according to Equation 1

$$X'_t(x', y', p) = X_t(x, y, p) \tag{1}$$

while Equation 2 represents the rotation around the centre of the object (x_o, y_o) by an angle ϕ.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \times \begin{bmatrix} x - x_o \\ y - y_o \end{bmatrix} \tag{2}$$

2) RESIZE

The aim of this paper is to augment data for a grasped object with a different size than the ones used in the captured data. For example, training data includes experiments for a small and a large object while the sensor must estimate the contact force for any intermediate object size. In order to augment the images to the desired size (e.g medium size), the original images are required to be resized considering a specific scaling ratio β. The scaling ratio is determined based on the real object sizes where $\beta > 1$ and $\beta < 1$ for resizing to larger and smaller sizes respectively. We choose linear interpolation to assign values to the pixels. To preserve the same image resolution for all samples, a margin with zeros is added to maintain the image size. As x and y dimensions are scaled with the same ratio, the resized samples preserve the aspect ratio of the object contact area.

B. NOISE

To establish a noise model, a set of experiments are recorded without any movements in the scene. Afterwards, the triggered events are considered noise which is accumulated over a time window over two different channels. Finally, the noise frames are added to the frames in the original dataset to generate samples with artificially added noise.

C. TIME-SERIES AUGMENTATION

In the grasping process, a lot of parameters such as Dynamic Vision Sensor (DVS) threshold, silicone material, sensor hysteresis, and uncertainty cause a varying delay between the applied force and the triggered events. Time-series augmentation methods aim to generate synthetic sequences by considering transformations along the time dimension.

1) FRAME SHIFTING

One of the simplest augmentation techniques in the time domain is to shift the index of the frames by a certain value (j) while preserving the ground truth. This approach assists the network to deal with a slight lag between different sequences. Since shifting frames remove j frames from the input, new frames are required to be added to keep the sequence length fixed and are all set to zero values. Equation 3 presents the frameshifting process where the new frames are denoted as X'_t and j presents the shifting value. The frameshifting is applicable in both directions (i) Left: The frames are shifted to the earlier timestamps ($j < 0$); (ii) Right: The frames are shifted to the future timestamps ($j > 0$).

$$\forall t, \quad X'_t(x, y, p) = X_{t+j}(x, y, p) \tag{3}$$

2) TEMPORAL EVENT SHIFTING

Similar to frameshifting, we propose a novel approach to shift events across the frames, called "Temporal Event Shift (TES)". In fact, Frame Shifting is a specific case of temporal event shifting where all the events are moved to the previous or next frames. The proposed method selects a fraction ζ of events ($0 < \zeta < 1$) randomly in each frame. These events are removed from the current frame and added to the next or previous j frames. Figure 3 demonstrates the procedure for temporal event shifting to the right while preserving the spatial information of events.

To shift the events to the past frames, j value is considered negative. This process is formulated in Equation 3 where the new frame is denoted as X'_t. $\forall t, p$, create a difference frame $Z_t(x, y, p)$ such as:

$$Z_t(x, y, p) \leq X_t(x, y, p), \quad \forall x, y \tag{4}$$

$$\sum_{x,y} Z_t(x, y, p) = \zeta \cdot \sum_{x,y} X_t(x, y, p) \tag{5}$$

$$X'_t(x, y, p) = X_t(x, y, p) - Z_t(x, y, p) + Z_{t+j}(x, y, p) \tag{6}$$

Based on the formulation, frame shifting is a special case of temporal event shifting where $Z_t = 1$.

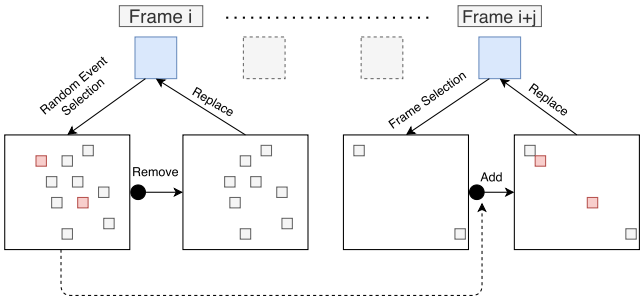


FIGURE 3. Temporal event shifting diagram when a ratio of events is shifted to the future frames ($j > 0$).

