Universal Access in the Information Society
https://doi.org/10.1007/s10209-023-01017-7

LONG PAPER

Greek sign language recognition for an education platform

Katerina Papadimitriou[1] · Gerasimos Potamianos[1] · Galini Sapountzaki[2] · Theodoros Goulas[3] · Eleni Efthimiou[3] · Stavroula-Evita Fotinea[3] · Petros Maragos[4]

Accepted: 5 July 2023
© The Author(s) 2023

Abstract
Recent advances in sign language (SL) technologies, along with demand for SL education, have led to increased interest in developing tools that enable automatic assessment of learners' SL video productions, helping both students and their instructors. At the very least, such tools should perform automatic SL recognition (SLR) of non-studio quality videos in a signer-independent (SI) fashion, thus providing simple binary feedback on learners' signing under realistic usage scenarios. Motivated by the above and the lack of any such tools for the Greek SL (GSL), we have been developing the "SL-ReDu" education platform for both receptive and productive GSL learning and student assessment. In this paper, we present our SLR module for GSL, developed for and integrated to the "SL-ReDu" system. The module incorporates state-of-the-art deep-learning based visual detection, feature extraction, and classification, operates in an SI mode on web-cam videos, and accommodates a small-size vocabulary of isolated signs and continuously fingerspelled letter sequences. We train the module on collected GSL data and demonstrate its superiority over a number of alternative SLR algorithms. We then conduct its objective evaluation within the "SL-ReDu" system and carry out a subjective evaluation of the overall platform, obtaining very satisfactory results in both.

Keywords Sign language recognition · Greek sign language · Education tools · Sign language assessment · MediaPipe · MobileNet · ResNet · BiLSTM

✉ Katerina Papadimitriou
aipapadimitriou@uth.gr

Gerasimos Potamianos
gpotam@ieee.org

Galini Sapountzaki
gsapountz@sed.uth.gr

Theodoros Goulas
tgoulas@athenarc.gr

Eleni Efthimiou
eleni_e@athenarc.gr

Stavroula-Evita Fotinea
evita@athenarc.gr

Petros Maragos
maragos@cs.ntua.gr

[1] Department of ECE, University of Thessaly, Volos, Greece

[2] Special Education Dept., University of Thessaly, Volos, Greece

[3] ILSP, Athena Research and Innovation Center, Athens, Greece

[4] School of ECE, National Technical University of Athens, Athens, Greece

1 Introduction

Sign languages (SLs) comprise a complex non-vocal form of communication, which occurs in the 3D visible space around the signer's upper torso, encapsulating both manual and non-manual articulation, each carrying gloss linguistic content [1]. Due to its intricacy, learning an SL as a second language (L2) constitutes a challenging and time-consuming process for both students and their instructors [2]. Importantly, students need regular feedback on their SL productions during learning, which may not be available at all times by their instructors (e.g., at home and at a leisurely learning pace). Further, the currently used assessment procedures yield significant instructor workload, as they rely exclusively on manual inspection of large amounts of video files of learners' productions or in vivo interaction with small groups of students, while also lacking both inter- and intra-instructor consistency and objectivity in grading [3, 4]. For example, at the Special Education Department at University of Thessaly (UTH-SED), the end-of-semester evaluation of L2 learners of Greek SL (GSL) that are enrolled in the corresponding

Published online: 23 July 2023

introductory curriculum compulsory course (typically 150 students) requires an average of 30 h for the in vivo procedure, while exam credibility suffers due to instructor fatigue during the process. In addition, the large student body forces the class focus to shift away from the resource-demanding instruction of GSL production, to the more manageable task of GSL perception at the expense of course coverage and quality. The above clearly demonstrate the demand for developing suitable automatic tools to support self-monitoring and objective evaluation for SL L2 learning.

Not surprisingly, there has been an increasing research activity to address such needs, enabled by continuously accelerating technological progress. For example, human-computer interaction tools have been developed to assess learners' perception of SL articulations [5, 6]. At the same time, recent deep-learning breakthroughs have propagated to automatic sign language recognition (SLR), allowing this technology to be incorporated into systems that address the need of assessing the validity of learners' signed productions. For example, "SignTutor" [7] constitutes an SLR-based system that enables teaching basic signs and evaluating learners' articulation, providing visual feedback on the correctness of the performed signs. In another work, "CopyCat" [8] comprises an educational adventure game for helping deaf children improve their language skills, and it is developed leveraging features from colored gloves and embedded accelerometers. The system introduced in [9] provides automatic feedback on handshape and movement correctness of Australian SL productions, based on input by a Kinect sensor. The work in [10] presents a game-based mobile application for sign vocabulary learning that provides immediate and appropriate feedback to the user based on machine learning and pattern recognition technologies. Further, another system is introduced in the context of the SMILE project [11], assessing productions of Swiss German SL, relying on automatic SLR. Finally, a web-based fingerspelled alphabet learning application for Indian SL is developed in [12], employing an automatic SLR approach.

Overall, the number of such implementations is rather small, with most requiring "enrolled" signers (known to the system) and / or employing special visual depth sensors or gloves for signing data input. Clearly, to better exploit SLR technology in the automatic assessment of SL productions under realistic use-case scenarios and larger learner populations, the SLR system module should operate in a signer-independent (SI) fashion and acquire video input from low-end cameras captured in non-studio quality conditions.

Motivated by the above, as well as the lack of learning tools in GSL, we have recently commenced the "SL-ReDu" project [13]. The project intends to pioneer a GSL learning and assessment system in an attempt to address the need for standardized L2 teaching, supporting both self-monitoring and objective evaluation of receptive and productive GSL learning. Covering such lag in GSL is the primary goal of this project, while we employ techniques that in future can be easily transferable to other SLs as well.

Toward these goals, in earlier work [14] we have presented our first version of the SL-ReDu prototype. As shown in Fig. 1, the developed interactive platform involves theory presentation of appropriate GSL learning material [15, 16] and enables both practice and testing via passive- and active-type exercises. Productive learning relies on our GSL recognition module presented in [14] that operates in an SI fashion on video input by a typical web-cam and provides binary feedback regarding the correctness of the signed productions. The module accommodates two recognition tasks, namely that of

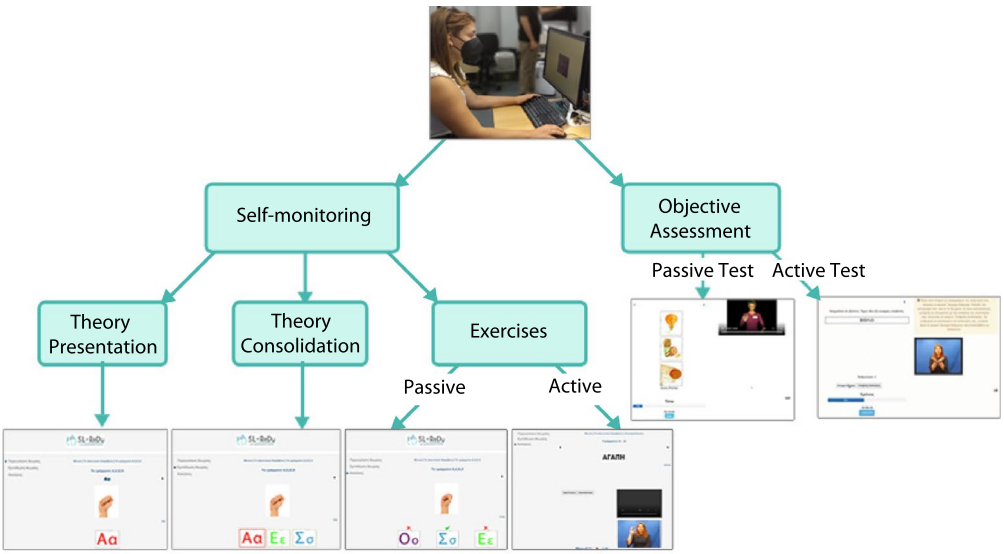

Fig. 1 Illustration of the SL-ReDu prototype system design

isolated signs within a small-size vocabulary and continuous fingerspelling (a crucial component of SLs [1]), aiming to validate our approach through objective and subjective evaluation, before broadening it in future to a richer GSL vocabulary and continuous signing. This article constitutes an extended version of [14] providing, in addition to our prior work:

- Extensive comparisons of the GSL recognition module to alternative SLR algorithmic approaches, demonstrating its superiority as evaluated on collected GSL data;
- A more detailed analysis of the module's objective evaluation and the SL-ReDu platform's subjective evaluation, abstracting new findings and demonstrating the success of our approach.

