Privacy-Preserving Motor Intent Classification via Feature Disentanglement

Jiahao Fan, Xiaogang Hu*

***Abstract*—Recent studies have revealed that sensitive and private attributes could be decoded from surface electromyogram (sEMG) signals, which can incur privacy threat to the users of sEMG-based pattern recognition applications. Most studies so far focus on improving the accuracy and reliability of sEMG classifiers, but much less attention has been paid to their privacy. To fill this gap, this study implemented and evaluated a framework to optimize sEMG-based data-sharing mechanism. Our primary goal was to remove sensitive attributes (i.e., identity-relevant information) in the sEMG features before sharing them with primary pattern recognition tasks. We disentangled the identity-insensitive, task-relevant representations from original sEMG features. We shared it with the downstream pattern recognition tasks to reduce the chance of sensitive attributes being inferred by potential attackers. The proposed method was evaluated on data from twenty subjects, with training and testing data acquired 3-25 days apart. Our results showed that the disentangled representations significantly reduced the success rate of identity inference attacks compared to the original feature and its sparse representations generated by the state-of-the-art feature projection methods. The disentangled representation was then evaluated in hand gesture recognition tasks. Our results revealed that the disentangled representations led to higher classification accuracy across classifiers compared with other feature implementations. This work shows that disentangled representations of sEMG signals are a promising solution for privacy-preserving motor intent recognition applications.**

***Index Terms*—surface electromyography, privacy preserving applications, hand gesture recognition, feature disentanglement.**

I. Introduction

Rapid development and significant progress have been made in pattern recognition using surface electromyography (sEMG). The acquired sEMG data are generally processed and analyzed to decode the intended motions of the users. However, recent studies have revealed that sensitive and private attributes, such as user identities [1] and health status [2] can also be decoded from raw sEMG signals, which raises security and privacy concerns for the users. Most studies so far focus on optimizing the accuracy and robustness of sEMG feature extraction or classifier models, but much less attention has been paid to privacy issues.

It is non-trivial to exploit feasible strategies to defend against potential privacy attacks in advance of the wide deployment of sEMG pattern recognition applications. To our knowledge, no studies consider privacy concerns in developing sEMG-based pattern recognition applications. This preliminary study focuses on privacy-preserving sEMG applications by implementing a data-sharing mechanism that uses the disentangled sEMG representations for downstream pattern recognition tasks. Specifically, we considered two coupling components in sEMG: one reflects the muscular motor patterns (termed task-relevant component), and another is related to the individual characteristics (termed identity-relevant component). We trained an autoencoder-like architecture with specialized loss functions on sEMG features to separate these two representations. As a result, only the decomposed task-relevant components are preserved and utilized in the downstream pattern recognition tasks. By doing so, the sensitive identity-relevant information in the raw signals is largely filtered out while maintaining the data utility. We have validated the proposed method on a high-density sEMG dataset of twenty subjects. The experimental results show that using the disentangled task-relevant representations could significantly lower the success rate of identity inference attacks. Meanwhile, the utility of the disentangled representations is also demonstrated by the high classification accuracy on gesture recognition tasks under a rigorous cross-day validation.

II. MATERIALS AND METHOD

A. Dataset

We used the Hyser dataset [3] to validate the proposed method. The Hyser dataset consists of HD-sEMG signal from forearm muscles (256 channels, $4\times$ 64 electrode arrays) acquired from twenty subjects (12 and 8 female, aged between 21-34 years) on two separate days with an interval between 3-25 days (8.50 $\pm$ 6.72 days on average). We refer to the experimental session on the first and second day as session 1 and session 2 below, respectively. The sEMG signals were acquired with a total gain of 150, a 16-bit resolution, and a sampling frequency of 2048 Hz.

The sEMG signals were acquired when subjects performed 34 hand gestures sequentially for two trials. For each trial, three dynamic tasks (from relaxing to the target gestures) with a 1-s duration were performed for each gesture. Overall, two trials of data from each subject were collected in one session.

This study was supported in part by the National Science Foundation (CBET-1847319, IIS-2106747) and the Department of Defense (W81XWH2110185).

