Emotion-Aware AI-Integrated Sign Language System Using Quantum Machine Learning for Inclusive Multimedia Communication

Dr. Mariyan Richard A[1]
Department of Computer Science and Applications
Atria Institute of technology, India
Email: mariyanrich01@gmail.com

Dr. Hemant Kumar Singh[2]
khwaja Moinuddin Chishti Language University,
Lucknow,India
Email: hemantbib@gmail.com

***Abstract*—Sign language is a complete form of communication in which hand movement, facial expression, body posture, and movement intensity work together to convey meaning. Many existing accessibility systems concentrate on recognizing signs or generating subtitles, but they do not adequately preserve the emotional information present in a signing person's ex-pression. This paper presents an emotion-aware sign language communication framework that combines computer vision, deep learning, multimodal feature fusion, and Quantum Machine Learning (QML). The proposed system first detects the hands, face, and body from video and extracts spatial, temporal, and emotional features. CNN-based processing is considered for visual feature extraction, while LSTM or Transformer models are used to capture the sequence of movements. The resulting compact feature representation is then supplied to a hybrid classical-quantum classifier using a Variational Quantum Circuit (VQC) or a quantum kernel. The recognized sign and emotional state are converted into text or an avatar-based sign representation. The system is intended for use in educational media, video content, news broadcasting, and other multimedia applications. The re-search will evaluate recognition accuracy, emotion classification, response time, and user acceptance, with a direct comparison between classical and hybrid classical-quantum models. The main aim is to move sign language accessibility from simple word-level translation toward communication that also retains emotional expression.**

***Keywords* —Sign language recognition, emotion recognition, multimodal learning, computer vision, deep learning, quantum machine learning, assistive technology, multimedia accessibility.**

I. INTRODUCTION

Communication is more than the exchange of words. In a normal conversation, people use facial expressions, body posture, hand movements, and changes in movement intensity to make their meaning clear. Sign language depends on these elements even more strongly. A signer may use the position and movement of the hands to express a word, while the face and body can provide additional information about the intention or emotion behind that sign.

For deaf and hard-of-hearing people, sign language is there-fore not simply a collection of hand gestures. It is a visual language with its own structure and expressive features. When communication is converted only into written subtitles, part of this information is lost. A subtitle may identify the words being communicated, but it normally does not show whether the signer is happy, serious, angry, surprised, or emotionally engaged.

Artificial Intelligence has created new opportunities for au-tomatic sign language recognition. Computer vision methods can locate hands and faces, while deep learning models can learn patterns from images and video sequences. Convolutional Neural Networks (CNNs) are useful for extracting spatial information, and sequence models such as Long Short-Term Memory (LSTM) networks and Transformers can learn how a gesture changes over time. These methods have made automatic recognition more practical, but emotion-aware in-terpretation remains a difficult problem.

The research proposal supporting this work identifies sev-eral limitations in current accessibility technologies. Subtitles provide linguistic information but do not adequately represent emotional context. Existing sign language recognition systems can also have limited expressive understanding, while gesture, facial, and motion information are difficult to model together. In addition, integration with mainstream multimedia platforms and efficient

processing of complex gesture data remain challenging.

In this paper, a practical hybrid architecture is proposed. Classical computer vision and deep learning are used for the first stages because video data is high-dimensional. Quantum Machine Learning is introduced after feature extraction, where a smaller feature vector can be encoded into a quantum circuit. This design allows the study to compare classical and quantum-assisted classification without assuming that quantum computing will automatically outperform classical learning.

The work has four main goals. First, it aims to recognize sign language from continuous visual input. Second, it aims to identify emotional information from facial expression, posture, and gesture dynamics. Third, it investigates whether a hybrid classical-quantum classifier is useful for these features. Finally, it considers how the resulting system can be connected to multimedia applications.

II. Background and Related Work

A. *Sign Language Recognition*

Sign language recognition has been studied using image processing, handcrafted features, machine learning, and deep neural networks. Earlier systems commonly described a hand using shape, contour, or orientation features. Such methods can work under controlled conditions, but changes in lighting, background, camera position, and signer style can reduce their performance.