In more detail, our GSL recognition module [14] in the case of isolated SLR is based on a 3D convolutional neural network (3D-CNN), namely the ResNet2+1D network [17], whereas in the case of continuous fingerspelling it relies on an efficient 2D-CNN, namely MobileNet [18], in conjunction with a bidirectional long short-term memory (BiLSTM) encoder [19]. Here, we demonstrate the superiority of our SLR models by investigating a number of alternative approaches for processing the input video and extracting visual spatio-temporal features. In particular, we explore various skeletal representations, such as the 2D skeleton of the signer derived using the HRNet model [20], 3D skeletal features extracted via the MediaPipe holistic model [21], and 3D expressive body pose and shape parameterization obtained by the ExPose algorithm [22]. Further, we investigate deep learning-based appearance representations extracted from the raw RGB video data, exploring two 2D-CNN image feature learners, namely the ResNeXt-101 [23] and the InceptionNet-V3 [18], as well as 3D-CNN spatio-temporal feature learners, such as the Pseudo-3D Residual Network (P3D) [24] and the 3D ConvNet (C3D) [25]. Finally, we extract motion representations via different optical flow models, namely the SpyNet [26], FlowNet2 [27], and PWC-Net [28]. In addition, we combine the resulting visual streams with a BiLSTM encoder for spatio-temporal feature extraction, before feeding them to the classifier. In the case of fingerspelling, apart from the BiLSTM, we also investigate the use of a bidirectional-GRU (BiGRU) encoder [29]. The various recognition networks are trained and evaluated under both multi-signer and signer-independent experimental paradigms on a suitable GSL corpus, collected in [14].

Moreover, we report the objective evaluation campaign results of the SL-ReDu platform for the GSL production exercises (active-type) available in the system, as well as the subjective evaluation based on user responses to a questionnaire. Finally, we split the participants evaluation pool into two groups according to the duration of their GSL learning exposure (less or more than five months), providing helpful insights concerning both objective and subjective evaluation results.

The remainder of the paper is organized as follows: The SL-ReDu platform is described in Sect. 2; the SLR algorithms are presented in Sect. 3; the GSL corpus and recognition experiments are covered in Sect. 4; the SL-ReDu prototype system user evaluation findings are discussed in Sect. 5; and the paper is concluded in Sect. 6.

2 The SL-ReDu prototype system

The SL-ReDu platform integrates our GSL recognition module [14], together with an appropriate human–machine interface, into an innovative environment for evaluating the signing performance of GSL learners. As shown in Fig. 1, the system involves two main modalities: self-monitoring and objective evaluation of receptive and productive GSL learning, thus attempting to address the shortcomings of traditional practice and assessment procedures in GSL L2 learning.

2.1 Platform Linguistic content

The material of the platform is inferred from the definition of the language material for levels A0-A1 of the Common European Framework for Languages (CEFRL) for GSL as L2. In particular, the linguistic content for both GSL perception and production of the initial system evaluation covers two subsets: (i) single word units, including both numerals and non-numerals, and (ii) GSL fingerspelling, including GSL alphabet units and continuous sequences of such.

2.2 Platform design

As shown in Fig. 1, the SL-ReDu interactive platform involves theory presentation of appropriate GSL learning material through videos and images, while enabling both practice and testing via passive-type assignments comprising ordinary multiple-choice questions relying on images, videos, and text to elicit a response from the user, as well as active-type exercises based on video recordings of user GSL productions. In the latter case, the integrated SLR technology provides a binary correctness assessment of the articulated signs, assuring an objective, fast, and reliable testing process.

2.3 Platform architecture

The SL-ReDu prototype system comprises a web-based learning and assessment application that runs on a web server that is responsible for the end-user interaction. Our web-based application architecture integrates two basic components, the front-end user interface and the back-end data-server, with the latter involving the system

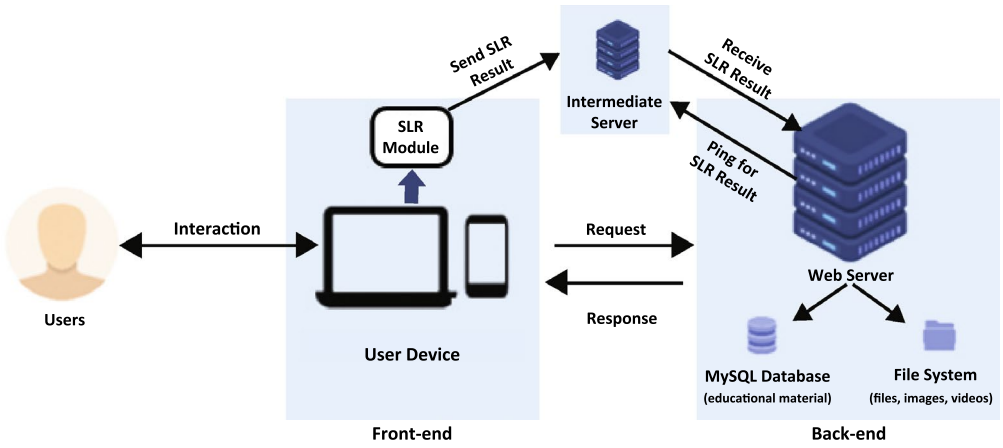

Fig. 2 Illustration of the SL-ReDu prototype system web-based architecture (Figure modified from [14])

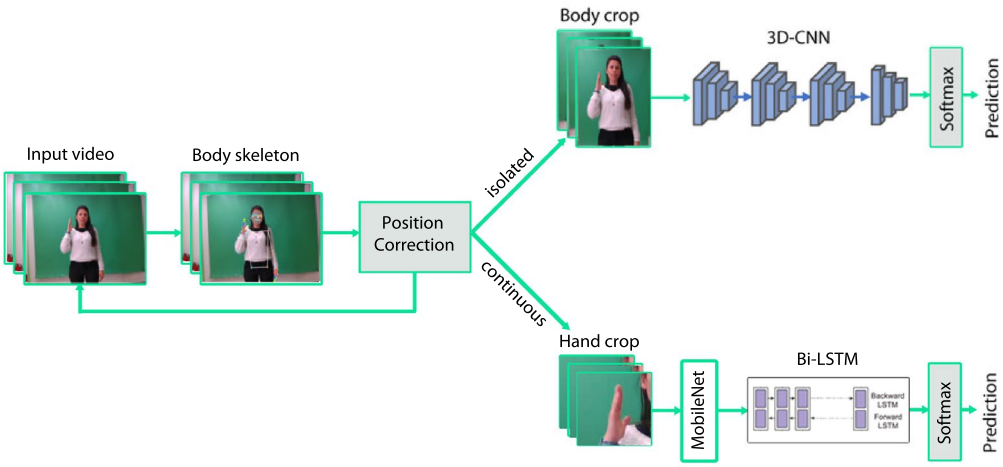

Fig. 3 Algorithmic flow-chart of the GSL recognition module of the SL-ReDu platform: It first employs the MediaPipe framework [21], estimating the signer's 3D pose to ensure correct positioning with respect to the camera and prompt for signer repositioning if necessary. Subsequently, the signer's video is further processed for region-of-interest extraction, and it is then fed to a suitable recognition model depending on the GSL recognition task (isolated signs or continuous fingerspelling)

database and the file storage system that incorporates images and videos corresponding to the educational material. In addition, the system incorporates an administration framework that can be leveraged by the instructor to generate assessment tests. The dynamic web platform is developed in the PHP programming language, in conjunction with HTML5, CSS3, and JavaScript, while the MySQL open-source database is employed for database construction. The web-based application is hosted in an Apache web server. Figure 2 depicts the employed architecture.

The SLR component of the system constitutes a separate module that runs on the learner-side device. Such device is currently a laptop with an available web-cam that records the GSL production, as well as computation acceleration by a graphics processing unit (GPU) to speed up GSL recognition of the recorded video due to the use of computationally intensive deep-learning SLR models.[1] The recognition results are sent from such device to the server that runs the web application, based on a protocol that involves an intermediate server, functioning as a "communication repository" between the two. Details of this process can be found in a related technical report of the SL-ReDu project [30]. Note that, in future, we plan to migrate the SLR module to a suitable server with GPU acceleration, allowing the use of low-end devices on the learner side.