Jiahao Fan (email: jxf5683@psu.edu) and Xiaogang Hu (email: xxh120@psu.edu) are with the Department of Mechanical Engineering at Penn State University. Xiaogang Hu is also with the Departments of Kinesiology and Physical Medicine Rehabilitation, the Huck Institutes of the Life Sciences, and Centerc for Neural Engineering at Penn State University.

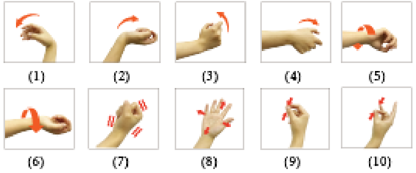

Fig. 1: Hand gestures selected in this study.

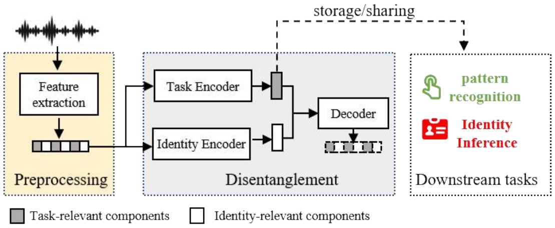

Fig. 2: The proposed data sharing mechanism. An autoencoder-like architecture first disentangles original features. Only the decomposed task-relevant representations are stored and shared with the downstream tasks. Typically, the legitimate downstream task is pattern recognition (marked in green), while the attempts to identify users are considered privacy attacks (marked in red).

In this study, we empirically selected ten hand gestures (see in Fig. 1) that were routinely used in myoelectric control [4].

B. Feature Extraction

The signals were split into 1-s segments for each gesture task based on the recorded markers. We then extracted four frequently used time-domain features from the sEMG segments. The extracted features are: 1) Root Mean Square (RMS), 2) Wave Length (WL), 3) Zero Crossing (ZC), and 4) Slope Sign Change (SSC). The values of features were calculated independently on each of the 256 channels, yielding a 1024-dimension (4 × 256) feature vector for each gesture task. Following [5], outlier detection was applied for each feature along the spatial axis to reduce the negative effect of channel occlusion.

C. Feature Disentanglement

Let $\{x_i^j\} \in \mathbb{R}^d$ denotes the sEMG features, where $\boldsymbol{x}$ denotes the extracted sEMG features with dimension $d = 1024$, and $i \in \{1, N_s\}$ and $j \in \{1, N_p\}$ are the subject identity and gesture patterns of $\boldsymbol{x}$, respectively. $N_s = 20$ and $N_p = 10$ denote the total number of subjects and patterns available in the dataset, respectively. As illustrated in Fig. 2, our implementation for feature disentanglement is based on an autoencoder-like architecture. We consider that sample x_i^j is composed of task-relevant components e^j, and identity-relevant component e_i. Our objective was to disentangle the task-relevant component from the original features, thus filtering out the identity sensitive information.

Specifically, our implementation consists of two encoders, E_P and E_S, and a decoder D. Thereinto, E_P is used to map the sample x^j to its task-relevant latent representation e^j, whereas E_S is applied to learn the map between x_i and its identity-relevant representation e_i. Sequentially, the decoder D reconstructs the sample from e_i and e^j. The process can be formulated as follow:

$$\begin{aligned} e^j &= E_P(x^j) \\ e_i &= E_S(x_i) \\ x_i^j &\approx D(e^j, e_i) \end{aligned} \tag{1}$$

To ensure the identity mapping of the autoencoder, the model is optimized to reconstruct the input x_i^j from the extracted latent representations:

$$\mathcal{L}_{recon} = \mathbb{E}[\|D(E_P(x_i^j), E_S(x_i^j)) - x_j^i\|] \tag{2}$$

We used the triplet loss to enable the separation of e_i and e^j to encourage the decomposed components to have much smaller intra-pattern/identity distances than inter-pattern/identity distance on the latent space. For E_P:

$$\begin{aligned} \mathcal{L}_{trip\_p} = \mathbb{E}[\|E_P(x_i^j) - E_P(x_l^j)\| - \|E_P(x_i^j) \\ - E_P(x_m^k)\| + \alpha]_+ \end{aligned} \tag{3}$$

where $k \neq j, l \neq i$, α is the margin. x_l^j denotes samples that share the same gesture pattern with x_i^j but have different identity, whereas x_m^k represents samples that have a different gesture pattern with x_i^j. We randomly pooled such sample pairs from the training sets during each training step. Likewise, for E_S, we applied a similar triplet loss $\mathcal{L}_{trip\_s}$ that shared a similar concept with $\mathcal{L}_{trip\_p}$.

