Abstract

In recent years, the convergence of human-in-the-loop (HITL) systems and intelligent automation has led to a significant rise in the development of interactive systems where humans and machines engage in seamless communication and perform tasks in the same working space. Human-computer interaction (HCI), human-machine interface (HMI), and human-robot interaction (HRI) form the foundational pillars of this multidisciplinary advancement, each addressing distinct yet overlapping aspects of interaction. HCI traditionally focuses on how human users interact with computational systems through graphical, tactile, or voice-based modalities, facilitated by various sensing interfaces. HMI extends this paradigm to encompass a broader range of electromechanical systems, including industrial devices and wearable sensors. Whereas, HRI is a more specialized domain within this spectrum that explores the dynamic interplay between humans and robotic agents. It emphasizes bidirectional perception, shared control, and adaptive learning mechanisms to foster intuitive and responsive interaction in real-world scenarios.

As these paradigms mature and are deployed in working environments, *human-robot collaborative tasks* (HRCTs) gain greater prominence, particularly in the domains requiring physical and cognitive cooperation between humans and robots. In such collaborative scenarios, robots are no longer mere tools for isolated tasks but are intelligent partners capable of interpreting human intent and responding to nuanced environmental cues. This advancement requires robust perception modules, context-aware decision-making, and mutual adaptability to manage uncertainty and ensure safety under various complex environments. The integration of diverse sensor modalities that capture physiological and behavioral human cues offers promising avenues for bridging the gap between human intention and robotic execution. Such meaningful interaction facilitates shared autonomy through various learning tasks for an assistive robot in applications such as rehabilitation, assistive manipulation, navigation, and various other collaborative operations in regular and unstructured environments.

In the context of human-robot interaction (HRI), the choice of sensor modalities plays a very crucial role in interpreting human cues. It enables responsive collaboration between human users and robotic agents and ensures safety in shared environments. Sensor modalities can be broadly classified into two main categories i.e., contact-based and non-contact-based, each offering specific advantages depending on the working environment. Contact-based or wearable sensors, such as data gloves, magnetic trackers, surface electromyography (sEMG), inertial measurement units (IMUs), accelerometers, etc., require physical attachment to the user's body to detect human intent in the form of muscle activity, kinematic movement parameters, applied force/torque, etc. These provide rich physiological information that is crucial for precise recognition of hand gestures, hand movement, hand activity, and other motion intent estimation. These sensors can provide high signal fidelity and better temporal resolution, making them particularly effective for capturing muscle activity, joint dynamics, and force-related human cues. Such cues are critical for tasks involving fine motor control such as prosthetic manipulation, assistive rehabilitation, health monitoring, gesture recognition, etc. However, these modalities require precise placement, good connectivity, and stability during data acquisition, and may cause user discomfort during prolonged and continuous use. Moreover, these are also susceptible to motion artifacts, electrode displacement, and signal drift over time, which can affect the data quality. Additionally, some systems may require calibration or recalibration across sessions or users due to inter-session and inter-subject variability, increasing complexity in the setup and handling. The need for direct skin contact, especially in bio-signal sensors like sEMG, can lead to skin irritation and introduce variability in skin-electrode contact resistance, especially during long-term deployments.

On the contrary, non-contact sensors, such as RGB cameras, infrared thermographic imagers, or depth sensors can acquire human intent information from a distant place without direct physical contact. These modalities enable vision-based cue detection, body posture analysis, human action recognition, and gesture tracking, making them particularly advantageous in scenarios where minimum physical interference and user comfort are desired. These are especially essential in scenarios such as remote monitoring, robot guidance in dynamic environments, teleoperation in hazardous environments, activity recognition in ambient assisted living, and interaction support in socially assistive robotics. While non-contact systems offer flexibility and ease of deployment, they are susceptible to environmental variations like lighting conditions, occlusions, or background clutter, which can degrade the

recognition accuracy.

This Thesis investigates both non-contact approaches, such as vision sensing for visual cue interpretation in robot navigation guidance and assistive robotics environments, and contact-based modalities, such as sEMG for muscular activation analysis in prosthetic control and rehabilitation strategies, across separate studies. These complementary explorations highlight how different sensing strategies can be effectively designed for specific human-robot interaction or human-machine interface contexts, contributing to the development of robust and application-specific HRI/HMI frameworks suitable for deployment in real-world environments.