[1] The currently used laptop is rather old (6-years) with only 12 GB RAM and a lower-end NVidia GeForce GTX 1050 Ti GPU with 4 GB of memory. Note that in the absence of the SLR module (i.e. system operation for GSL receptive exercises only), user interaction is possible via even lower-end computers or smartphones.

3 The GSL recognition module

We next detail the GSL recognition module adopted in the SL-ReDu prototype system, also schematically illustrated in Fig. 3. In addition to its depicted components, a number of alternative algorithms are also considered in this paper and evaluated in Sect. 4.

3.1 Pre-processing

The SL-ReDu recognition module commences with the detection of the signer in order to inspect the relative position with respect to the field-of-view of the device camera, providing instructions for rectification in case of wrong positioning (e.g., occluded manual articulators). For signer detection and tracking, we employ the MediaPipe holistic framework [21], which is a multi-stage pipeline that takes as input an image frame and returns 543 whole-body landmarks, namely 33 body pose keypoints, 21 joints for each hand, as well as 468 facial keypoints.

When the articulators are occluded, namely the detected landmarks of the two hands, face, and upper torso are missing for more than a specific number of frames (12 frames), the system prompts the signer to reposition, informing of incorrect hands, face, and body positioning with respect to the camera. If the positioning is acceptable, the detected landmarks are used to generate the region-of-interest (RoI) for the next stage, i.e. the visual representation generation (see also Fig. 3). Particularly, in the case of isolated signs, when multiple articulators may be involved in signing, the entire upper body is cropped in order to produce the RoI. In the case of fingerspelling though, the RoI is limited to the signing hand (as determined by its motion). In more detail, the x and y landmarks of the whole-body or hands with values ranging in [0.0,1.0] are normalized to the image plane based on the image width and height respectively and stored in a list. Afterward, the maximum and minimum values of the corresponding x and y landmarks of the list are employed for RoI cropping.

3.2 Feature extraction

In the case of isolated signing, our GSL recognition module [14] adopts a 3D-CNN model for spatio-temporal feature extraction, namely the ResNet2+1D network [17], while in the case of continuous fingerspelling it relies on the MobileNet 2D-CNN-based feature learner [18] combined with a BiLSTM encoder [19]. Here, we demonstrate the superiority of our SLR models by investigating a number of alternative approaches for processing the input video and extracting visual spatio-temporal features, focusing on articulator shape or appearance and static or motion patterns. Specifically, we first investigate 2D and 3D skeletal features, as well as 3D expressive body pose and shape features. Further, we explore deep learning-based appearance representations, considering various 2D- and 3D-CNN image feature learners. Finally, we investigate motion representations via different optical flow models. All these different feature extraction models are detailed next.

3.2.1 Pose features in 2D and 3D

Here, we exploit the HRNet framework [20] to detect the human skeletal joints of the signer's body, face, and hands in 2D. In addition, we investigate Google's MediaPipe holistic framework [21] for simultaneous perception of 3D human body pose, face and hand landmarks, yielding the corresponding 3D human skeletal feature representations. Moreover, we obtain 3D shape, pose, and facial expression features of the signer from single RGB images using the ExPose method [22]. Further details follow next.

HRNet [20]. The HRNet full-body pose estimator relies on a high-resolution CNN that preserves the high-resolution representation by combining high- and low-resolution convolutions in parallel and repeating multiscale fusion across parallel convolutions. The adopted parallel processing strategy allows HRNet to maintain high resolution throughout the neural network, resulting in more accurate representations. HRNet estimates in total 133 2D whole-body landmarks from the RGB images, namely 23 body-pose skeletal joints, 21 joints for each hand, as well as 68 facial keypoints (see also Fig. 4b). Since manual and non-manual SL articulation occurs around the signer's upper-body, we disregard 10 lower body joints, as well as 68 facial keypoints of the 133 skeletal joints returned by HRNet. Specifically, in the isolated sign recognition task we employ 13 body-pose skeletal joints and 21 joints for each hand, all provided in 2D coordinates on the image plane, generating a 110-dimensional (dim) feature vector, while in the fingerspelling task we obtain a 42-dim (21×2) feature vector.

MediaPipe [21]. The MediaPipe holistic pipeline integrates separate models of the pose, face, and hand components, optimized for their respective domains. MediaPipe holistic estimates 543 3D whole human body joints (see also Fig. 4c). Specifically, 33 3D joints of these correspond to the body skeleton, 21 to the 3D joints of each hand, while the remaining 468 correspond to the 3D coordinates of the facial region. Note that x and y landmark coordinates are normalized to the [0.0,1.0] range based on the image width and height respectively, while the magnitude of the depth landmark, z, uses the same scale as x with the midpoint of the hips being the origin. Here we exploit only 67 of the 543 MediaPipe 3D coordinates, removing 8 human body joints associated with the invisible lower body parts of the signer, as well as the 468 facial landmarks. In particular, we retain

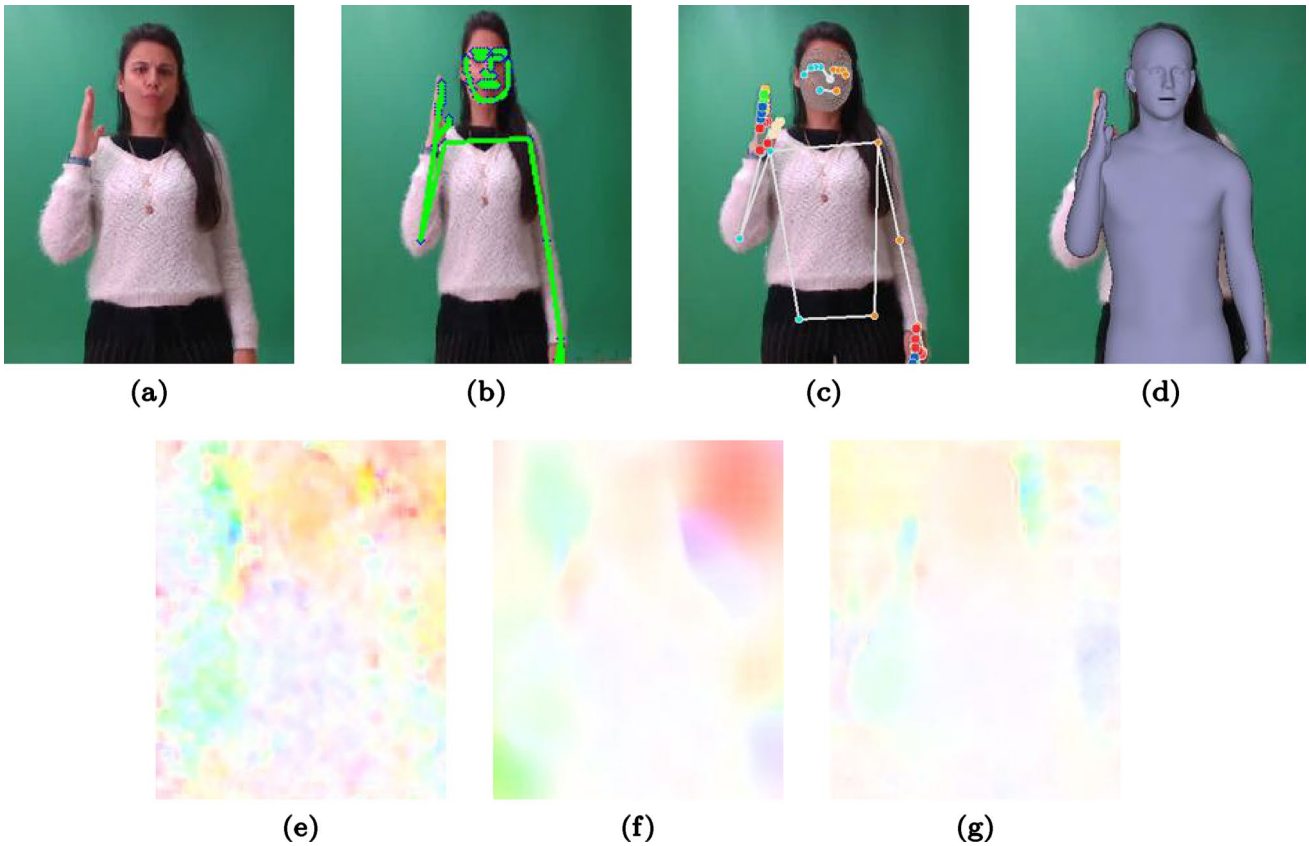

Fig. 4 Various feature extraction components considered in this paper: **a** Sample frame from the ITI GSL corpus [31], **b** 2D human pose regression via HRNet [20], **c** 3D human skeleton representation using the MediaPipe holistic pipeline [21], **d** 3D human body shape, pose, hand articulation, and facial expression representation via the ExPose regression model [22]; and optical flow motion informative images generated by **e** the SpyNet [26], **f** the FlowNet [27], and **g** the PWC-Net [28] optical flow models

25 body-pose and 42 hand-pose ones, all provided in 3D coordinates on the image plane, thus resulting in a 201-dim feature vector for the isolated sign recognition task and a 63-dim (21×3) vector for continuous fingerspelling.