We also used the cross-pattern and cross-identity reconstruction to enhance the separation of these two representations. The condition is formulated as follows,

$$\mathcal{L}_{cross\_recon} = \mathbb{E}[\|D(E_S(x_i^l), E_P(x_m^j)) - x_i^j\|] \tag{4}$$

where $i \neq m, l \neq j$. This condition encourages the extracted task-relevant and identity-relevant components to be combined into reconstructions that share the corresponding patterns and identities with the input. Summing the above terms, we obtain our total loss as:

$$\mathcal{L} = \mathcal{L}_{recon} + \lambda_1 \mathcal{L}_{cross\_recon} + \lambda_2 \mathcal{L}_{trip_p} + \lambda_3 \mathcal{L}_{trip_s} \tag{5}$$

where $\{\lambda_1, \lambda_2, \lambda_2\}$ are the balance weights. In all the experiments, the weights are set to $\{1, 0.5, 0.5\}$, respectively.

D. Implementation Details

All the components in the autoencoder were implemented based on 2-D convolution layers (see in table I). It is worth noting that we rearranged the 1024-dimension feature vectors to a $4 \times 16 \times 16$ dimension tensor to preserve its original spatial layout. The encoders progressively downsample the input by a stride of 2, whereas the decoder upsamples the latent features with a factor of 2 by bi-linear upsampling on each layer. The

TABLE I: The full architecture of the proposed network

Module	Layers	k	s	In/Out
E_P	Conv+ IN + LReLu	3	2	4/32
	Conv+ IN+ LReLu	3	2	32/64
	Conv+ IN+ LReLu	3	2	64/128
E_S	Conv+ IN+ LReLu+ AP	3	1	4/32
	Conv+ IN+ LReLu + AP	3	1	32/64
	Conv+ IN+ LReLu+ AP	3	1	64/128
D	US+ Conv+ DO+ LReLu	3	1	256/128
	US+ Conv+ DO+ LReLu	3	1	128/64
	US+ Conv	3	1	64/4

Conv, IN, LReLu, AP, UpS and DO denote convolution, Instance Normalization, Leaky ReLU, average pooling, upsampling and Dropout layers, respectively. All of the convolution layers use zero padding. k and s denote the kernel width and stride, respectively. In/Out in the rightmost column represents the channel number of the input and output.

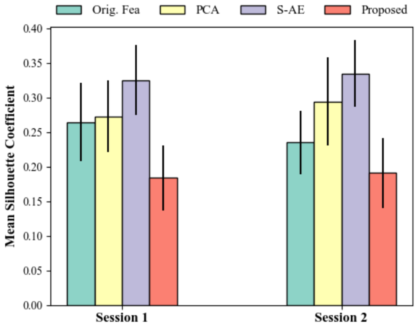

Fig. 3: The mean SIL coefficient (labeled by identities) obtained on original feature (Orig. Fea), feature projection by PCA, Standard Autoencoder (S-AE), and disentangled representations.

TABLE II: The success rate of identity inference attacks with 5% identities labels disclosed to attackers (lower accuracy is better).

method	Session 1			Session 2		
	SVM	KNN	RF	SVM	KNN	RF
Orig.Fea	56.41	30.28	38.72	50.99	28.30	32.76
PCA	45.91	30.07	44.13	39.73	27.74	40.23
S-AE	50.0	39.08	42.39	43.49	34.86	37.39
Proposed	**23.22**	**20.13**	**22.91**	**24.35**	**19.62**	**23.10**

learned two latent vectors from each encoder are combined by concatenating the two parts along the channel axis before feeding into the decoder. A data augmentation strategy in [5] was applied to the training data.