Deep learning changed this approach by allowing models to learn useful representations directly from visual data. CNNs can identify local patterns such as hand shape and appearance. For video, however, a single frame is not sufficient. The order and speed of movements also matter. LSTM networks and Transformer-based models are therefore suitable for represent-ing temporal information.

Continuous sign language recognition is harder than isolated sign recognition because the system must process a sequence and identify transitions between signs. Koller *et al.* investi-gated continuous sign language recognition using statistical and deep learning approaches [1], [2].

B. *Emotion Recognition*

Emotion recognition usually uses facial expression, speech, posture, or a combination of these signals. In sign language, facial and body information is particularly important because expression is part of the communication itself. A system that uses only the hand region can miss changes in facial expression and movement intensity.

This motivates a multimodal design in which gesture, face, posture, and temporal information are processed together. The objective is to recognize not only the linguistic meaning of a sign but also the emotional context associated with it.

C. *Quantum Machine Learning*

Quantum Machine Learning is an emerging approach that combines machine learning with quantum computation. Two techniques considered in this research are Variational Quantum Circuits and quantum kernels.

A hybrid model is more suitable for the proposed applica-tion than an entirely quantum model because current quantum devices are still limited in qubit count, noise level, and circuit depth. The quantum part can therefore be used as a classifier or feature transformation after classical preprocessing.

Quantum machine learning in feature spaces has been studied by Schuld and Killoran [9]. Circuit-centric quantum classifiers and variational quantum algorithms have also been investigated as potential approaches for classification and optimization [10], [11].

III. Problem Statement

Most accessible multimedia communication systems are designed around text. Although captions and subtitles are useful, they do not preserve the complete visual and emotional information of sign language. Existing recognition systems may identify a gesture correctly but still fail to capture the way in which the gesture is expressed.

The problem addressed in this research is the development of a real-time, multimodal sign language system that recognizes both linguistic content and emotional context and that can be integrated into multimedia applications. The system must combine several types of information without making the processing pipeline too slow for interactive use.

A. Research Gap

The main research gap is the limited connection between sign recognition, emotion recognition, and practical multime-dia integration. There is also an opportunity to investigate QML within this multimodal setting.
The research therefore focuses on the following question:

Can a hybrid classical-quantum model improve or efficiently support classification of multimodal sign and emotion features when compared with a classi-cal deep-learning baseline?

PROPOSED METHODOLOGY

The proposed methodology follows seven phases:

1) Data collection and preprocessing
2) Feature extraction
3) Multimodal fusion
4) Quantum Machine Learning integration
5) Emotion-aware recognition
6) Gesture synthesis and translation
7) System integration and evaluation

This workflow follows the methodological direction of the research proposal.

A. *Data Collection and Preprocessing*

The study will use sign language datasets such as Ameri-can Sign Language (ASL) and Indian Sign Language (ISL), together with emotion-labelled gesture data where appropriate. The selected datasets will be divided into training, validation, and test sets.

Care will be taken to avoid data leakage between these sets, particularly when multiple samples belong to the same signer. For

each video, frames are sampled and passed through computer vision modules. Hand tracking is used to obtain hand

landmarks, face detection provides facial regions, and pose estimation provides body landmarks.

The extracted coordinates can be normalized relative to the body so that the model is less dependent on camera distance and absolute position. Data augmentation may be applied to improve robustness to small changes in scale, position, and motion.

B. *Feature Extraction*

Three groups of features are considered.

1) *Spatial Features:* Spatial features include:

- Hand shape
- Hand orientation
- Hand position
- Facial landmarks
- Body posture

CNN-based models can be used for visual feature extraction.

2) *Temporal Features:* Temporal features describe how the gesture changes over time. They include:

- Movement direction
- Gesture duration
- Movement speed
- Acceleration
- Repeated movement patterns

LSTM or Transformer models can be used to learn temporal dependencies.

3) *Emotional Features:* Emotional features can be extracted from:

- Facial expression
- Eyebrow and mouth movement
- Head movement
- Gesture intensity
- Body posture

The resulting features are combined into a multimodal representation.

C. Multimodal Fusion

Let the extracted feature groups be represented as

$$F = [F_g, F_f, F_p, F_t] \quad (1)$$

where F_g represents gesture features, F_f represents facial features, F_p represents posture features, and F_t represents temporal features.