ExPose [22]. ExPose obtains expressive 3D body pose and shape of the signer, operating directly on image pixels (see also Fig. 4d). The ExPose framework extracts the 3D joint rotation parameterization generated using 3D whole body reconstruction. In particular, both shape and expression are described by 10 coefficients derived from the principal component analysis space, while the body pose includes 53 joints with 6 degrees of freedom, i.e. 22 body-pose joints, 15 joints per hand, and 1 for the jaw, yielding a 338-dim feature vector for the isolated recognition task. In the case of fingerspelling, we employ only the 15 joints with 6 degrees of freedom of the signing hand, resulting in a 90-dim feature vector.

Note that in case of 2D and 3D skeletal joint estimation failure, the missing features are substituted by the previous existing ones. Further, to achieve translation and scale invariance, we normalize all extracted human skeletal joints derived from the HRNet and MediaPipe frameworks by converting the image to a local coordinate system with the neck keypoint being the origin, and apply further normalization based on the distance between the left and right shoulder keypoints.

3.2.2 Appearance features

Further, we adopt deep learning-based appearance representations extracted from the entire RoI, as generated by the pre-processing phase (see also Sect. 3.1). Specifically, we first investigate three 2D-CNN based neural networks, namely the ResNeXt-101 [23], the MobileNet [18] that is employed for the continuous fingerspelling recognition task, and the Inception-V3 [32]. In addition, we consider three 3D-CNN models in our experiments, i.e., the ResNet2+1D [17], which is also incorporated in our isolated SLR module, the Pseudo-3D Residual Network (P3D) [24], as well as the 3D ConvNet (C3D) [25]. Note that their 3D convolutions extract both spatial and temporal components relative to signing motion, thus not being limited to purely appearance representation as in the case

of 2D-CNNs. More details of these feature learning models follow next.

ResNeXt-101 [23]. ResNeXt-101 is a neural network that requires less hyperparameters than a traditional ResNet. This is accomplished through the use of cardinality, which is an additional dimension on top of the width and depth of ResNets. The network is pre-trained using the ImageNet corpus [32], which contains around $1k$ tagged images for each of $1k$ categories. The input RoI to the network must be rescaled to 224×224 pixels. Feature maps are constructed by using the output of the last global average pooling layer to produce 2048-dim representations. A linear layer is employed to reduce dimensionality, resulting in a 512-dim feature vector.

MobileNet [18]. This is a CNN-based architecture that is based on an inverted residual structure with residual connections between bottleneck layers. The MobileNet architecture consists of an initial fully convolutional layer with 32 filters, followed by 19 bottleneck layers. The model is pre-trained on the ImageNet database [32] and requires the input image to be rescaled to 224×224 pixels. The model yields 1024-dim output that, following a linear layer, results to a 512-dim feature vector.

Inception-V3 [32]. This network is a 48-layer deep CNN pre-trained on ImageNet [32], and it requires data rescaling to 299×299 pixels. Feature maps are generated by taking the output of the last global average pooling layer, yielding 2048-dim representations. To reduce dimensionality, a linear layer is used, resulting in 512-dim features.

ResNet2+1D [17]. The ResNet2+1D feature learner is a 18-layer model that separates 3D convolutions into spatial 2D convolutions followed by temporal 1D convolutions. The ResNet2+1D network comprises five (2+1)D convolutional blocks, with each composed of one spatial and one temporal convolution, and a 3D average pooling layer that operates on both space and time for nonlinear downsampling of the output tensor. The model weights are pretrained on the Kinetics dataset [33]. The network uses 16 frames with size 112×112 as input clips and yields a 512-dim feature vector.

Pseudo-3D Residual Network (P3D) [24]. In this network, 3D convolutions are decoupled into $1 \times K \times K$ convolutional filters on the spatial domain and $t \times 1 \times 1$ convolutions tailored to the temporal domain. Specifically, the architecture consists of 199 layers involving different variants of pseudo-3D convolution blocks. In the first one, temporal 1D convolutional filters follow spatial 2D convolutional filters in a cascaded manner, while in the second one both filter types operate in a parallel fashion. The network uses a 16-frame clip and 160×160 input resolution, yielding 2048-dim representations. To reduce dimensionality, a linear layer is used, resulting in 512-dim features.

3D ConvNet (C3D) [25]. C3D is a deep 3D-CNN with an homogenous architecture containing 8 convolutional and 5 pooling layers, followed by 2 fully connected layers. The C3D model is given an input video segment of 16 frames and 112×112 input resolution and results in a 4096-dim feature vector. For dimensionality reduction, a linear layer is used, resulting in 512-dim features.

3.2.3 Optical flow features

In addition to the above, we also adopt optical flow features, which play a crucial role in SLR. To acquire them we employ three approaches, namely the SpyNet [26], the FlowNet2 [27], and the PWC-Net [28] that extract motion information of the RoI using deep-learning models. More details follow next.

SpyNet [26]. The optical flow is computed by combining an image-pyramid formulation with deep learning. This optical flow method is based on warping the second image of a pair of image frames at each pyramid level using the current flow estimate and producing an optical flow update. At each pyramid level, one deep neural network is trained in order to estimate the flow that is upsampled to the next pyramid level. SpyNet is "lean" in terms of model parameters with 1.2M in total.

FlowNet2 [27]. FlowNet2 adopts a stacked architecture involving different types of networks in order to compute both large and small displacements of optical flow. More precisely, large displacement of optical flow is computed by stacking two adjacent image frames as input to a network and warping the second image toward the first in the pair using the current flow, while in a second network the two image frames are separately convoluted for small displacement optical flow estimation. Finally, the outputs are fused by a correlation layer. The FlowNet model is "heavy," having almost 160 M parameters.

PWC-Net [28]. This method follows three main principles: pyramidal processing, warping, and the use of a cost volume. Initially, it warps the CNN features of the second image of the pair, employing the current optical flow estimate. Subsequently, a cost volume is generated, using the warped features of the second image, as well as the features of the first one. Finally, the cost volume is further processed by a CNN for optical flow estimation. PWC-Net is small in size, having 9.7M model parameters.

Once the optical flow is estimated using any of the aforementioned models, it is scaled to the input image size and stored as a 2-band float image for both horizontal and vertical flow components in an optical flow file. Subsequently, motion informative images are generated by colorizing the displacement vectors acquired by the optical flow files (see also Figs. 4(e)-(g)). It should be noted that the optical flow informative image frames are then fed to a 2D-CNN MobileNet [18] image feature learner, yielding 512-dim motion features.

3.3 SL recognition

As mentioned in Sect. 1, the SLR module accommodates two recognition tasks, namely that of isolated GSL signs within a medium-size vocabulary, developing separate models for numerals and non-numerals, and that of continuous sequences of fingerspelled letters of the Greek alphabet. Specifically, for the isolated SLR task we employ the ResNet2+1D network, whereas for the continuous fingerspelling recognition task the 2D-CNN MobileNet architecture is adopted as the visual feature learner of each video frame, as well as a BiLSTM encoder [19]. Next we describe the two SLR modes and we provide their implementation details.

