Multisource Approaches to Italian Sign Language (LIS) Recognition: Insights from the MultiMedaLIS Dataset

Gaia Caligiore*†[1], Raffaele Mineo†[2], Concetto Spampinato[2], Egidio Ragonese[2], Simone Palazzo[2], Sabina Fontana[2]

[1] *University of Modena Reggio-Emilia, Italy.*
[2] *University of Catania, Italy.*

Abstract

Given their status as unwritten visual-gestural languages, research on the automatic recognition of sign languages has increasingly implemented multisource capturing tools for data collection and processing. This paper explores advancements in Italian Sign Language (LIS) recognition using a multimodal dataset in the medical domain: the MultiMedaLIS Dataset. We investigate the integration of RGB frames, depth data, optical flow, and skeletal information to develop and evaluate two computational models: Skeleton-Based Graph Convolutional Network (SL-GCN) and Spatiotemporal Separable Convolutional Network (SSTCN). RADAR data was collected but not included in the testing phase. Our experiments validate the effectiveness of these models in enhancing the accuracy and robustness of isolated LIS signs recognition. Our findings highlight the potential of multisource approaches in computational linguistics to improve linguistic accessibility and inclusivity for members of the signing community.

Keywords

Italian Sign Language, Sign Language Recognition, Deep Learning, Computer Vision

1. Introduction

Italian Sign Language (LIS- Lingua dei Segni Italiana) is the primary means of communication within the Italian signing community. Due to their visual-gestural modality, sign languages (SLs) were initially not considered fully-fledged linguistic systems. However, since the 1960s, beginning with Stokoe's pioneering works [1], the contemporary study of SLs has evolved into a robust field of research. Over the past half-century, significant societal and scientific advancements have transformed the perception and status of SLs, now recognized as natural and complete languages, having received legal recognition in many countries.

In the Italian context, the study of signed communication began in the early 1980s, involving both hearing and deaf researchers. At that time, what we now call LIS was still mostly unnamed and was often referred to as 'mime' or 'gesture' by both signers and non-signers alike [2]. The first significant publications on LIS [3] [4], along with the collaborative efforts of deaf and hearing researchers, initiated a transformative period in SL research in the Italian context [5]. This shift in perspective was influenced by factors beyond the language itself, such as increased meta-linguistic awareness and greater visibility of the community and its language to the wider public. In fact, from a societal perspective, the visibility of SL in Italy, especially in media, has significantly changed with technological advancements, mirroring global trends.

In the late 1980s, Italy introduced subtitles in movies on television, marking a step toward content accessibility. The importance of media accessibility, through subtitles or LIS interpreting, was accentuated during the COVID-19 pandemic. The need for equitable access to critical information for deaf individuals became evident, with efforts born within the community

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

* Corresponding author.
† These authors contributed equally.
gaia.caligiore@unimore.it (G. Caligiore); raffaele.mineo@phd.unict.it (R. Mineo); concetto.spampinato@unict.it (C. Spampinato); egidio.ragonese@unict.it (E. Ragonese); simone.palazzo@unict.it (S. Palazzo); sfontana@unict.it (S. Fontana).

000-0002-7087-1819 (G. Caligiore), 0000-0002-1171-5672 (R. Mineo); 0000-0001-6653-2577 (C. Spampinato); 0000-0001-6893-7076 (E. Ragonese); 0000-0002-2441-0982 (S. Palazzo); 0000-0003-3083-1676 (S. Fontana)

stressing the central role of LIS in ensuring that the deaf signers received accessible information during challenging times [6], highlighting the significant communication barriers that deaf individuals face, especially when in-person interactions were restricted. This increased visibility, along with persistent advocacy by the signing community, played a crucial role in the official recognition of LIS and Tactile LIS (LISt) in May 2021.

Within this evolving societal and linguistic framework, the increased media visibility of LIS and the introduction of video capturing tools in daily lives, language collection emerges as a central issue. For SLs, the need for comprehensive collections is particularly significant. Unlike oral languages, which in some cases have developed standardized written systems, SLs must rely on video collections to capture signed communication accurately. These videos, whether raw or annotated, are essential for analyzing SLs with both qualitative and quantitative evidence.