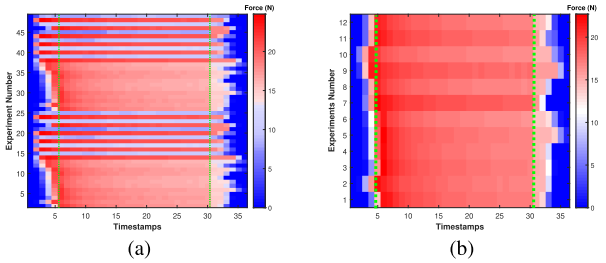


FIGURE 4. Each row demonstrates the force values that is captured by the F/T sensor over time. (a) Training set: 48 experiments are conducted using the small and large objects. (b) Validation set: 12 experiments considering the medium size object.

III. EXPERIMENTAL SETUP

The experimental environment is not fully controlled to mimic real-world grasping applications and show the sensor performance under uncertainty. Real-life experiments are conducted on a Baxter robot including a F/T sensor, silicone membrane, DVS, and 3D-printed transparent planes. The transparent silicone membrane has 50 shore hardness and 8mm depth. Furthermore, the range of contact force is set to 0-25N which is significantly higher than the force range in [11], [17]. Figure 2(a) presents the experimental setup for the grasping task.

Three bolts with 12, 15, and 18mm diameters are used for the grasping process as shown in Figure 2(b). In this paper, all objects are painted in black to increase the contrast between the environment and the object's surface. Alternatively, a black silicone membrane with fixed lighting conditions can be considered like [40]. The fixed lighting conditions and DVS thresholds lead to standardizing the threshold of the event for all experiments in various environments.

This study aims to reduce the cost and time of the data collection process by investigating the impact of augmentation methods. Therefore, we assume that experiments for two object sizes are given while the network learns to estimate the contact force for unseen object sizes (e.g medium). The main reason for choosing medium size objects for validation and testing is to ensure the right interpolation between the smallest and the largest distribution. In practice, the collection of data for two sizes (smallest and largest) is applicable and other sizes can be augmented with the proposed method. Therefore, we choose the small and large bolts for training (48 sequences) while the medium bolt experiments (6 sequences for validation and 6 sequences for the test set) are considered for validation. Furthermore, the augmented data for the desired size (medium bolt) are added to the training set to evaluate the network performance and compare augmentation techniques. Figure 4 presents the force values recorded by the F/T sensor for the training (a) and validation sets (b).

Two configurations are set for the gripper to grasp the object with a different applied force. The experimental setup is not fully controlled which results in a slight variation of force between experiments with the same configuration. Therefore, a slight variation of force over time is visible.

A. PREPARATION OF FRAMES

The experiments have a maximum length of 360ms. In this paper, 36 frames are conducted for each sequence by the accumulation of events over a 10ms window. The frames are cropped to 140×150 to reduce the noise and eliminate the background which is selected based on the largest object contact area. Afterwards, the frames are resized to 70×75 pixels considering the accumulation of neighborhoods to reduce memory requirements. The resizing ratio is selected based on the maximum saturation level of each pixel over the time window. The force readings have a resolution of 2ms which is measured by the F/T sensor. After the synchronization, force measurements are read every 10ms to synchronize them with the frames.

B. TRAINING CONFIGURATIONS

We have studied various architectures including LSTM, CNN-LSTM and Convolutional LSTM architectures in [17] comprehensively. In this paper, we validate the effectiveness of the augmentation techniques on the best-performing architecture (ConvLSTM) only to have a fair comparison between the augmentation methods without changing the network architecture.

To select the hyperparameters, firstly we performed experiments on the original dataset to find the best optimizer, early stopping value and learning rate and ensure network convergence. Secondly, we train the model on each augmented dataset 10 times with a different random seed while keeping the same hyperparameters and network architecture to remove any influence of randomness. Finally, we evaluate the results by considering the average error accross the 10 trained models. Figure 5 presents examples of training and validation loss for different augmentation methods.