3.3.1 Isolated sign recognition

For isolated GSL recognition, the SL-ReDu prototype system [14] employs the ResNet2+1D network for spatio-temporal visual feature extraction (see also Fig. 3). As discussed in Sect. 3.2.2, ResNet2+1D is a 18-layer model that includes five (2+1)D convolutional blocks coupled with a 3D pooling layer that is employed for nonlinear downsampling of the output tensor in both spatial and temporal dimension. The pooling layer is followed by the classifier, i.e. a fully connected layer coupled with a softmax layer, which produces the desired probability scores. For label prediction, the cross-entropy loss function is employed with label smoothing [34]. Model training (fine-tuning) is carried out via the Adam optimizer [35] with initial learning rate set to 0.0001 and weight decay 0.0001. The mini-batch size is fixed to 16.

We compare the above with a multitude of networks described in Sect. 3.2 employed for visual stream extraction, namely the 2D and 3D skeletal joints, the 3D body pose and shape information, as well as the appearance- and motion-based representations, with a BiLSTM encoder [19] for spatio-temporal feature modeling and the classifier. Note that, in the case of 3D-CNN feature learners, which allow sign classification from the spatial and temporal encoded information of RGB sequences, we abstain from using an additional encoder in an attempt to retain a light-weight SLR model for the SL-ReDu prototype system. The models are trained employing a dropout rate of 0.1 with a mini-batch size fixed to 32. All models are implemented in PyTorch [36], and experiments are carried out using GPU acceleration for both training and evaluation.

It should also be noted that separate models are built for the recognition of isolated numeral signs (with a vocabulary size of 18) and the recognition of isolated non-numeral signs (with a vocabulary size of 36). As already stated in Sect. 1, in the first phase of SL-ReDu evaluation the isolated SLR module serves for a small-size vocabulary, while in the second phase of the project we plan to extend the sign vocabulary to about 400 (numerals and non-numerals).

3.3.2 Continuous fingerspelling recognition

For continuous GSL fingerspelling recognition, the SL-ReDu prototype system [14] employs the 2D-CNN MobileNet architecture [18], serving as visual feature learner of each video frame, and a BiLSTM encoder [19], which learns their temporal relations (see also Fig. 3). The output feature maps are propagated to a last fully-connected layer followed by a softmax, yielding the probabilities distribution for aligning the signing videos to letter sequences, modeled via the connectionist temporal classification (CTC) decoding model [37]. We also add a label smoothing term equal to 0.2, in order to penalize low-entropy distributions. Specifically, a two-layer BiLSTM encoder is employed with 512-dim hidden states, followed by CTC decoding for letter sequence prediction.

In this work, we compare this approach with recent state-of-the-art techniques in SLR. Specifically, we substitute the MobileNet architecture with a number of additional networks described in Sect. 3.2. To this end, we exploit 2D and 3D skeletal representations, and we also extract expressive 3D pose and shape information of the signer. Further, we adopt deep learning-based appearance and motion representations of the signing activity. Apart from the BiLSTM encoder, we also investigate the contribution of a bidirectional-GRU (BiGRU) encoder [29]. In particular, the model constitutes a 2-layer BiGRU encoder with 512 hidden units.

All aforementioned models are trained employing a dropout rate of 0.1 with a mini-batch size fixed to 16. During training the Adam optimizer is used with an initial learning rate of 0.001 decreased by a factor of 0.1, if the validation score remains consistent for 9 steps. During inference, the beam search strategy is employed with a beam width of 3. Note also that no letter language model is employed. For model implementation, PyTorch [36] is employed, and all experiments are performed on GPU acceleration for model training and inference.

4 GSL data and experiments

To accomplish the development and evaluation of the GSL recognizer, a suitable database has been collected and presented in [14]. We describe it next, followed by the adopted experimental framework and our GSL recognition experiments on it.

4.1 The GSL database

To enable isolated GSL recognition of numerals (18-sign vocabulary) and isolated SLR of non-numerals (36-sign GSL vocabulary), as well as continuous recognition of fingerspelled sequences of the 24 Greek alphabet letters, signing data from numerous volunteer informants (both native and non-native in GSL) have been collected, as presented in [14]. The data were captured by a Logitech C615 webcam indoors, under realistic, non-studio conditions with varying background and lighting, at a frame rate of 30 Hz, in YUV411 video format, and a resolution of 640×480 pixels.

Regarding numeral signs, video data from 20 signers have been gathered (see also Fig. 5). Each signer articulated the 18 numerals 5 times in a row, yielding a total of 1,800 video snippets. In the case of non-numeral signs, data from 17 signers have been collected (see again Fig. 5). The 36 signs were articulated 5 times by each informant. Further, videos from the publicly accessible ITI GSL corpus [31] have been added to these data, resulting in 7 more informants signing the same set of 36 signs 5 times (see also Fig. 6). It is

Fig. 5 Sample video frames from the non-studio data collected for isolated GSL recognition of numerals and non-numerals

Fig. 6 Example video frames from the publicly accessible ITI GSL corpus [31]

important to note that the latter were recorded in a studio environment with an Intel RealSense D435 RGB-D camera, but just the RGB stream is used here. Thus, there are a total of 24 signers (17 + 7) and 4,320 videos in the combined dataset.

Finally, video data from 12 informants has been collected while fingerspelling (see Fig. 7). Specifically, all 24 Greek alphabet letters were signed by each signer once, as well as another 50 fingerspelled words (unique to each signer) composed of 4–5 letters. In addition, 3 signers performed an additional 71 words with a letter sequence length of 4 or 5, as well as 16 words with a letter sequence length of 3–7. Note that each informant has signed each letter at least 4 times. To summarize, 1071 videos have been collected in this process.

4.2 Experimental framework

Since SL-ReDu platform learners are often "unseen" during GSL model training, signer-independent (SI) SLR is of particular interest to us. We also report multi-signer (MS) recognition results for comparison; in this scenario, data from all signers are used in both training and test sets (with the sets remaining disjoint).

For the MS case, we adopt ten-fold cross-validation, with 80% of all videos used for training (1440 for numerals, 3456 for non-numerals, and 857 for fingerspelling), 10% for validation (180 for numerals, 432 for non-numerals, and 107 for fingerspelling), and the remaining 10% for testing (same number of videos as in validation).

For the SI scenario, we use cross-validation with 20-folds in the numerals case, 24 folds in the non-numerals case, and 12 ones for fingerspelling. Each fold comprises one "test" signer, while the rest are used to train the model.

In addition to these paradigms, our GSL recognition models are trained for use by the SL-ReDu platform in its user-evaluation. To that end, 1620 numeral videos, 3888 non-numeral videos, and 964 fingerspelling videos are employed in training, while the remaining 1000 are used for validation (numerals: 180; non-numerals: 432; fingerspelling: 107).

4.3 Recognition results

For the task of isolated GSL recognition, we evaluate the performance of the various networks described in Sect. 3, reporting experimental results on the isolated GSL datasets of Sect. 4.1 under both MS and SI experimental paradigms. Results are reported in word accuracy (WAcc), %.

Fig. 7 Sample video frames from the non-studio fingerspelling data

In our first experiments, reported in Table 1, we employ the feature representations of Sect. 3.2 individually, namely the 2D human skeletal features, the 3D human skeleton representation, the 3D human pose and shape parameterization, the 2D-CNN based appearance representation, and the optical flow motion features. These streams are fed to a BiLSTM encoder and a cross-entropy based decoder for word prediction. Note that the BiLSTM inclusion does not seem to help our 3D-CNN models, and thus is not integrated to our SLR pipeline. For example, we evaluated the performance of the proposed R(2+1)D model coupled with a 2-layer BiLSTM encoder with 512-dimensional hidden units, and the MS WAcc on non-numerals degraded from 99.44 to 97.89%.

In Table 1, it can be observed that the 3D human pose and shape feature stream yields the best results on all skeleton-based sequence learning models among all tasks and experimental paradigms, showcasing the robustness of 3D pose and shape parameterization. Further, the ResXt-101 performs well, but our 3D-CNN model turns out superior to the considered alternatives in both isolated SLR tasks and experimental paradigms. Moreover, optical flow representations seem to constitute a more powerful

Table 1 Word accuracy (%) of isolated SLR (of numeral and non-numeral signs) on the corresponding datasets of Sect. 4.1, under both MS and SI training/testing paradigms, employing various networks of Sect. 3. Model size (approximate parameters in millions) and inference time (in seconds) are also shown.