2. Automatic Sign Language Recognition

The development and use of preferably annotated SL datasets or corpora are crucial for training and validating automatic recognition models, and access to high-quality data from diverse SLs and cultural contexts enhances the generalizability of these solutions. Comprehensive data collections of this kind ensures that models can effectively understand and process the wide range of linguistic and cultural nuances present in different SLs.

In the domain of automatic sign language recognition (SLR) of LIS, the integration of visual and spatial information presents a complex challenge. As mentioned, LIS operates through the visual-gestural channel. More precisely, it is characterized as multimodal[2] (signed discourse is comprised of manual and body components) and multilinear (manual and body components are performed simultaneously) [2]. Recent advancements in SLR have been significantly driven by annotated datasets, which serve as the basis for training and validating models [7, 8, 9, 10, 11].

Machine learning technologies, particularly deep learning neural networks, have facilitated the development of more precise and robust models for SL interpretation. These models are able to refine their performance through training on diverse and complex datasets. Additionally, computer vision plays a central role in this field by enabling real-time analysis and interpretation of body and manual components [2] that is hand movements, facial expressions, and body posture [12, 13, 14, 15].

A significant challenge in applying deep learning and computer vision methods to SLR lies in ensuring the quality and adequacy of training data, which is essential for achieving optimal model performance.

Therefore, in this study, we focus on evaluating the efficacy of the MultiMedaLIS Dataset (Multimodal Medical LIS Dataset) and assessing various deep learning models for SLR which employ advanced deep learning techniques to interpret isolated signs by integrating diverse data types such as RGB video, depth information, optical flow, and skeletal data.

We benchmark our Dataset with two models: the Skeleton-Based Graph Convolutional Network (SL-GCN) and the Spatiotemporal Separable Convolutional Network (SSTCN). These models are trained on the MultiMedaLIS Dataset, showcasing how the incorporation of multisource data can enhance the accuracy of sign recognition. This approach aims at testing the potential of integrating different data modalities to improve the robustness and performance of SLR systems.

3. State of the Art

In this section, we discuss the state of the art from two perspectives considered during our work on the Dataset: LIS data collection and SLR tools

3.1. LIS Data Collections

SL researchers in Italy have been actively engaged in the creation of LIS corpora and datasets. This effort involves a complex process of video data collection and annotation, as SL datasets can vary significantly depending on their intended use. Within this context, SL data collections can be categorized into two main types. The first type includes datasets that feature videos depicting continuous signing, capturing the flow and context of natural SL usage. The second type comprises datasets that focus on isolated signs, which are individual signs presented separately from continuous discourse.

The scarcity of available LIS data collections has prompted researchers to develop their own resources. Several smaller-scale LIS corpora have been

[2] Given our group's interdisciplinarity, we found "multimodal" can mean different things depending on one's background: in linguistics, it refers to the employment of manual and body components while signing, while in computer vision, it means using multiple capturing tools. To differentiate, we use "multisource" for capturing tools. Thus, "multimodal" in this text follows SL linguistics terminology.

independently established, each serving distinct purposes based on the type of data collected.

The methodologies employed for collecting LIS data encompass a diverse array of approaches, ranging from naming tasks to semi-structured and spontaneous interviews with deaf signers, to video recording sessions involving hearing individuals learning LIS as a second language (L2) or second modality (M2) [16]. These documentations serve equally diverse purposes, ranging from documenting the language itself to creating tools for automatic translation highlighting the ongoing commitment of researchers to expand and enrich the available resources for studying LIS [17, 18, 19, 20, 21, 22, 23, 24].

Despite the predominant private nature of corpora collections, an exception to the accessibility challenge is found in the online dictionary SpreadTheSign, a project originating in 2004. Initially conceived as a dictionary for SLs, SpreadTheSign has evolved into a versatile resource for language documentation [25]. Another significant resource is the Corpus LIS, recognized as the largest collection of spontaneous, semi-structured, and structured videos in LIS by deaf signers. The primary objectives of this corpus were twofold: to collect a substantial quantity of data suitable for quantitative analysis and to establish a comprehensive representation of LIS usage in Italy [26, 27, 28].

3.2. SLR Tools

Like SL data collections, SLR approaches can be broadly classified into two main categories: those that rely on specialized hardware and those that use visual information. The former employ specialized hardware, such as gloves able to capture precise hand movements. While these systems can provide detailed data, they are often considered intrusive and can compromise the natural flow of communication. Additionally, they are unable to capture the full spectrum of SLs, which includes manual and body components. In contrast, vision-based approaches use visual information captured by cameras, including RGB, depth, infrared, or a combination of these. These methods are less intrusive for users, as they do not require the use of special equipment.