The fused representation is supplied to the classification module.

The use of multiple modalities is intended to reduce the limitations of relying only on hand gestures.

IV. Quantum Machine Learning Integration

A. Hybrid Classical-Quantum Architecture

The proposed system uses a hybrid classical-quantum ar-chitecture rather than attempting to process raw video directly using a quantum computer.

The classical component performs visual processing and generates a compact feature vector. The feature vector is then encoded into a quantum circuit and processed using a Variational Quantum Circuit (VQC) or quantum kernel.

The overall flow is:

A parameterized circuit is subsequently applied:

$$|\psi(x, \vartheta)\rangle = V(\vartheta)U(x)|0\rangle^{\otimes n} \quad (4)$$

where ϑ represents trainable circuit parameters.

The parameters can be optimized using

$$\vartheta^* = \arg\min L(y, \hat{y}) \quad (5)$$

where L is the classification loss, y is the target class, and $\hat{y}$ is the predicted class.

C. Quantum Kernel

A second approach is to use a quantum kernel. A quantum feature map transforms the classical input x into a quantum state:

$$x \rightarrow |\phi(x)\rangle \quad (6)$$

The similarity between two samples can then be represented by

$$K(x_i, x_j) = |\langle\phi(x_i)|\phi(x_j)\rangle|^2 \quad (7)$$

The resulting kernel matrix can be supplied to a classical classifier.

Both approaches will be considered experimentally rather than assuming that one approach will always be superior.

V. EMOTION-AWARE SIGN LANGUAGE RECOGNITION

The key feature of the proposed framework is the combi-nation of linguistic recognition and emotional recognition.

A conventional sign recognition system may produce: “Congratulations”

The proposed system attempts to provide the linguistic meaning together with an emotional interpretation: “Congratulations” + positive/excited emotional con-text.

The final prediction can be expressed as

$$Y = f(F_g, F_f, F_p, F_t) \qquad (8)$$

where Y contains the predicted sign and emotional information.

Possible emotion classes include neutral, happy, sad, angry, fear, and surprise. The final classes will depend on the annotation available in the selected dataset.

B. Variational Quantum Circuit

For an input feature vector x, the quantum encoding operation can be represented as

$$|\psi(x)\rangle = U(x)|0\rangle^{\otimes n} \quad (3)$$

where $U(x)$ is the feature encoding operation and n is the number of qubits.

VI. GESTURE SYNTHESIS AND TRANSLATION

Following recognition, the system generates an appropriate output representation. Three communication directions are considered.

A. Sign-to-Text

The input sign-language video is converted into text:

Sign Video → Recognition → Text (9)

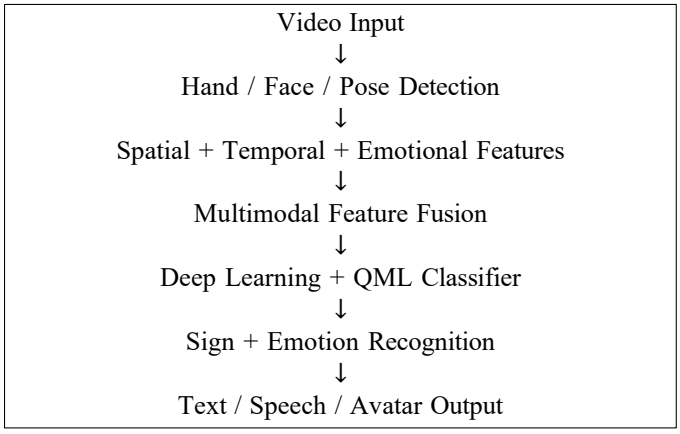


Fig. 1. Proposed processing architecture.

B. Speech-to-Sign

Speech is converted into text and then mapped to a sign sequence:

Speech → Speech Recognition → Text → Sign Sequence (10)

C. Text-to-Sign

Text is mapped to sign-language gestures and displayed using an animated avatar:

Text → Sign Mapping → Avatar (11)

An avatar-based module can provide a consistent visual representation of the generated sign sequence. Neural approaches to sign language translation have pre-viously demonstrated the feasibility of learning mappings between spoken-language representations and sign sequences [4], [5].