Visual representations		SLR task ⟶	Numerals		Non-numerals			
		Models ↓	MS	SI	MS	SI	Size	Time
Pose	2D skel	HRNet/BiLSTM	93.33	86.67	91.67	78.33	32 M	3.54
	3D skel	MediaPipe/BiLSTM	94.44	88.89	92.13	81.67	11 M	2.89
	3D pose/shape	ExPose/BiLSTM	95.56	90.00	93.75	86.11	11 M	12.25
RGB	2D-CNNs	ResXt-101/BiLSTM	96.67	93.33	97.45	88.40	56 M	1.14
		MobileNet/BiLSTM	94.44	91.11	95.14	86.67	15 M	1.01
		Inception/BiLSTM	95.56	92.22	96.06	87.78	38 M	1.09
	3D-CNNs	P3D	91.11	87.78	97.45	73.28	66 M	0.55
		C3D	94.44	90.00	98.38	81.64	78 M	0.62
		R(2+1)D	**97.78**	**94.48**	**99.44**	**96.20**	42 M	0.30
Flow	2D-CNNs	SpyNet/BiLSTM	95.56	91.11	92.36	84.44	12 M	2.28
		FlowNet/BiLSTM	94.44	90.00	91.20	82.22	50 M	3.29
		PWC-Net/BiLSTM	96.67	92.22	94.44	85.56	20 M	2.35

The proposed R(2+1)D model is in bold

Table 2 Word accuracy (%) of continuous fingerspelling on the dataset of Sect. 4.1, under both MS and SI training/testing paradigms, employing various networks of Sect. 3. Model size (approximate parameters in millions) and inference time (in seconds) are also shown.

Visual representations		Models ⟶	(&BiLSTM)				(&BiGRU)			
		↓	MS	SI	Size	Time	MS	SI	Size	Time
Pose	2D skel	HRNet	62.62	42.22	32 M	3.54	58.22	41.07	29 M	3.53
	3D skel	MediaPipe	64.49	45.56	11 M	2.89	59.53	43.76	8 M	2.89
	3D pose/shape	ExPose	64.77	50.28	11 M	12.25	62.43	48.87	8 M	12.25
RGB	2D-CNNs	ResXt-101	73.29	62.22	56 M	1.14	71.96	62.80	53 M	1.13
		MobileNet	**75.22**	**65.30**	15 M	1.01	73.83	60.00	12 M	1.01
		Inception	74.76	64.86	38 M	1.09	72.90	61.98	36 M	1.08
	3D-CNNs	P3D	70.09	37.84	77 M	0.56	65.79	36.20	74 M	0.56
		C3D	71.96	61.11	89 M	0.64	70.28	61.07	87 M	0.63
		R(2+1)D	73.83	63.33	54 M	0.31	70.56	60.88	51 M	0.30
Flow	2D-CNNs	SpyNet	71.02	62.16	12 M	2.28	70.65	60.45	10 M	2.27
		FlowNet	68.22	59.46	50 M	3.29	66.82	57.52	47 M	3.29
		PWC-Net	72.89	63.51	20 M	2.35	71.96	61.99	17 M	2.34

The proposed SLR model is marked in bold (MobileNet & BiLSTM)

representation than the skeletal features and in some cases outperform the 2D- or 3D-CNNs. Note also that performance is worse in the SI case as compared to the MS one, which can also be readily observed in Figs. 8a, b, where WAcc for the isolated tasks in the SI case is lower but remains nevertheless well above 80%, even for the worse performing ones, demonstrating the potential of utilizing our model in learning platforms like SL-ReDu. Finally, it can be observed that the proposed model is the fastest model among the considered alternatives,[2] albeit at a considerable increase in model size as compared to leaner ones.

In Table 2, we evaluate the performance of the various sequence learning models described in Sect. 3 and report the recognition performance of the continuous fingerspelling task on the aforementioned dataset (see Sect. 4.1), under both MS and SI training/testing paradigms. Results are reported in WAcc, %, taking into account the number of substitutions, deletions and insertions in the predicted hypotheses. It may be observed that the BiLSTM encoder-based recognition model turns out superior to the considered alternatives in terms of WAcc, revealing the power of exploiting BiLSTMs in continuous SLR. It is obvious that our model (MobileNet and BiLSTM) outperforms all considered alternatives. Moreover, relying on skeletal representations performs the worse. Surprisingly, 3D-CNN feature representation do not seem important contributors to the system performance. In all cases, performance degrades in the SI case, compared to the MS scenario, which is not surprising. It should be noted that the R(2+1)D with a BiGRU encoder is the fastest model (0.30 s) among the considered alternatives,[3] but we employ the MobileNet feature learner with BiLSTM encoder as our recognizer, since it provides a trade-off between accuracy and speed.

Finally, in Fig. 8c the WAcc, % and letter accuracy (LAcc), % for the fingerspelling task in both MS and SI cases is shown. The performance suffers at the WAcc level, which is natural since letter recognition errors (including insertions and deletions) accumulate at the word level, especially for longer letter sequences. This effect is exacerbated due the lack of a language model in the recognizer, as well as the significantly smaller amount of collected data and number of signers compared to the isolated tasks. As expected, LAcc results are higher, but clearly further improvement is needed.

5 User evaluation of the SL-ReDu platform

Next, we proceed with the user-based system evaluation during a campaign conducted at UTH-SED, reporting both the GSL recognition results obtained (objective evaluation), as well as the subjective assessment of the SL-ReDu prototype as a whole, based on user responses to a questionnaire. Before reporting our results, we briefly describe the pool of the evaluation participants.

5.1 Volunteer users

The evaluation campaign involved two student groups from UTH-SED, as well as two professional volunteers. A total

[2] Reported times in Table 1 refer to model inference on a 30 Hz video of length 5 s, as run on an NVidia GeForce RTX3090 GPU, thus all models are faster than real time, except the ExPose-based one. It should be noted that the older laptop used in the evaluation of Sect. 5 is about 8.9× slower, but still acceptable to users in terms of speed.

[3] Reported inference times in Table 2 are reported for a 30 Hz, 5 s long fingerspelling video.

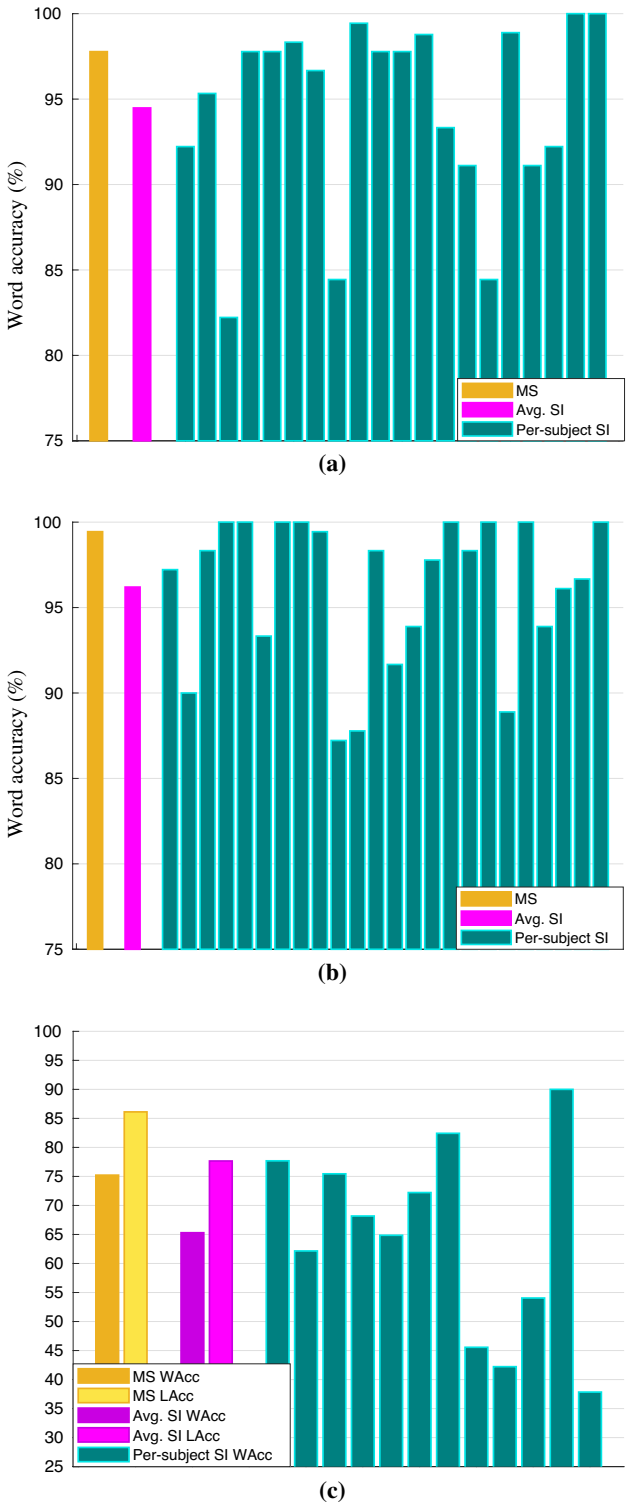

Fig. 8 MS, average SI, and per-subject SI isolated word recognition accuracy (WAcc, %) for **a** numerals; **b** non-numerals; and **c** fi ngerspelling signs. In the case of fingerspelling MS and average SI letter recognition accuracy (LAcc, %) is also included