In SLR, a challenge lies in effectively capturing both body movements and specific motions of hands, arms, and face. For instance, [29] introduces a multi-scale, multi-modal framework that focuses on spatial details across different scales. This approach involves each visual modality capturing spatial information uniquely, supported by a system operating at three temporal scales. The training methodology emphasizes precise initialization of individual modalities and progressive fusion via ModDrop, which enhances overall robustness and performance.

Another study proposes an iterative optimization alignment network tailored for weakly supervised continuous SLR [30]. The framework employs a 3D residual convolutional network for feature extraction, complemented by an encoder-decoder architecture featuring LSTM decoders and Connectionist Temporal Classification (CTC).

[31] introduces a 3D convolutional neural network enhanced with an attention module, designed to extract spatiotemporal features directly from raw video data. In contrast, [32] combines bidirectional recurrence and temporal convolutions, emphasizing temporal information's effectiveness in sign tasks, although not covering the full spectrum of movements. Moreover, [33] employs CNNs, a Feature Pooling Module, and LSTM networks to generate distinctive visual representations but falls short in capturing comprehensive movements and signing.

However, as previously noted, RGB-based SLR systems can raise privacy concerns, particularly when processing visual data in cloud environments or for machine learning training [34]. Addressing these issues, radio frequency (RF) sensors have emerged as a promising alternative, ensuring privacy preservation while enabling innovative data representations for SLR. In the literature, deep learning techniques have been applied to various RF modalities such as ultra-wideband (UWB) [35], Doppler [36], continuous wave (CW) [37], micro-Doppler [38], frequency modulated continuous wave (FMCW) [14], multi-antenna systems [39], and millimeter waves [40].

As part of the Dataset discussed in this work, we have also collected RADAR data and are actively analyzing it. However, preliminary results are not available at this time, so they are not included in this report. Currently, RADAR-based solutions have demonstrated robust performance across diverse environmental conditions, highlighting the productivity of incorporating this sensor technology in data collection efforts. Nevertheless, many existing RADAR solutions are tailored to recognizing a limited set of signs, highlighting the ongoing challenge of expanding vocabulary recognition capabilities in datasets like the one discussed in the following section.

4. The MultiMedaLIS Dataset

The MultiMedaLIS [41] Dataset was created thanks to the interdisciplinary collaboration established between the Department of Humanities (DISUM) and the Department of Electrical, Electronic and Computer Engineering (DIEEI) of the University of Catania (Unict). It aims to offer a multimodal collection of LIS signs specifically focused on medical contexts.

For the data recording protocol, the DIEEI group developed a customized recording software to collect the

LIS data, supplemented with a desktop computer and a modified keyboard transformed into a pedal board. This pedal board, equipped with two pedals, allowed hands-free navigation of the software, enabling users to move forward (by pushing on the right pedal) or backward (by pushing on the left pedal) while maintaining a neutral recording position[3]. During sessions, one of 126 Italian labels or alphabet letters was displayed on a screen, with adjustable display time for preparation and transition from one sign to the other. Each recording started from a neutral position, and the right pedal marked the completion of a sign. If errors occurred, the left pedal allowed re-recording. The software's interface features a color-coded background: yellow for preparation and green for recording. Additionally, it supports flexible data expansion, accepting word lists from text files for easy customization in future collections.

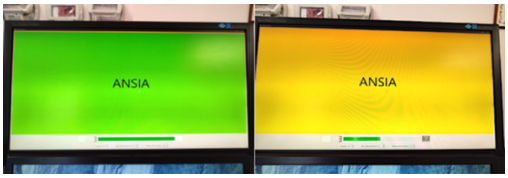

Figure 1: User interface display presented during the recording phase (green) and preparation phase (yellow).