Despite the advancements in sensor technologies and data acquisition systems for human-robot interaction, several practical issues are encountered in real-world deployments. One of the foremost issues is sensor-generated and environmental noise, which can significantly impair the signal or image quality and deteriorate their subsequent processing. Vision-based systems are particularly vulnerable to photometric irregularities such as lighting conditions, contrast variations, and background clutter, all of which can disrupt the consistency in visual cue detection. Similarly, wearable and contact-based sensors such as sEMG, IMU, and accelerometer are prone to motion artifacts, crosstalk, sensor noise, and electrode displacement. These interferences often introduce random fluctuations or irregular distortions that reduce the reliability of feature extraction, especially during prolonged sessions or dynamic task execution in unstructured environments. Another critical challenge lies in intra-subject and inter-subject variability. Visual patterns, physiological signals, and behavioral expressions of intent can vary widely across individuals due to anatomical, neuromuscular, and habitual differences. Visual cues may be inconsistently performed or perceived due to human variability and contextual differences across social and geographical boundaries. With wearable sensors, even within the same subject, variations may occur across sessions due to muscle fatigue, inconsistent sensor placement, or deviated motion dynamics. This inherent diversity in data characteristics makes the development of generalizable models challenging and necessitates adaptive or personalized strategies. Additionally, background clutter in the environment can significantly affect the performance of non-contact modalities like vision sensors. Complex or dynamic backgrounds make it challenging to isolate relevant visual features, such as hand gestures or body posture, which are critical for accurate interaction modeling. This affects the robustness of visual feature extraction and can lead to higher false positives or recognition failures in cluttered or real-world environments.

A further concern is the high dimensionality of sensor data, particularly when dealing with large-sized image frames or rich signals such as EMG waveforms. Processing such data in real-time is computationally expensive and often impractical for embedded or robotic platforms. This highlights the need for effective dimensionality reduction (DR) techniques that can retain the discriminative and structural information of the data while reducing the computational burden. The use of manifold learning techniques, such as those explored in this Thesis, provides a promising direction for balancing information preservation and computational efficiency. In particular, this Thesis thoroughly investigates the family of linear approximation-based manifold learning-inspired approaches, with a primary focus on the *locality preserving projection* (LPP), which effectively preserves the local structural information of the data in the projected subspace.

Moreover, a persistent challenge remains in the generalization of models across different application scenarios. Algorithms trained on controlled environments or synthetic datasets often underperform when exposed to unstructured environments with unseen conditions. This lack of knowledge transferability necessitates the development of adaptive learning strategies that can dynamically adjust to changing scenarios, sensor drifts, and user-specific variations without requiring exhaustive retraining and offering more robustness to the system. Eventually, real-world deployment of HRI systems must keep a balance between accuracy and computational latency. High recognition accuracy is essential for reliable interaction, but the requirement of computational resources must be sufficiently low to support the practical deployment of such algorithms on embedded systems. This trade-off becomes particularly crucial in scenarios involving real-world control, shared autonomy, or safety-critical tasks. In the context of developing countries, where access to high-performance computing resources may be limited, the need for cost-effective, energy-efficient, and resource-constrained solutions becomes even more pronounced. Consequently, designing models that are both efficient, reliable, and will be able to work well without relying on costly hardware remains a major challenge in building practical and widely usable interaction strategies for HRI systems.

As mentioned above, dimensionality reduction techniques play a crucial role in simplifying the input representations by projecting the original high-dimensional data onto lower-dimensional subspaces. They need to do so while preserving the essential features of the data, required for interpretation and classification in the reduced space. Effective dimensionality reduction not only enhances computational efficiency but also improves model

generalizability and robustness by filtering out redundant or irrelevant information from the data. Traditional global dimensionality reduction methods such as principal component analysis (PCA), linear discriminant analysis (LDA), and independent component analysis (ICA) have been widely used due to their ease of implementation and computational efficiency. However, these techniques often work under the assumption of global linearity and may fail to preserve the local geometric relationships inherent in complex sensor data. Being linear dimensionality reduction techniques, these techniques rely on the assumption that the underlying structure of the data can be effectively captured using linear transformations. However, sensor data in HRI applications such as vision-based cues or sEMG signals often exhibit nonlinear characteristics due to various factors like anatomical differences, complex motion dynamics, and environmental hazards. In such cases, linear techniques may be inadequate for uncovering the true latent structure of the data. This has led to the exploration of nonlinear dimensionality reduction approaches, particularly the family of manifold learning techniques such as isomap, locally linear embedding (LLE), and Laplacian eigenmaps (LE), which attempt to discover low-dimensional embeddings that preserve the intrinsic geometry of the data. However, many of these methods lack an explicit mapping function, making it difficult to handle new samples under real-world implementations. To address these challenges, this Thesis focuses on locality preserving projections (LPP), a DR technique that explores the nonlinear local structure of the data by linear approximation strategies, while enabling efficient computation and generalization to new data points. LPP captures the intrinsic local structure of the high-dimensional data by constructing a nearest-neighbor graph and projecting the data onto a subspace that preserves such local relationships. Unlike PCA or LDA, LPP does not seek global variance maximization or class separability alone, but rather emphasizes preserving intrinsic neighborhood proximity, making it highly effective in scenarios involving dynamic motion cues, irregular visual cues, sensor disturbances, and human variability. Its ability to produce an explicit linear mapping and enable straightforward projection for the new test samples are critical for real-world deployment in embedded systems.