To ensure all the networks reach the stabilization point of training and validation loss, we set the early stopping parameter to 20 based on trial and error. Therefore, the training process finishes when the validation loss stops improving after 20 consecutive epochs. Adam optimizer is used to minimize the training loss (MSE) for the training set

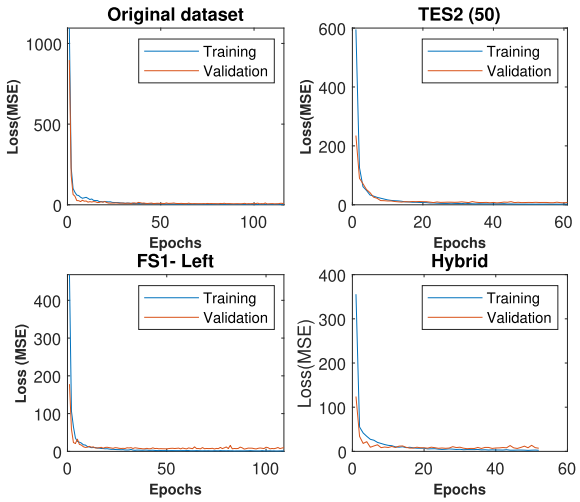


FIGURE 5. Examples of training and validation loss for the original dataset, TES2-50, FS-1 and hybrid augmentation methods.

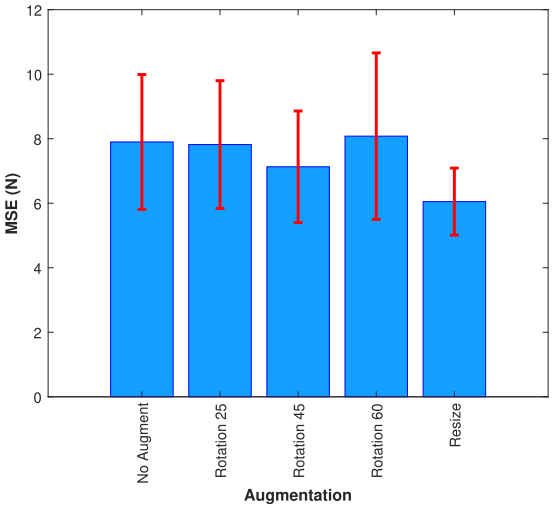


FIGURE 6. MSE for geometric augmentation methods. *y*-axis shows the average of MSE(N) for the trained networks after 10 repetitions with random initialization. The red lines represent the standard deviation (STD) of MSE(N) for each method.

while monitoring the validation loss for selecting the best network. The training process finishes when the validation loss stops improving after 20 consecutive epochs. All the models are trained with the same configuration to provide a fair comparison. Keras framework is used to set the training configuration using an NVIDIA 1080 GPU. Figure 4 presents the training and validation set of the data set.

IV. RESULTS AND DISCUSSION

To evaluate the augmentation techniques, the training data size is doubled with the synthetic sequences while preserving the ground truth. Since the random initialization of weights affects the training process, the random seed is controlled for 10 runs. The final results are obtained by averaging the lowest error on the validation set using the same random initialization. Figure 6 presents the average of MSE for the validation set where the red line shows the standard deviation of MSE for the image-based augmentation methods.

The network trained without augmentation (No Augment) achieves MSE of 7.89N with STD of 2.09N. The standard deviation of more than 25% indicates instability of the training process with respect to random initialization. The rotation of images between 0 and 45 degrees (Rot45) provides a slight improvement in network accuracy. The best result using the geometric augmentation approaches is achieved by the resizing method for the desired object size. The scaling factor of resizing is considered as 1.25 and 0.83 for small and large objects respectively.

On the other hand, we consider background noise for further augmentation. The background noise includes both event polarities which are added to the original frames to double the training samples. The results indicate a slight improvement of 10% in MSE and the standard deviation is comparable to the networks that are trained without augmentation.

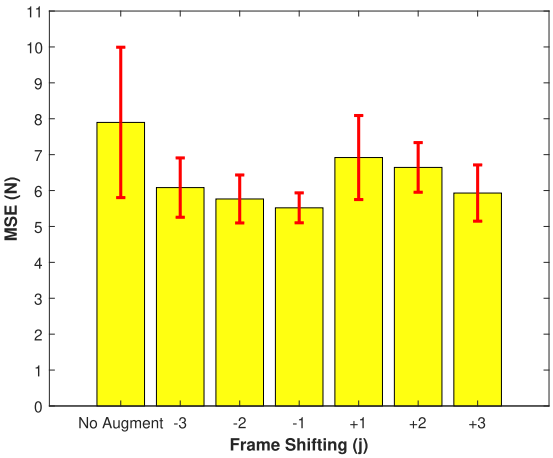


FIGURE 7. Comparison of average MSE for frame shifting methods. *x*-axis shows the *j* value for frame shifting and the red bar illustrates the standard deviation (STD) of MSE over 10 runs.