VII. MULTIMEDIA SYSTEM INTEGRATION

The proposed framework is designed as an API-based architecture to facilitate integration with external applications.

Potential application areas include:

· Online video platforms

- News broadcasting
- Educational multimedia
- Digital classrooms
- Public information systems
- Accessible entertainment

A multimedia platform can send a video stream or selected

Algorithm 1 Emotion-Aware Sign Recognition

1: Capture or load sign-language video
2: Detect hands, face, and body landmarks
3: Normalize frame sequences
4: Extract spatial and temporal features
5: Extract emotional features
6: Fuse multimodal representations
7: Reduce feature dimensionality
8: Encode features for QML
9: Apply VQC or quantum kernel
10: Predict sign and emotional category
11: Generate text or avatar output
12: Record performance metrics

quantum framework such as Qiskit or Cirq can be used for the QML experiments.

The implementation will be divided into independent mod-ules so that the classical baseline and hybrid model can be tested using the same extracted features.

The first baseline will use a conventional deep-learning clas-sifier. The second model will use multimodal fusion without a quantum component. The third model will use a VQC or quantum-kernel classifier.

Keeping preprocessing and data splits consistent across these configurations will make the comparison more mean-ingful.

A. Overall Processing Algorithm

X. Experimental Evaluation

The proposed system will be evaluated using several complementary metrics.

A. Recognition Accuracy

Recognition accuracy is defined as

$$Accuracy = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \times 100 \quad (12)$$

Accuracy will be reported for sign recognition and, where appropriate, emotion classification.

B. Precision, Recall and F1-score

Precision is defined as

$$Precision = \frac{TP}{TP + FP}$$

Recall is defined as frames to the recognition engine and receive text, sign, and emotional information in return.

VIII. Implementation Approach

A prototype can be implemented using a Python-based machine learning environment. Computer vision components can be connected to a deep-learning framework, These measures are useful when the emotion classes are not equally represented.

TABLE I
EXPERIMENTAL COMPARISON PLAN

Model	Input	Classifier	Results
M1	Gesture	Classical DL	To be measured
M2	Gesture + Face + Pose	Multimodal DL	To be measured
M3	Gesture + Face + Pose	DL + QML	To be measured

C. *Response Latency*

Real-time performance will be measured using response latency:

$$Latency = T_{output} - T_{input} \quad (16)$$

where T_{input} is the time at which an input segment is received and T_{output} is the time at which the corresponding result is produced.

D. *User Acceptance*

User evaluation will consider:

- Ease of use
- Translation clarity
- Emotional expressiveness
- Response speed
- Overall satisfaction

XI. COMPARATIVE EXPERIMENTAL DESIGN

Three model configurations will be compared.

The first model represents a gesture-only baseline. The second model investigates whether multimodal information improves recognition. The third model adds the QML com-ponent.

Numerical values are intentionally not reported at this stage because the research proposal does not contain measured experimental results. The values in Table I will be replaced by actual results after implementation and testing.

XII. RESULTS AND DISCUSSION

At the current research stage, the system is a proposed framework and numerical performance values are therefore not presented as measured results.

The expected outcome is that multimodal processing will provide more information than gesture-only recognition be-cause facial expression and movement dynamics contribute additional context.

The QML experiment will determine whether the quantum classifier can match or improve the classical baseline under the same feature representation and whether the additional quantum processing affects response time.

A successful prototype should provide a clear sign predic-tion together with an emotion label and should produce the result quickly enough for interactive communication.

The research proposal identifies improved communication clarity, emotional expressiveness, scalable multimedia

integra-tion, and contributions to inclusive technology and assistive AI as important expected outcomes.

A. Interpretation of Possible Outcomes

If the multimodal model performs better than the gesture-only model, this would support the importance of combining gesture, face, and posture.

If the QML model performs similarly to the classical model but requires more time, the result would indicate that QML is feasible but may not yet be advantageous for real-time deployment.

If the QML model provides better classification with ac-ceptable latency, the finding would justify further investigation using larger datasets and more advanced quantum circuits.

The study will therefore treat QML as an experimentally testable component rather than as a guaranteed improvement.

XIII. APPLICATIONS AND SIGNIFICANCE

The proposed framework has direct applications in accessi-ble education. Recorded lectures and online learning content could provide automatic sign-language interpretation in addi-tion to captions.