of 10 university students at the so-called "A0" level who had been exposed to GSL for less than five months made up the first user group, whereas a total of 11 students at the "A1" level with more than five months of experience were part of the second group. The third group consisted of the two GSL professionals (experts), involved in GSL instruction. The volunteer demographics were in line with those of the student/instructor population at UTH-SED, with females outnumbering males (21 females and 2 males) and ages ranging from 19 to 22 years old for the undergraduates, a 35-year old graduate student, and two experts in the 40-45 year old range. Note that all volunteers had signed consent forms prior to the evaluation.

5.2 Objective evaluation of the GSL recognizer

Objective evaluation was conducted by means of active signing, where learner SL productions were captured by a camera of a specially equipped laptop at the user side (see also Fig. 2) and were subsequently recognized by the on-laptop SLR module, providing feedback to the learner through the SL-ReDu platform, as well as feedback in case of incorrect placement with respect to the camera. A small number of tests were made available for each recognition task for the evaluation participants to choose from, each including few production exercises. In particular, for isolated GSL recognition of numerals three tests with six GSL production questions each were integrated in the platform, while for non-numerals six corresponding six-question production assignments were incorporated. In addition, continuous fingerspelling tests were also available, namely six six-question exercises that included letters in addition to words (the latter were absent from the fingerspelling training set of Sect. 4.1). The participants were permitted to perform each exercise up to twice (the second time in case of negative feedback by the system) within the time duration constraints of the selected tests (slightly different per task), and the system automatically graded their efforts providing the cumulative test scores.

A subset of the volunteers of Sect. 5.1 was used for the active GSL production and recognition evaluation. This included a total of 12 users (all females), with 7 users being students at the "A0" level (referred to as group "G1"), as well as 4 of the "A1"-level students and 1 expert (referred to as group "G2"). Each subject completed three six-question assignments, one for each of the three GSL tasks stated above, performing 18 exercises (216 in total for the

Table 3 Scores achieved on the six-question GSL production assignments by the 12 volunteers in the objective SL-ReDu system evaluation of SLR

Task	Isolated numerals		Isolated non-numerals		Fingerspelling	
User id	Selected test	Score	Selected test	Score	Selected test	Score
1	test5.php	6 / 6	testSignRecog1.php	6 / 6	activefstest1.php	**5** / 6
2	test6.php	6 / 6	testSignRecog2.php	6 / 6	activefstest2.php	6 / 6
3	test7.php	6 / 6	testSignRecog3.php	**5** / 6	activefstest3.php	**5** / 6
4	test5.php	6 / 6	testSignRecog4.php	**5** / 6	activefstest4.php	**5** / 6
5	test6.php	**5** / 6	testSignRecog5.php	6 / 6	activefstest5.php	6 / 6
6	test7.php	6 / 6	testSignRecog6.php	6 / 6	activefstest6.php	6 / 6
7	test5.php	6 / 6	testSignRecog1.php	6 / 6	activefstest1.php	**5** / 6
8	test6.php	6 / 6	testSignRecog2.php	6 / 6	activefstest2.php	6 / 6
9	test7.php	6 / 6	testSignRecog3.php	6 / 6	activefstest3.php	6 / 6
10	test5.php	6 / 6	testSignRecog4.php	6 / 6	activefstest4.php	6 / 6
11	test6.php	6 / 6	testSignRecog5.php	6 / 6	activefstest5.php	6 / 6
12	test7.php	6 / 6	testSignRecog6.php	6 / 6	activefstest6.php	**3** / 6

12 volunteers). Table 3 lists specifics of the scores they attained (incorrect recognitions by the GSL recognizer are shown in bold). Further, Table 4 provides a cumulative and task-specific summary of the results (over all users). Bold emphasizes the correct signings that have been recognized as incorrect by the GSL recognizer on all 3 tasks during the objective SL-ReDu system evaluation of SLR.

As it can be observed in Table 4, out of the 216 exercises in total, 200 correct GSL productions were determined by the system to be valid and 16 to be invalid at the first user attempt. The latter 16 exercises include 10 correct GSL productions that were wrongly recognized as incorrect by the system and 6 incorrect GSL productions that were accurately determined as incorrect by the system. In the second user attempt, all 16 exercises were correctly signed, but the system inaccurately recognized 10 of those as incorrect. These included 1 numeral, 2 non-numerals, and 7 fingerspelling exercises (see also Table 3), demonstrating that the last task is the most challenging due to its continuous nature. Note that there is a very small number of instances in which the user signed correctly but the system incorrectly identified the sign. It is interesting that this did not occur in the opposite scenario, where an inaccurate user signing was recognized by the system as correct. This is primary due to the fact that the incorrect signings happened to be very different to valid ones, i.e. the signings that were employed to train the GSL recognition model.

Further, in Table 5, we compare the above results between the less experienced (group "G1") and more experienced GSL volunteers (group "G2") defined earlier, reporting SLR results in terms of Wacc (all tasks) and LAcc (fingerspelling only) with best scores in each task being shown in bold. We observe that the results of "G2" users are clearly better than those obtained by "G1" volunteers. This is likely due to the more careful and clearer signing performed by the more experienced GSL users. This difference becomes even larger in the fingerspelling task, due to the additional fact that the corresponding exercises require the production of continuous sign sequences.

Table 4 A summary of the objective SL-ReDu system evaluation of SLR, cumulatively presented for all GSL recognition tasks and each task separately for both first and second (if required) sign production attempt

Attempts		First attempt		Second attempt	
(#216 first, #16 s)		System response		System response	
Task	User sign	Correct	Incorrect	Correct	Incorrect
All 3 tasks	Correct	200	**10**	6	**10**
	Incorrect	0	6	0	0
Isolated Numerals	Correct	69	**1**	2	**1**
	Incorrect	0	2	0	0
Isolated Non-numerals	Correct	69	**2**	1	**2**
	Incorrect	0	1	0	0
Continuous Fingerspelling	Correct	62	3	3	**7**
	Incorrect	0	**7**	0	0

Table 5 GSL recognition performance for all three SLR tasks during the SL-ReDu system objective evaluation, reported for the two user groups ("G1"-less experienced, "G2"-more experienced) and overall. Results are shown in word accuracy (WAcc, %) for all tasks, as well as letter accuracy (LAcc, %) for fingerspelling

GSL recognition task	Metric	User groups		All users
		"G1"	"G2"	
Isolated numerals	WAcc	95.56	**100.00**	97.33
Isolated non-numerals		91.11	**100.00**	94.67
Continuous fingerspelling	WAcc	76.00	**93.75**	82.93
	LAcc	86.92	**91.04**	88.51

Table 6 Median, maximum, and minimum values of the subjective evaluation of the SL-ReDu platform for each of the eight questions over all 23 evaluation participants

Subjective question	Values of answers over all subjects		
	Median	Maximum	Minimum
User-friendly platform	Much	Very much	Much
Adequate time	Very much	Very much	Medium
Difficulty level meets user needs	Much	Very much	Medium
Satisfaction with platform design	Much	Very much	Medium
Feeling safe using platform	Very much	Very much	Medium
Performance improvement	Very much	Very much	Medium
Signing educational support	Much	Very much	Medium
Positive signing experience	Very much	Very much	Much