After the recording process, Dataset included synchronized data capturing facial expressions, hand and body movements and comprises a total of 25,830 sign instances. This includes 205 repetitions of 100 different signs and the 26 signs of the LIS alphabet [41]. Beyond these 26 signs, the signs included in the MultiMedaLIS Dataset can be broadly categorized into two groups [42]: semantically marked signs related to health and health issues, and non-semantically marked signs. It is important to note that while the first group of signs is categorized as semantically marked, this classification does not imply that these signs belong exclusively to a specialized jargon lexicon. The decision to categorize signs as semantically marked was driven by their significance in contexts related to health and medical interactions in the post-pandemic world (hence, when the Dataset was first theorized). However, it was also important to include additional signs that could contribute to constructing meaningful utterances in patient-doctor interactions. During the creation of the MultiMedaLIS Dataset, careful consideration was given to selecting signs that could be combined to form coherent and meaningful utterances.

Regarding the specific form of signs, the MultiMedaLIS Dataset includes a lexicon of standard, isolated signs that are not combined within utterances. These signs reflect forms commonly found in online dictionaries and educational materials. To ensure the accuracy of the data, sign variants performed by a professional LIS interpreter during the collection of a test dataset were compared with the same variants found in the online dictionary SpreadTheSign. This comparison aimed to select documented versions of each sign for inclusion in the Dataset. By incorporating these documented variants, we aimed to enhance its precision, reliability, and real-world applicability. This approach contributed to ensuring that the Dataset aligns with established standards and supports effective research and application in the field of LIS.

When discussing recording tools for state-of-the-art multimodal corpora in the Italian context, such as the Corpus LIS [27] and the CORMIP [43] the emphasis is placed on the portability and non-invasiveness of these tools. This approach ensures minimal interference with the signer's natural environment and activities.

Portable and non-invasive recording tools are chosen specifically for their ability to capture data in familiar, and sometimes domestic, settings without disrupting the signer's surroundings, aiming to maintain the authenticity of the signed interactions and minimize any discomfort or distraction for the participants.

To capture LIS for recognition with minimal invasiveness we integrated a combination of recording tools. A 60GHz RADAR sensor, employed to capture detailed manual motion data, provided Time- and Frequency-Domain data and Range Doppler Maps for distinguishing moving objects at 13 fps. For more structured depth and facial recognition data, the Realsense D455 depth camera and Kinect v1 were incorporated. The Realsense D455, equipped with dual infrared cameras and RGB mode, captured depth data at 848x480 pixels and RGB data at 1280x720 pixels, both at 30 fps, enabling the tracking of facial expressions through 68 facial points. The Zed v1 and Zed v2 cameras provided high-resolution stereoscopic data, recording at 1920x1080 pixels and 25 fps, with capabilities for generating depth maps and 3D point clouds. Additionally, the Zed v2 offered tracking for 18 body points in both 2D and 3D [41].

[3] The neutral recording position referenced is a seated position in which the user has their arms extended along the sides of the torso, elbows bent at 90°, and palms facing downward [41].

Figure 2: Combination of synchronized infrared and depth data from the MultiMedaLIS Dataset.

By prioritizing portability and non-invasiveness, high-quality data can be still collected, while respecting the privacy and comfort of the individuals recorded. Anonymization is achieved through the use of the RADAR sensor, which we introduced specifically to address privacy concerns inherent in face-to-face signed communication.

5. Testing the Dataset

The MultiMedaLIS Dataset was designed with the aim of supporting the development of SLR models by enabling the collection and integration of information through various data modalities:

- RGB frames: images extracted from videos.
- Depth data: three-dimensional information for each RGB frame
- Optical flow: to emphasize movement
- Skeletal data: face landmarks and body joints

One of the main components of the Dataset are RGB frames, which are images extracted from videos. These frames provide a two-dimensional visual representation of the signs performed by the signer, capturing details such as hand positions and facial expressions. The Dataset includes depth data, providing a three-dimensional aspect to the images. allowing for more detailed information on the distance and relative position of elements in the scene. This type of data is particularly useful for understanding the spatial dynamics of signs.

Alongside RGB and depth data, the MultiMedaLIS Dataset also contains optical flow information, which describes the movement between consecutive frames. Optical flow is essential for capturing the direction and speed of movements, providing a more detailed understanding of the transitions between various signs. Finally, the Dataset includes skeletal data, representing face landmarks and body joints, allowing for precise tracking of joint and body segment positions, facilitating the analysis of signs in terms of joint movements.

Managing this multimodal data is an emerging topic in computational linguistics. By combining different sources of information, it is possible to significantly improve the performance of SLR models. For example, integrating depth data with RGB frames can provide a more complete representation of signs, while adding optical flow and skeletal data can further enrich the analysis of movement's temporal structure. In our view, the MultiMedaLIS Dataset provides a solid foundation for exploring these combinations, allowing researchers to develop more effective and accurate solutions for SLR.