E. Validation protocols and Metrics

We estimated the level of the identity information preserved in disentangled task-relevant representations from two metrics. The first is the clustering indicator based on the mean silhouette (SIL) coefficient labeled by identities. The second is the success rate of identity inference, defined as the probability of identifying the person in the dataset to whom a feature vector belongs. In addition, we implemented the disentangled features for hand gesture recognition tasks. We evaluated the performance of the disentangled representations with a set of classifiers on hand gesture recognition tasks, under a rigorous cross-day validation. By doing so, the long-term usability of the decomposed task-relevant components was considered.

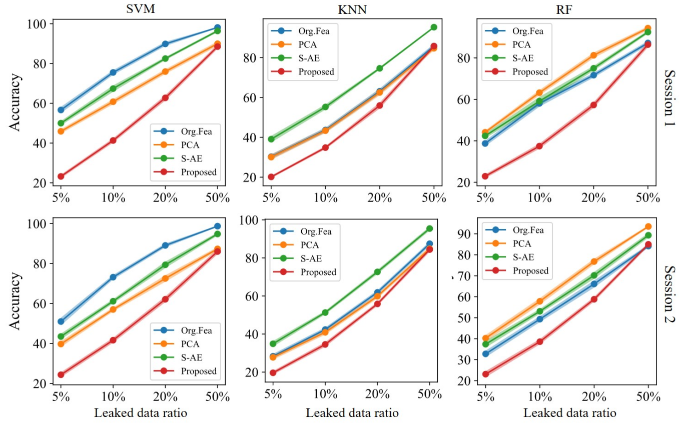

Fig. 4: The success rate of identity inference attacks in scenarios where a fraction of labeled data is disclosed to the attackers. Different classifiers are used to test the original features, features transformed by PCA, S-AE, and the proposed method. Shaded areas represent the 95% confidence interval.

Specifically, data acquired in session 1 was used to train the proposed disentangling model. The trained model was subsequently utilized to transform the original features from both session 1 and session 2 into their task-relevant representations. The SIL coefficient was computed on the yielded disentangled representations. The user identification performance was evaluated in simulation scenarios where attackers obtained a small fraction of identity labels in the dataset. In this work, the identity label disclosure ratio was set to $\{5\%, 10\%, 20\%, 50\%\}$. Each setting was tested for 10 independent runs. For gesture recognition, the decomposed task-relevant features from session 1 were used to train the classifiers, and data from session 2 were used for testing.

We used linear support vector machine (SVM), Random Forest (RF), and K-Nearest Neighbor (KNN) as classifiers for both identity recognition and hand gesture recognition tasks. Results obtained by original features (**Orig. Fea**), projected features by Principal Components Analysis (**PCA**) with a 95% variance ratio, and latent vectors generated by standard autoencoder (**S-AE**) are served as standard benchmarks to be compared with. The paired t-test between results obtained by the proposed method and its counterparts was used for statistical analysis. The differences were considered statistically significant for $p < 0.05$.

III. RESULT

A. Clustering Indicators on Identities Space

We computed the SIL coefficient to measure how similar a feature vector from an individual was to its own cluster compared to other clusters on the identity space. A lower mean SIL coefficient indicates that it is more difficult to find a decision boundary to determine the identities of the feature vectors in the dataset. As shown in Fig. 3, the SIL coefficient obtained on the original features is 0.264 and 0.235 in session 1 and session 2, respectively. After applying PCA, the projected features show higher SIL coefficient values in session 1 and session 2 with a significant difference ($p < 0.001$). The significant improvement of SIL scores is also observed on the latent vectors obtained by S-AE ($p < 0.001$). In contrast, the decomposed task-relevant components using the proposed methods contribute to the lowest SIL score in both sessions among all methods. The differences between SIL score obtained on the original features and the decomposed task-relevant components are significant in both sessions ($p < 0.001$). This demonstrates the capacity of proposed methods to filter out the identity-related information from the original features, thus improving the ability to defend against privacy attacks.