The results indicate that the MSE of the network is reduced to 6.05N and the standard deviation is decreased to 1.04N, a decrease of 50%. Therefore, resizing is the most effective image-based augmentation method, which makes sense as the challenge in our experiments was to train the networks for unseen object size.

Two time-series augmentation methods, mentioned in the section, are tested: Frame Shifting (FS) and the proposed Temporal Event Shifting (TES). In most of the experiments, the majority of events are fired within three frames (30ms) for the grasping and releasing phase. Therefore, our time-series augmentation considers a maximum shifting of 3 frames. For the FS method, *j* is varied between -3 and 3 to find the most effective value to shift the frames. Figure 7 presents the effectiveness of frameshifting augmentation with different *j* values.

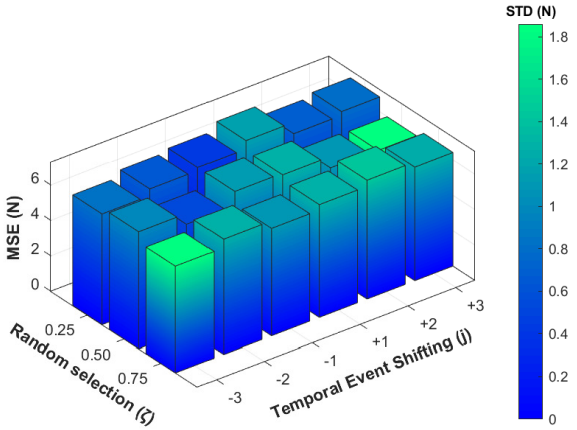


FIGURE 8. Comparison of average MSE of the networks for Temporal Event Shifting augmentation technique. x-axis and y-axis present the j and ζ values respectively. The MSE value of each method is illustrated on z-axis. The color of each bar surface represents the standard deviation (STD) of MSE over 10 runs.

Shifting one frame to left (FS-1) results in the lowest MSE of 5.51N which is 30% less than the MSE achieved without augmentation. Furthermore, the STD of errors is reduced significantly to one fourth (0.41N) of the networks trained on the real data.

For the TES method, the fraction of the events to be moved (ζ) is considered as 0.25, 0.50 and 0.75 with the same j variations as in the FS method. Figure 8 demonstrates the average MSE of the validation set considering different j and ζ. Among the TES augmentation configurations, two frames shift to the left with 50% threshold (TES-2(0.50)) resulting in the minimum MSE of 5.98N with 30% reduction of standard deviation (0.53N) compared to the results without augmentation.

In FS-based augmentations, the amount of new data generated is limited to one new sequence for the original sequence considering a fixed j value. On the other hand, in TES-based augmentations, the random seeds affect the selection of events, and as a consequence, an unlimited number of new samples can be produced for specific values of j and ζ. We produced an experiment to generate 480 synthetic sequences by varying the seed for TES-2(50) method. The results show that increasing the generated samples does not improve the network performance where an average MSE of 6.25N with 0.82N standard deviation is achieved. The main reason for this phenomenon is that the ground truth remains the same, despite the significant variation in the input.

Most of the augmentation techniques in the time domain improve the networks' performance. The main factors that affect the events along the time dimension are the F/T sensor hysteresis, the non-linear behavior of the silicone membrane, uncertainty, and vibrations. These factors are inevitable in real-world applications which show the benefit of the augmentation methods in the time domain.

A typical grasping task includes three phases: (i) Grasping phase is defined where the contact force increases to the maximum level (The first 5 frames); (ii) Holding phase