News broadcasting is another application in which real-time sign output could make public information more accessible.

For online video and entertainment, an avatar-based sign representation could be generated alongside the original con-tent.

The same recognition engine could also support communi-cation between signing and non-signing users by converting signs into text and text or speech into signs.

The significance of the work is both technical and social. Technically, it investigates multimodal AI and QML in a chal-lenging sequence-recognition problem. From an accessibility perspective, it aims to preserve information that is normally lost when sign language is reduced to subtitles.

XIV. CHALLENGES AND LIMITATIONS

The quality of the dataset will strongly affect the system. Sign languages differ across regions, and ASL and ISL cannot simply be treated as the same language.

Emotion labels may also be subjective, making annotation consistency important.

Real-time performance is another challenge. Hand tracking, face analysis, pose estimation, and temporal modeling must operate continuously.

Quantum circuits can add additional execution overhead, particularly when run on remote quantum hardware. The number of qubits and circuit depth also restrict the amount of information that can be encoded directly.

Finally, user evaluation should include people who actually use sign language. A technically accurate model may still be difficult to use if the interface is slow, unclear, or unnatural.

XV. FUTURE WORK

Future work will focus on collecting larger multimodal datasets with synchronized sign, facial, posture, and emotion annotations.

Continuous sign-language translation using Transformer ar-chitectures can be investigated after establishing a reliable baseline.

On the quantum side, different feature maps, VQC struc-tures, and quantum kernels can be compared. The effect of noise and circuit depth can also be studied using simulators and available quantum hardware. Edge-based inference is another possible direction for re-ducing latency and keeping video data closer to the user. The avatar module can also be improved by using realistic three-dimensional models and learned motion generation. In the longer term, the complete system can be tested with real multimedia streams rather than only prepared datasets.

XVI. CONCLUSION

This paper proposed an emotion-aware AI-integrated sign language system for inclusive multimedia communication. The central idea is to recognize sign language as a multimodal form of communication in which hand movement, facial expression, posture, and temporal dynamics should be considered together. The proposed architecture uses classical computer vision and deep learning for visual processing and introduces Quan-tum Machine Learning at the compact feature-classification stage. VQC and quantum-kernel approaches provide two possible QML implementations.

The system also includes sign-to-text, speech-to-sign, and text-to-sign pathways with an avatar-based output mechanism. The research will compare gesture-only, multimodal clas-sical, and hybrid classical-quantum models using recognition

accuracy, F1-score, latency, and user acceptance.

Rather than assuming a quantum advantage, the study will establish whether QML offers a measurable benefit for this problem. The overall goal is to develop a practi-cal accessibility framework that communicates not only the meaning of a sign but also more of the emotion carried by the signer.

REFERENCES

[1] O. Koller, J. Forster, and H. Ney, "Continuous Sign Language Recog-nition: Towards Large Vocabulary Statistical Recognition Systems Han-dling Multiple Signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, 2015.

[2] O. Koller, H. Ney, and R. Bowden, "Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3793–3802.

[3] O. Koller, C. Camgoz, H. Ney, and R. Bowden, "Deep Sign: Hybrid CNN-HMM for Continuous Sign Language Recognition," in *Proc. British Machine Vision Conference (BMVC)*, 2016.

[4] C. A. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Neural Sign Language Translation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7784–7793.

[5] C. A. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign Lan-guage Transformers: Joint End-to-End Sign Language Recognition and Translation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10023–10033.

[6] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] A. Vaswani *et al.*, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436–444, 2015.

[9] M. Schuld and N. Killoran, "Quantum Machine Learning in Feature Hilbert Spaces," *Physical Review Letters*, vol. 122, no. 4, 2019.

[10] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, “Circuit-Centric Quantum Classifiers,” *Physical Review A*, vol. 101, no. 3, 2020.
[11] M. Cerezo *et al.*, “Variational Quantum Algorithms,” *Nature Reviews Physics*, vol. 3, pp. 625–644, 2021.
[12] A. Peruzzo *et al.*, “A Variational Eigenvalue Solver on a Photonic Quantum Processor,” *Nature Communications*, vol. 5, 2014.