B. Identity Inference Attacks

We evaluated the success rate of identity inference attacks under the stimulation that 5% of the data with identity labels was disclosed to the attackers. As shown in table II, using a very small quantity of labeled data, the attackers can manage to train classifiers that map the remaining original features to their identities with a high accuracy of 56.41%, which was ten times higher than the random guess baseline $(1/N_s, 5\%)$. This indicates that direct sharing of the original features may impose serious privacy threats on the enrolled users. Though PCA and S-AE compressed the original features onto a lower-dimensional feature subspace, they failed to lower the success rate of the identity recognition, indicating that component selection based on variance alone is ineffective in eliminating identity information. In contrast, with semantic information considered, the disentangled task-relevant representations could withstand the re-identification attacks to a greater extent, which was evidenced by the sharp decrease in the success rate of identity inferences. Fig.4 depicts the trend of the success rate of identification recognition by increasing the amount of leaked data. As expected, the accuracy increased sharply as the percentage of leaked data grows. Nevertheless, the trained classifiers always show the least success on features transformed by our proposed method in almost all cases, showing its potential to preserve users' privacy information.

C. Data Utility on Primary Pattern Recognition Tasks

Despite the good performance in defending against privacy attacks, the performance of disentangled representations should be maintained to ensure their utility in pattern recognition tasks. We used the disentangled representations from data in session 1 as training data and session 2 as testing data. The accuracy of hand gesture recognition under cross-day validation is shown in Table III. We observed that the utilized classifiers consistently achieve the best classification accuracy on the disentangled task-relevant representations, significantly outperforming the original features and the generated features by PCA and S-AE ($p < 0.01$ for all testing pairs). This demonstrates the feasibility of the proposed method in improving the current data-sharing mechanism for developing privacy-preserving sEMG-based pattern recognition systems.

TABLE III: The accuracy of hand gesture recognition.

Methods	SVM	KNN	RF
Orig. Fea	80.81	74.03	77.74
PCA	85.13	78.73	77.81
S-AE	91.0	89.99	88.76
Proposed	**93.85**	**93.16**	**93.51**

IV. Discussion

In this work, we investigated the potential privacy threats in sEMG-based pattern recognition applications. We extracted the subject-invariant task-relevant representations from original sEMG features by feature disentanglement. The disentangled task-relevant representations show their superior capacity to resist privacy attacks over other state-of-the-art feature projection methods while maintaining high accuracy in the downstream pattern recognition tasks. Overall, this work provides a new perspective in optimizing current sEMG data-sharing mechanism, and thus can boost the development of more secure and reliable sEMG-based applications. Future works should focus on further optimizing the current method to bring its performance to defend against identity inference attacks to the next level by reducing the success rates to approximately that of random guess.

References

[1] J. Fan, X. Jiang, X. Liu, X. Zhao, X. Ye, C. Dai, M. Akay, and W. Chen, "Cancelable hd-semg biometric identification via deep feature learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1782–1793, 2022.

[2] G. R. Naik, S. E. Selvan, and H. T. Nguyen, "Single-channel emg classification with ensemble-empirical-mode-decomposition-based ica for diagnosing neuromuscular disorders," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 7, pp. 734–743, 2016.

[3] X. Jiang, X. Liu, J. Fan, X. Ye, C. Dai, E. A. Clancy, M. Akay, and W. Chen, "Open access dataset, toolbox and benchmark processing results of high-density surface electromyogram recordings," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1035–1046, 2021.

[4] U. Côté-Allard, C. L. Fall, A. Drouin, A. Campeau-Lecours, C. Gosselin, K. Glette, F. Laviolette, and B. Gosselin, "Deep learning for electromyographic hand gesture signal classification using transfer learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp. 760–771, 2019.

[5] X. Jiang, X. Liu, J. Fan, X. Ye, C. Dai, E. A. Clancy, D. Farina, and W. Chen, "Optimization of hd-semg-based cross-day hand gesture classification by optimal feature extraction and data augmentation," *IEEE Transactions on Human-Machine Systems*, pp. 1–11, 2022.