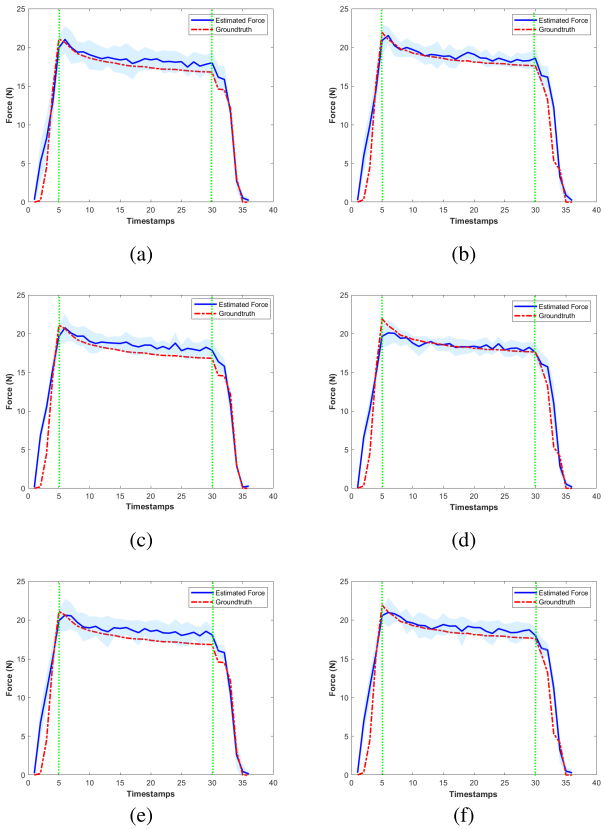


FIGURE 9. The highlighted area illustrates the standard deviation of the estimated force over 10 runs. The phases of a grasp are differentiated by the green line in each figure. Each column presents an experiment from the validation set.Top row presents the average of estimated force and groundtruth without augmentation. The middle row demonstrates the output of the network for FS-1 augmentation method. The bottom row (e,f) presents the average of estimated force and groundtruth for TES-2(0.50) augmentation method.

includes a slight variation of force during the time from 6th frame to 30th frame. (iii) Releasing phase where the force values are decreased continuously to zero (The last 5 frames). Figure 9 presents the average of estimated force (blue) and ground truth (red) for two examples of the validation set over 10 runs. The top row (a,b) demonstrates the average of the force predictions for training without augmentation while the middle row (c,d) presents the average of estimated force considering FS-1 method. The bottom row (e,f) demonstrates the average of estimated force and ground truth using TES-2(0.50) method.

The results indicate that both frameshifting and temporal event-shifting augmentation reduce the standard deviation of the predictions in all three phases. In fact, the impact of random initialization is decreased by augmenting the training data. In Figure 9 (b) and (d), a clear improvement of the estimated force in the majority of the vibration phase is visible. Even though the frameshifting results in a lower MSE and standard deviation, the temporal event shifting method captures the maximum contact force (at $5th$ timestamp) more accurately in most of the cases.

In order to investigate the impact of augmentation methods on all the measurements, 12 predictions of 10 models are

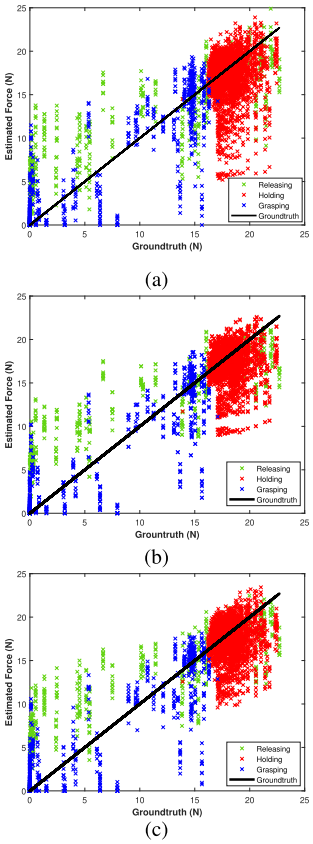


FIGURE 10. Estimated force using models a) trained without augmentation, b) trained with FS-1, c) trained with TES-2(50).

considered for grasping, holding, and releasing phases. The final results include 4320 points which are demonstrated in Figure 10. The black line presents the contact force measured by F/T sensor. The estimated force is presented by blue, red, and green for the grasping, holding, and releasing phases respectively. Figure 10 compares the estimated force using FS-1 and TES-2(50) augmentation techniques.

As observed in Figure 10, FS-1 and TES-2(50) augmentations improve the force estimation in the holding phase. Both augmentation methods shift events to the earlier frames to create synthetic sequences. The main reason for this phenomenon is that the number of triggered events increases significantly after applying a certain amount of force. Therefore, shifting the events to the left allow the network to relate more events to the contact force in the early frames. Furthermore, the silicone membrane has a non-linear deformation that absorbs a ratio of the contact force, particularly in the transition phases. The force absorption coupled with the F/T sensor hysteresis introduces a variable delay between the triggered events and the contact force.