6. Models and Architectures

In the context of automatic SLR, various approaches and model architectures have been tested to leverage the characteristics of multimodal data in the MultiMedaLIS Dataset.

The SL-GCN (Skeleton-Based Graph Convolutional Network) represents a significant innovation in this field. This model generates skeletal data from videos and creates temporal graphs that capture the spatiotemporal relationships between joint movements. Through fine-tuning and the combination of different data streams, SL-GCN has demonstrated high accuracy in sign recognition [44] [45].

Another prominent architecture is the SSTCN (Spatiotemporal Separable Convolutional Network) [46], which excels in feature extraction from videos using HRNet [47]. This approach has shown an accuracy of 96.33%, highlighting its effectiveness in capturing spatial and temporal dynamics of LIS signs.

RGB frames are crucial for the visual representation of signs. The process of splitting videos into frames, cropping, and normalization optimally prepares the data for analysis by deep learning models. The use of dense optical flow presents significant challenges in sign recognition. Optical flow extraction using the Farneback algorithm [48] led to 56% accuracy, highlighting difficulties in capturing precise details of movements, alongside computational limitations. Depth data encoded with Height, Horizontal disparity, Angle (HHA) represent another crucial resource in the MultiMedaLIS Dataset. Applying HHA encoding to depth frames achieved 88% accuracy using the ResNet(2+1)D architecture [49], substantiating importance of three-dimensional information in enhancing understanding and interpretation of signs, offering a more detailed perspective compared to two-dimensional data.

7. Training and Evaluation Procedure

For the training of the models, we employed a multi-stream approach that integrates skeletal, RGB, and depth data to improve sign recognition accuracy. The models were trained on a NVIDIA Tesla T4 16GB GPU using the Adam optimizer with an initial learning rate of 0.001 and a batch size of 8. We applied cross-validation to ensure the robustness of the results, splitting the Dataset into training (70%) and validation (15%) subsets and data augmentation techniques, such as color jittering, changing the brightness, contrast, saturation and hue, to

increase the diversity of the training data and improve generalization.

The loss function adopted for training was categorical cross-entropy, appropriate for multi-class classification tasks. The models were trained for a maximum of 100 epochs, with an early stopping criterion set to terminate training if no improvement in validation loss was observed for 10 consecutive epochs. For evaluation, we used a test set comprising 15% of the Dataset, ensuring that the models were tested on unseen data.

8. Results

The results demonstrate the model's efficiency in leveraging multi-modal data for improved outcomes. As can be seen in Table 1, the SL-GCN multi-stream model achieved the best accuracy, with a Top-1 accuracy of 97.98% and a Top-5 accuracy of 99.94%, surpassing the performance of models using single data streams such as skeletal joints, bones, or motion alone. This demonstrates the advantage of combining multiple streams of information to capture both spatial and temporal dynamics of signs.

Table 1
Performance of SL-GCN multi-stream on the test set

Data	Accuracy Top-1 (%)	Accuracy Top-5(%)
Joints	96.24	99.84
Bones	95.82	99.84
Joint Motion	90.37	99.15
Bone Motion	92.69	99.52
Multi-stream	97.98	99.94

In Table 2, datasets trained on the SL-GCN model are compared. Our Dataset produced the highest accuracy (97.98%) among the datasets evaluated, outperforming larger datasets like AUTSL (95.45%).

Table 2
Comparison of different datasets on SL-GCN model

Dataset	Number of signs	Accuracy (%)
MultiMedaLIS	126	97.98
AUTSL	226	95.45
ASLLVD	20	61.04
Alphabet	26	85.19

Table 3 presents a comparison of different methods across the entire Dataset. The SL-GCN trained on RGB frames achieved the highest accuracy (97.98%), followed by the SSTCN model with 96.33%. The ResNet(2+1)D architecture showed strong performance when applied to RGB frames (97.29%), but struggled when using optical flow data alone, reaching just 56.31% accuracy, suggesting that while the optical flow provides valuable information on motion, it lacks the richness of spatial features found in RGB and depth data. The HHA-encoded depth data, when processed with the ResNet(2+1)D model, achieved an accuracy of 88.04%, confirming that depth information is complementary, but not as effective as RGB data in isolation.