While the locality preserving projection offers clear advantages in preserving local geometric structures of high-dimensional data, several issues still persist with the traditional form of this technique, under irregular data conditions. One of the major concerns lies in the sensitivity of graph construction to noise and outliers. Traditional LPP relies on Euclidean distance-based similarity graphs, which can be heavily influenced by lighting variation and

background clutter in vision-based systems, or artifacts arising from sensor disturbances and motion dynamics. This reliance on fixed-distance kernels makes the resulting weight matrix vulnerable under non-ideal conditions, thereby reducing the discriminative power of the projected subspace. Another important aspect is that standard LPP-based approaches ignore the feature-specific importance of the data. Across various sensing modalities in HRI applications, not all the features contribute equally to human intent recognition or task discrimination. However, the conventional similarity graph used in LPP construction does not account for this relevance or discriminative capacity of individual features from the data. It focuses solely on spatial proximity between data samples, thereby discarding valuable information from the feature domain. As a result, the generated subspace may remain inadequate to emphasize semantically meaningful components of the data, especially in multimodal sensor data with heterogeneous characteristics. Additionally, the similarity matrix in standard LPP is restricted to Euclidean geometry, which assumes that the underlying samples lie on a flat, linear manifold. In practical scenarios, such as with biological signals e.g., sEMG, or with natural visual scenes, the data often resides on complex, nonlinear manifolds (e.g., Riemannian). The Euclidean assumption thus limits LPP's ability to fully encode spatial and contextual relationships among samples, leading to suboptimal embedding, especially in cases of data corruption. Encoding of nonlinear data structures requires models that are capable of adapting to curvature and topology in the data space. Furthermore, the use of inflexible similarity matrices, such as fixed Gaussian kernels, restricts LPP's adaptability across different sensing conditions and modalities. A fixed kernel may not effectively model relationships across varying sensor anomalies, noise levels, or inter-class/intra-class separations. The absence of adaptive distance modeling mechanisms limits LPP's scalability in heterogeneous datasets, such as those containing new data samples with previously unseen corruption. This creates a strong necessity for context-aware variants of LPP that can dynamically modulate the similarity function based on the statistical and geometric properties of the input data.

Taken together, these limitations motivate the need for robust, adaptive, and semantically aware extensions of LPP that can handle the diverse and noisy nature of HRI data. Several variants are developed in this Thesis work by incorporating adaptive spatial kernels, granular computing-aided kernel fusion, noise-resilient distance metrics (e.g., Euler, Grassmannian), uncertainty-aware similarity fusion, sparsity-inducing regularizations, etc. These adaptations not only enhance class separability and subspace stability but also ensure that the learned

embeddings remain interpretable and computationally viable for deployment in real-world, embedded platforms inside HRI frameworks. The entire work can be broadly perceived in two research verticals i.e., (i) LPP-based strategies for vision sensor-based modalities and (ii) LPP-based strategies for wearable sensor-based modalities. The augmented LPP variants from both these research verticals majorly focus on offering more accurate, reliable, robust, and computationally inexpensive solutions. Their performance was extensively evaluated across both vision and wearable sensor modalities, and their performance outcomes are summarized below.

The proposed LPP-based frameworks demonstrated notable performance across both vision and wearable sensor modalities. For vision-based symbolic cue data, the conventional LPP achieved about 99% accuracy under normal illumination but deteriorated to nearly 55% under the darkest condition. The proposed ALPSK-BLPP framework improved this to approximately 71%, while granular computing-based extensions such as REGF-2DLPP and dNG-2DRLPP further enhanced performance up to 81% and 85%, respectively. Under noisy conditions, the proposed LTrP-BLPP algorithm achieved around 85% accuracy at 35% salt-and-pepper noise density and 89% under speckle noise with $variance = 0.35$, outperforming the traditional LPP by $12 - 15\%$ accuracy margins. Similarly, variants like HRLTP-BLPP maintained about $80 - 84\%$ accuracy under simultaneous illumination degradation (dark level-2) and noise corruption (either of salt-and-pepper or speckle noise), and dNG-2DRLPP achieved $\sim 82\%$ under Gaussian noise ($variance = 0.35$). For wearable sEMG data, the proposed UaBMA-OLPP and RPNG-OLPP models respectively achieved around 88% and 89% recognition accuracies, showing higher resilience than the classical LPP ($\sim 81\%$) framework. The robust RPNG-OLPP variant maintained $77 - 84\%$ accuracy even under noisy conditions for the sEMG dataset.