The image-based and time-domain augmentation methods synthesize the training data from different perspectives. Therefore, a combination of both methods provides both spatial and time-domain variations in the generated samples. Since the best accuracy is achieved by resizing and FS-1,

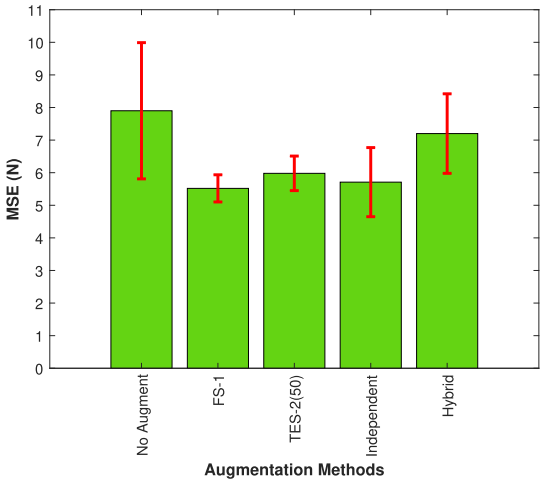


FIGURE 11. Comparison of average MSE for the proposed augmentation method. The independent and hybrid methods are considered for FS-1 and resizing methods that achieved the lowest error for time-domain and image-based techniques respectively. The STD of each method is presented as a red line.

these two methods are combined to generate a new set of synthetic samples. There are two ways to combine the two methods: (i) Perform each augmentation method independently to generate synthetic sequences; (ii) Hybridise both augmentations methods on samples to generate a set of synthetic samples. The results indicate that independent augmentation of each sample achieves better accuracy than a simultaneous combination of methods. The independent sample generation method reduces the average MSE of the networks to 5.71N with a standard deviation of 1.06N which is slightly higher than FS-1 method. The hybrid augmentation method results in a high MSE of 7.20N with a standard deviation of 1.22N, a significantly higher error compared to FS-1 method. Figure 11 demonstrates the average MSE of the proposed augmentation methods where the standard deviation is highlighted as a red line.

In image-based augmentation techniques, resizing the object to a desired size results in the best accuracy. Since the network learns the relationship between the applied force and triggered events based on the contact area, resizing the training data simulates the experiments for the new size of an object.

A noticeable delay was observed in the releasing phase where the network always responds faster than the F/T sensor. In fact, the responding time of the silicone membrane has a significant impact on the delay between the triggered events and the contact force. For example in [11], we demonstrated that same shape objects with different elasticity generate a different number of events which can be used to classify the objects' material. Therefore, the augmentation methods in the time domain improve the network accuracy remarkably whereas FS-1 results in the lowest average of MSE.

V. CONCLUSION

This paper proposed a novel event-based method to generate synthetic data for vision-based force estimation considering spatial and temporal domains. The experiments are

performed on three objects' sizes where the smallest and the largest objects are considered for training and the middle size object is used for testing. A novel augmentation technique event-shifting is proposed to generalize the network on unseen experiments. We demonstrated that algorithmic augmentation methods improve the network accuracy significantly without performing new experiments.

REFERENCES

[1] K. Shimonomura, "Tactile image sensors employing camera: A review," *Sensors*, vol. 19, no. 18, p. 3933, 2019.

[2] B. Ward-Cherrier, N. Pestell, and N. F. Lepora, "NeuroTac: A neuromorphic optical tactile sensor applied to texture recognition," 2020, *arXiv:2003.00467*.

[3] C. Sferrazza, A. Wahlsten, C. Trueeb, and R. D'Andrea, "Ground truth force distribution for learning-based tactile sensing: A finite element approach," *IEEE Access*, vol. 7, pp. 173438–173449, 2019.

[4] N. F. Lepora, "Soft biomimetic optical tactile sensing with the TacTip: A review," *IEEE Sensors J.*, vol. 21, no. 19, pp. 21131–21143, Oct. 2021.

[5] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2017, pp. 23–30, doi: 10.1109/IROS.2017.8202133.

[6] C. Sferrazza and R. D'Andrea, "Sim-to-real for high-resolution optical tactile sensing: From images to three-dimensional contact force distributions," *Soft Robot.*, vol. 9, no. 5, pp. 926–937, Oct. 2021.

[7] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2022.

[8] T. Taunyazov, W. Sng, H. H. See, B. Lim, J. Kuan, A. F. Ansari, B. C. K. Tee, and H. Soh, "Event-driven visual-tactile sensing and learning for robots," 2020, *arXiv:2009.07083*.

[9] A. Rigi, F. B. Naeini, D. Makris, and Y. Zweiri, "A novel event-based incipient slip detection using dynamic active-pixel vision sensor (DAVIS)," *Sensors*, vol. 18, no. 2, p. 333, Jan. 2018.