Table 3
Performance of various methods on the MultiMedaLIS Dataset

Methods	Dataset	Accuracy(%)
SS-CGN	RGB	97.98
SSTCN	RGB	96.33
ResNet(2+1)D Optical Flow	RGB	56.31
ResNet(2+1)D Frame	RGB	97.29
ResNet(2+1)D Encoding HHA	Depth	88.04

The results highlight importance of combining multiple data modalities, especially RGB and skeletal data, for improving the accuracy and robustness of SLR systems. The performance of the SL-GCN model with multi-stream data shows the model's ability to effectively capture signs, as well as the Dataset's value.

9. Discussion and Conclusion

In this study, our goal was to demonstrate our first steps into testing the efficacy of the MultiMedaLIS Dataset in contributing to the advancement of the field of SLR through multisource approaches. The integration of RGB frames, depth data, optical flow, and skeletal data has provided a comprehensive basis for developing and evaluating SLR models. Our experiments with the SL-GCN and SSTCN architectures have highlighted advancements in recognizing isolated LIS signs in medical semantic contexts, given the domain of our Dataset.

The SL-GCN model, trained on skeletal data to construct temporal graphs, achieved accuracy in capturing spatiotemporal relationships critical to sign recognition. This approach not only enhances the precision of rendering LIS signs but is also reinforced by a Dataset able to support robust graph-based convolutional networks in multimodal SLR tasks. At the same time, our Dataset proved robust, precise and variable enough for SSTCN model testing, focusing on spatiotemporal separable convolutions, revealing robust performance in extracting spatial dynamics from RGB frames.

Having validated the visual modalities on the mentioned models, we have promising preliminary results on adapting these models to accept RADAR data. We plan to extract the pre-trained RADAR data

processing module and use it independently during inference. This approach will eliminate the need for RGB visual data. Furthermore, we plan to expand the Dataset by applying the same protocol with 10 deaf signers. This will effectively increase the current Dataset, enhancing the generalizability across different signers. Our goal is to develop an autonomous, resource-constrained system (thanks to the exclusion of RGB data) that operates on-edge or even offline. This cost-effective solution can be used in any emergency contexts where direct access to interpreting is not available.

References

[1] W. Stokoe, Sign language structure: an outline of the visual communication systems of the American deaf, University of Buffalo, Buffalo, New York, 1960.

[2] V. Volterra, M. Roccaforte, A. Di Renzo, S. Fontana, Italian Sign Language from a Cognitive and Socio-semiotic Perspective. Implications for a general language theory, John Benjamins Publishing Company, Amsterdam-Philadelphia, 2022.

[3] M. Montanini, M. Facchini, L. Fruggeri, Dal Gesto al Gesto: il bambino sordo tra gesto e parola, Cappelli, Bologna, 1979.

[4] V. Volterra, I segni come le parole: la comunicazione dei sordi, Boringhieri, Torino, 1981.

[5] S. Fontana, S. Corazza, P. Boyes-Braem, V. Volterra, Language research and language community change: Italian Sign Language (LIS) 1981-2013, in volume 236 of the International Journal of the Sociology of Language, 2015.

[6] E. Tomasuolo, T. Gulli, V. Volterra, S. Fontana, The Italian Deaf Community at the Time of Coronavirus, in volume 5 of Frontiers in Sociology, 2021.

[7] D. Li, C. R. Opazo, X. Yu and H. Li, Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison, in proceedings of the 2020 IEEE WACV, Snowmass, CO, USA, 2020, pp. 1448-1458.

[8] O. Mercanoglu Sincan, H. Yalim Keles, AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods, IEEE Access, 2020. https://doi.org/10.48550/arXiv.2008.00932

[9] H. R. Vaezi Joze, O. Koller, MS-ASL: A large-scale data set and benchmark for understanding American sign language, arXiv preprint arXiv, 2018.

[10] U. von Agris, M. Knorr and K. F. Kraiss, The significance of facial features for automatic sign language recognition, proceedings of the 8th IEEE International Conference on Automatic Face & Gesture Recognition, Amsterdam, Netherlands, 2008, pp. 1-6.

[11] S. Tornay, O. Aran, M. Magimai Doss, An HMM Approach with Inherent Model Selection for Sign Language and Gesture Recognition, In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 2020, pp. 6049-6056.