[10] O. Faris, R. Muthusamy, F. Renda, I. Hussain, D. Gan, L. Seneviratne, and Y. Zweiri, "Proprioception and exteroception of a soft robotic finger using neuromorphic vision-based sensing," *Soft Robot.*, Oct. 2022.

[11] F. B. Naeini, A. M. AlAli, R. Al-Husari, A. Rigi, M. K. Al-Sharman, D. Makris, and Y. Zweiri, "A novel dynamic-vision-based approach for tactile sensing applications," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 1881–1893, May 2020.

[12] R. Muthusamy, A. Ayyad, M. Halwani, D. Swart, D. Gan, L. Seneviratne, and Y. Zweiri, "Neuromorphic eye-in-hand visual servoing," *IEEE Access*, vol. 9, pp. 55853–55870, 2021.

[13] X. Huang, M. Halwani, R. Muthusamy, A. Ayyad, D. Swart, L. Seneviratne, D. Gan, and Y. Zweiri, "Real-time grasping strategies using event camera," *J. Intell. Manuf.*, vol. 33, no. 2, pp. 593–615, Feb. 2022.

[14] A. Ayyad, M. Halwani, D. Swart, R. Muthusamy, F. Almaskari, and Y. Zweiri, "Neuromorphic vision based control for the precise positioning of robotic drilling systems," *Robot. Comput.-Integr. Manuf.*, vol. 79, Feb. 2023, Art. no. 102419.

[15] P. I. Corke and S. A. Hutchinson, "Real-time vision, tracking and control," in *Proc. Millennium Conf. IEEE Int. Conf. Robot. Automat. Symposia (ICRA)*, Apr. 2000, pp. 622–629.

[16] X. Wang, G. Fang, K. Wang, X. Xie, K.-H. Lee, J. D. L. Ho, W. L. Tang, J. Lam, and K.-W. Kwok, "Eye-in-hand visual servoing enhanced with sparse strain measurement for soft continuum robots," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 2161–2168, Apr. 2020.

[17] F. B. Naeini, D. Makris, D. Gan, and Y. Zweiri, "Dynamic-vision-based force measurements using convolutional recurrent neural networks," *Sensors*, vol. 20, no. 16, pp. 1–15, 2020.

[18] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[19] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 3, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2014, pp. 2672–2680.

[20] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *Proc. Int. Interdiscipl. Ph.D. Workshop (IIPhDW)*, May 2018, pp. 117–122.

[21] D. Yasmina, R. Karima, and A. Ouahiba, "Traffic signs recognition with deep learning," in *Proc. Int. Conf. Appl. Smart Syst. (ICASS)*, Nov. 2018, pp. 1–5.

[22] I. Sato, H. Nishimura, and K. Yokoi, "APAC: Augmented pattern classification with neural networks," 2015, *arXiv:1505.03229*.

[23] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "AutoAugment: Learning augmentation strategies from data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 113–123.

[24] H. Bagherinezhad, M. Horton, M. Rastegari, and A. Farhadi, "Label refinery: Improving ImageNet classification through label progression," 2018, *arXiv:1805.02641*.

[25] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2015, pp. 1–6.

[26] A. L. Guennec, S. Malinowski, and R. Tavenard, "Data augmentation for time series classification using convolutional neural networks," in *Proc. ECML/PKDD Workshop Adv. Analytics Learn. Temporal Data*, 2016, pp. 1–9.

[27] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Computer Vision—ECCV*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 597–613. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46454-1_36

[28] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, 2015, pp. 1486–1494.

[29] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–16.

[30] O. Mogren, "C-RNN-GAN: Continuous recurrent neural networks with adversarial training," 2016, *arXiv:1611.09904*.

[31] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional GANs," 2017, *arXiv:1706.02633*.

[32] G. Ramponi, P. Protopapas, M. Brambilla, and R. Janssen, "T-CGAN: Conditional generative adversarial network for data augmentation in noisy time series with irregular sampling," 2018, *arXiv:1811.08295*.

[33] J. Yoon, D. Jarrett, and M. van der Schaar, "Time-series generative adversarial networks," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'textquotesingle Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 5508–5518.

[34] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*.

[35] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, "Differential data augmentation techniques for medical imaging classification tasks," in *Proc. AMIA Annu. Symp.*, 2017, pp. 979–984.

[36] F. Gu, W. Sng, X. Hu, and F. Yu, "EventDrop: Data augmentation for event-based learning," in *Proc. 30th Int. Joint Conf. Artif. Intell. (IJCAI)*, Z.-H. Zhou, Ed, Aug. 2021, pp. 700–707. [Online]. Available: https://www.ijcai.org/proceedings/2021/0097.pdf, doi: 10.24963/ijcai.2021/97.

[37] Y. Li, Y. Kim, H. Park, T. Geller, and P. Panda, "Neuromorphic data augmentation for training spiking neural networks," Yale Univ., New Haven, CT, USA, Tech. Rep., 2022. [Online]. Available: https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136670623.pdf

[38] N. Khan, K. Iqbal, and M. G. Martini, "Lossless compression of data from static and mobile dynamic vision sensors-performance and trade-offs," *IEEE Access*, vol. 8, pp. 103149–103163, 2020.

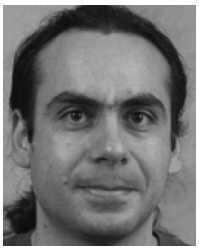
[39] A. I. Maqueda, A. Loquercio, G. Gallego, N. Garcia, and D. Scaramuzza, "Event-based vision meets deep learning on steering prediction for self-driving cars," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5419–5427.

[40] V. Kakani, X. Cui, M. Ma, and H. Kim, "Vision-based tactile sensor mechanism for the estimation of contact position and force distribution using deep learning," *Sensors*, vol. 21, no. 5, p. 1920, Mar. 2021.

FARIBORZ BAGHAEI NAEINI received the B.Sc. degree in computer hardware engineering from Azad University, Central Tehran Branch, Tehran, Iran, in 2015, the M.Sc. degree (Hons.) in embedded systems from Kingston University London, London, U.K., in 2017, where he received the Ph.D. degree in artificial intelligence from the School of Computer Science and Mathematics, in 2021. He has been active in many research groups, such as computer vision, deep learning, game theory, and smart energy. His current research interests include haptic, vision-based measurement, deep learning, neuromorphic vision, and robotics.

SANKET KACHOLE received the B.Sc. degree in mechanical engineering from Pune University and the M.Sc. degree in advanced industrial manufacturing from Kingston University, where he is currently pursuing the Ph.D. degree in artificial intelligence with the School of Mathematics and Computer Science. His research interests include deep learning, image segmentation, neuromorphic vision, and haptic.

RAJKUMAR MUTHUSAMY received the B.E. degree in electrical and electronics engineering from Anna University, India, in 2009, the M.S. degree (Hons.) in electrical engineering from Yuan Ze University, Taiwan, in 2013, and the D.Sc.(Tech) degree in automation systems and control engineering from Aalto University, Finland, in 2018. He is a Senior Robotics Scientist with Dubai Future Labs, currently focusing on the development of autonomous robotic systems and solutions for mobile manipulation and industrial applications. He contributed to the Development of Self-Driving Vehicles, as a Senior Robotics Engineer at Sensible4 Oy, Finland, in 2018. Before joining Dubai Future Foundation in August 2021, he was a Post-Doctoral fellow at Khalifa University, UAE and carried out RnD on advanced robotic perception and manipulation for Industry 4.0. His research interests include AI, autonomous vehicles and systems, assistive, collaborative, and general robotics. He is one of the recipients of the Finnish Engineering Award 2020. He is an enthusiast in providing endtoend solutions for the next generation of autonomous robots operating instructured and unstructured environments.



DIMITRIOS MAKRIS received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1999, and the Ph.D. degree in computer vision from City University, London, U.K., in 2004. He is currently a Professor with the School of Computer Science and Mathematics, Kingston University, London, U.K. His current research interests include image processing, computer vision and machine learning, and particularly motion analysis. He is a member of the British Machine Vision Association and the IET Vision and Imaging Network.

YAHYA ZWEIRI (Member, IEEE) received the Ph.D. degree from King's College London, in 2003. He is currently an Associate Professor with the Department of Aerospace Engineering and the Deputy Director of the Advanced Research and Innovation Center, Khalifa University, United Arab Emirates. He was involved in defense and security research projects in the last 20 years at the Defense Science and Technology Laboratory, King's College London, and the King Abdullah II Design and Development Bureau, Jordan. He has published over 130 refereed journals and conference papers and filed ten patents in USA and U.K. His main research interests include robotic systems for extreme conditions with particular emphasis on applied AI aspects and neuromorphic vision systems.