[12] Y. Chen, C. Shen, X. -S. Wei, L. Liu and J. Yang, Adversarial PoseNet: A Structure-Aware Convolutional Network for Human Pose Estimation, 2017 IEEE ICCV, 2017, pp. 1221-1230.

[13] E. Barsoum, C. Zhang, C. Canton Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, in Proceedings of the 18th ACM ICMI, 2016, pp. 279–283.

[14] Y. Wang, A. Ren, M. Zhou, W. Wang and X. Yang, A Novel Detection and Recognition Method for Continuous Hand Gesture Using FMCW Radar in volume 8 of IEEE Access, 2020, pp. 167264-167275.

[15] O. Yusuf, M. Habib, M. Moustafa, Real-time hand gesture recognition: Integrating skeleton-based data fusion and multi-stream CNN, 2024.

[16] A. Cardinaletti, L. Mantovan, Le Lingue dei Segni nel 'Volume Complementare' e l'Insegnamento della LIS nelle Università Italiane, 2, volume 14 of Italiano Lingua Seconda. Rivista internazionale di linguistica italiana e educazione linguistica, 2022, pp. 113-128.

[17] T. Russo Cardona, Iconicity and Productivity in Sign Language Discourse: An Analysis of Three LIS Discourse Registers, 2, volume 4 of Sign Language Studies, 200), pp. 164-197.

[18] A. Ricci, C. Bonsignori, A. Di Renzo, Che giorno è oggi? Prime analisi e riflessioni sull'espressione del tempo in LIS [Poster presentation], IV Convegno Nazionale LIS 'La Lingua dei Segni Italiana: una risorsa per il futuro', Rome, 2018.

[19] E. Fornasiero, La morfologia valutativa in LIS: una descrizione preliminare [Poster presentation], IV Convegno Nazionale LIS 'La Lingua dei Segni Italiana: una risorsa per il futuro', Rome, 2018.

[20] A. Di Renzo, A. Slonimska, L'uso delle Strutture di Grande Iconicità nei testi narrativi segnati: primi dati su bambini prescolari, scolari e adulti [Poster presentation], IV Convegno Nazionale LIS 'La Lingua dei Segni Italiana: una risorsa per il futuro', Rome, 2018.

[21] S. R. Conte, Nomi di persona e di luogo nella comunità sorda in Italia: interviste, analisi e primi risultati [Poster presentation], IV Convegno Nazionale LIS 'La Lingua dei Segni Italiana: una risorsa per il futuro', Rome, 2018.

[22] S. Fontana, E. Raniolo, Interazioni tra oralità e unità segniche: uno studio sulle labializzazioni nella Lingua dei Segni Italiana (LIS), in: G. Schneider, M. Janner, B. Élie (Eds.), Proceedings of the VII Dies Romanicus Turicensis, Peter Lang, Bern, 2015, pp. 241-258.

[23] V. Cuccio, G. Di Stasio, S. Fontana, On the Embodiment of Negation in Italian Sign Language: An Approach Based on Multiple Representation Theories, in volume 1 of Frontiers in Psychology, 2022.

[24] S. Fontana, Grammar and Experience: The Interplay Between Language Awareness and Attitude in Italian Sign Language (LIS), 5, volume 14 of the International Journal of Linguistics, 2022, pp. 1-18.

[25] M. Hilzensauer, K. Krammer, A multilingual dictionary for sign languages: 'SpreadTheSign', in proceedings of ICERI , Seville, 2015.

[26] C. Cecchetto, S. Giudice, E. Mereghetti, La raccolta del Corpus LIS, in: A. Cardinaletti, C. Cecchetto, C. Donati (Eds.), Grammatica, Lessico e Dimensioni di Variazione della LIS, FrancoAngeli, Milan, 2011, pp. 55-68.

[27] C. Geraci, K. Battaglia, A. Cardinaletti, C. Cecchetto, C. Donati, S. Giudice, E. Mereghetti, The LIS Corpus Project, in volume 11 of Sign Language Studies, 2011, pp. 528-571.

[28] M. Santoro, F. Poletti, L'Annotazione del Corpus, in: A. Cardinaletti, C. Cecchetto, C. Donati (Eds.), Grammatica, Lessico e Dimensioni di Variazione della LIS, FrancoAngeli, Milan, 2011, pp. 69-78.

[29] N. Neverova, C. Wolf, G. Taylor and F. Nebout, ModDrop: Adaptive Multi-Modal Gesture Recognition, in volume 8 of IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2016, pp. 1692-1706.

[30] J. Pu, W. Zhou, and H. Li, Iterative alignment network for continuous sign language recognition, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4165–4174.

[31] J. Huang, W. Zhou, H. Li and W. Li, Attention-Based 3D-CNNs for Large-Vocabulary Sign Language Recognition, in volume 29 of IEEE Transactions on Circuits and Systems for Video Technology, 2019, pp. 2822-2832.

[32] D. Bragg, T. Verhoef, C. Vogler, M. Morris, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort, N. Caselli, M. Huenerfauth, H. Kacorri, Sign language recognition, generation, and translation: An interdisciplinary perspective, in Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility, 2019, pp. 16 – 31.

[33] O. Mercanoglu Sincan, A. O. Tur and H. Yalim Keles, Isolated Sign Language Recognition with Multi-scale Features using LSTM, in proceedings of the 27th Signal Processing and Communications Applications Conference (SIU), Sivas, Turkey, 2019, pp. 1-4.

[34] S. Z. Gurbuz, A. C. Gurbuz, E. A. Malaia, D. J. Griffin, C. Crawford, M. M. Rahman, R. Aksu, E. Kurtoglu, R. Mdrafi, A. Anbuselvam, T Macks, E. Ozcelik, A linguistic perspective on radar micro-doppler analysis of American sign language, in proceedings of the 2020 IEEE International Radar Conference (RADAR), Washington, DC, USA, 2020, pp. 232-237.

[35] B. Li, Sign language/gesture recognition based on cumulative distribution density features using UWB radar, in volume 70 of IEEE TIM, 2021, pp. 1-13.

[36] H. Kulhandjian, Sign language gesture recognition using Doppler radar and deep learning" in proceedings of the 2019 IEEE Globecom Workshops (GC Wkshps), Waikoloa, HI, USA, 2019, pp. 1-6.

[37] Y. Lu, Y. Lang, Sign language recognition with CW radar and machine learning, proceedings of the 21st International Radar Symposium (IRS), Warsaw, Poland, 2020, pp. 31-34.

[38] J. McCleary, Sign language recognition using micro-doppler and explainable deep learning, in volume 139 of Computer Modeling in Engineering & Sciences 2024, 2024, pp. 2399-2450.

[39] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, volume 39 of IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2016, pp. 1137-1149.

[40] O. O. Adeoluwa, S. J. Kearney, E. Kurtoglu, C. J. Connors, S. Z. Gurbuz, near real-time ASL recognition using a millimeter wave radar, Proceedings of Volume 11742 of Radar Sensor Technology XXV, SPIE, 2021.

[41] R. Mineo, G. Caligiore, C. Spampinato, S. Fontana, S. Palazzo, E. Ragonese, Sign Language Recognition for Patient-Doctor Communication: A Multimedia/Multimodal Dataset, Proceedings of the IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI), 2024.

[42] G. Caligiore, Codifying the body: exploring the cognitive and socio-semiotic framework in building a multimodal Italian sign language (LIS) dataset [Ph.D. thesis], University of Catania, Catania, 2024.

[43] L. Lo Re, Corpus Multimodale dell'Italiano Parlato: basi metodologiche per la creazione di un

prototipo [Ph.D. thesis], University of Florence, Florence, 2022.

[44] C. Correia de Amorim, C. Macedo, C. Zanchettin, Spatial- Temporal Graph Convolutional Networks for Sign Language Recognition, Proceedings of the 2019 International Conference on Artificial Neural Networks, Munich, Germany, 2019, pp. 646-657.

[45] Ayas Faikar Nafis and Nanik Suciati, Sign language recognition on video data based on graph convolutional network. 18, volume 99 of Journal of Theoretical and Applied Information Technology, 2023, pp. 4323-4333.

[46] S. Jiang, B. Sun, L. Wang, Y. Bai, K Li, Y. Fu. Skeleton aware multi-modal sign language recognition, Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2021, pp. 5693-5703.

[47] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5693-5703.

[48] G. Farneback, Two-frame motion estimation based on polynomial expansion. Volume 2749 of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, pp. 363-370.

[49] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, & M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6450-6459.