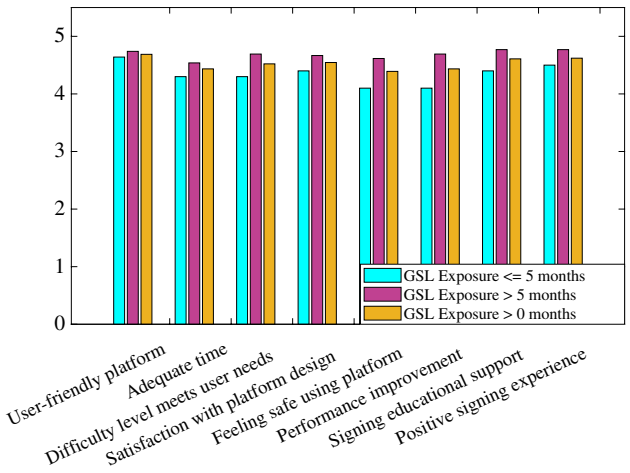

Fig. 9 Mean values (on the 1–5 Likert scale) of the platform subjective user assessment along eight aspects over the "A0"-level ("GSL experience $\leq$ 5 months") and "A1"-level ("GSL experience $>$ 5 months") groups and both ("GSL experience $>$ 0 months")

5.3 Subjective assessment of the platform

Following the completion of the self-monitoring and GSL production sessions on the SL-ReDu platform, participants were given an anonymous subjective experience questionnaire. This questionnaire measures eight aspects related to ease of use, usefulness, design, and user trust on a Likert scale that ranges from one to five.

The analysis of the subjective experience questionnaires that the participants submitted at the conclusion of their evaluation sessions yielded insightful data in the form of both statistical patterns and textual comments. As deduced from Table 6, for four of the eight subjective evaluation questions, the majority of participants gave the best rating ("very much") in their responses. The majority of participants, in particular, expressed complete satisfaction with the amount of time the platform gave them, a sense of safety while using it, a conviction that they will use it to enhance their GSL performance, and an overall positive experience. Additionally, only six questions had one or more "medium" answers, which are the lowest ratings returned. These observations lead to the conclusion that the subjective evaluation findings are very satisfactory.

Table 7 Median values of questionnaire responses to four questions computed separately over the "A0"- and "A1"-level groups of evaluation participants

Subjective question	"A0"-level	"A1"-level
Difficulty level meets user needs	Much	Very much
Using platform for performance improvement	Much	Very much
Signing educational support	Much	Much/very much
Positive signing experience	Much	Very much

In addition, we divide the 23-volunteer evaluation pool into users with GSL experience less than 5 months ("A0" level) and those with more than 5 months GSL exposure ("A1" level). As can be deduced from Fig. 9, in all cases the mean opinion score of both groups remained within the 4-5 range. The fact that there are some ratings that differ between the two groups is interesting to note. Due to extended GSL exposure of the "A1" volunteers, one anticipates that such users will be more confident utilizing the platform. Indeed, as it can be observed, this group provided more positive feedback in all questions, whereas the first group participants ("A0" volunteers with shorter GSL exposure) showed some reluctance and gave less favorable answers to most questions. Further, in Table 7 we provide median values of questionnaire responses to four questions computed separately over the "A0"- and "A1"-level participants. As it can be deduced, responses varied across the two groups: the latter were more confident using the platform and gave the highest positive feedback in questions regarding level of difficulty, performance improvement, and educational support, while "A0"-level students showed a small degree of reservation.

Finally, personal free-text written comments were provided by 16 volunteers at the bottom of the subjective evaluation form. Table 8 provides a list of these remarks. In our future work, the SL-ReDu prototype human-computer interface will benefit from the comments highlighted in bold.

6 Conclusion

In this paper we report our ongoing work on the SL-ReDu GSL education tool, which is built for both self-monitoring and objective evaluation of GSL perception and production. Specifically, we present the SL-ReDu prototype

Table 8 Subjective evaluation comments returned by 16 evaluation participants (in addition to the questionnaire ratings) in the SL-ReDu system evaluation. Comments in bold will lead to improvements in the human-computer interface of the SL-ReDu system

Subjective evaluation comments
"Easy to manage application with a variety of exercises and a very friendly environment."
"I think this is a very pleasant platform that will help students who want more practice and contact with GSL outside the UTH-SED course."
"The platform is very good and gives us the ability to practice and fill any gaps we may have."
"The platform is quite supportive and enables practice."
"It was a pleasant and unique experience."
"Very nice experience; it helps you a lot."
"It was an interesting experience. The platform is useful and easy to use."
"It was a pleasant experience that I would try again."
"Great program, easy to use and simple. **Maybe it would be better to not scroll down to the page bottom to view the exercise response**."
"I liked the platform functionally. **Some points need a better design. The numbering 1/6 is placed on the bottom right, while personally I was looking for it between the navigation arrows where there is the exercise indication**."
"The platform seems pretty easy to use, **however when I stood in front of the camera the system could not detect my face. I made several attempts, but only in 1-2 cases it detected me**."
"**Letter "E" was not distinct enough as rendered in the fingerspelling signing section.** The environment was very friendly. **The time was not clear enough as it was confused with the number of exercises**."
"**Fingerspelled letter** "Π" **needs to be corrected. Exam time is too long. Reduce the exercises from 50 to 25 in the general perception exam**."
"**I would like more colors in the main menu, where we select the sections. Also, when it shows me the correct or incorrect answer, I would like the box to turn green or yellow respectively**."
"**Some videos play in low quality**."
"**"Numbers → Numbers 1–10.000 → Summarization". It is not clear which button to click on. I would have liked exercise descriptions. Images accompanying the exercises are unnecessary. Numbers should be included in the "Thematic Vocabulary" menu tab**."

system, overviewing its built-in interface, linguistic content, and architecture design. Most importantly, we present the GSL recognizer that is integrated to the prototype, and which is capable of recognizing isolated signs within a small vocabulary and continuous fingerspelled letter sequences. In addition, we provide comparative evaluation results of the developed recognition models against state-of-the-art SLR approaches. In particular, the experimental results demonstrate that our recognition module performs well under a signer-independent framework in non-ideal visual settings, outperforming alternative architectures that rely on skeletal, appearance, and motion features. Finally, we present the evaluation campaign of the prototype, discussing results concerning objective assessments of the GSL production aspects of the system, as well as a subjective assessment of the entire platform based on an appropriate questionnaire. The findings of the evaluation can be regarded as very satisfying, validating our approach and the viability of the system design. A larger evaluation campaign is planned in future concerning the next version of the SL-ReDu prototype currently under development, aiming to facilitate significantly richer GSL material and enable continuous GSL production assessment.

Acknowledgements This research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "1st Call for H.F.R.I. Research Projects to support Faculty Members & Researchers and the procurement of high-cost research equipment grant" (Project "SL-ReDu," Project Number HFRI-FM17-2456).

Author Contributions All authors contributed equally.

Funding Open access funding provided by HEAL-Link Greece. This research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "1st Call for H.F.R.I. Research Projects to support Faculty Members & Researchers and the procurement of high-cost research equipment grant" (Project "SL-ReDu," Project Number HFRI-FM17-2456).

Declarations

Conflict of interest There are no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will

need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Armstrong, D.F., Karchmer, M.A., VanCleve, J.V.: The Study of Signed Languages: Essays in Honor of William C. Stokoe. Gallaudet University Press, Washington, DC (2002)
2. Kemp, M.: Why is learning American Sign Language a challenge? Am. Ann. Deaf **143**(3), 255–259 (1998)
3. Haug, T.: Web-based sign language assessment: challenges and innovations. In: ALTE International Conference: Learning and Assessment—Making the Connections (Panel Presentation) (2017)
4. Paludneviciene, R., Hauser, P.C., Daggett, D.J., Kurz, K.B.: Issues and trends in sign language assessment. In: Assessing Literacy in Deaf Individuals: Neurocognitive Measurement and Predictors, pp. 191–207. Springer, New York, NY (2012)
5. Bochner, J.H., Samar, V.J., Hauser, P.C., Garrison, W.M., Searls, J.M., Sanders, C.A.: Validity of the American Sign Language discrimination test. Lang. Test. **33**(4), 473–495 (2016)
6. Hauser, P.C., Paludneviciene, R., Riddle, W., Kurz, K.B., Emmorey, K., Contreras, J.: American Sign Language comprehension test: A tool for sign language researchers. J. Deaf Stud. Deaf Educ. **21**(1), 64–69 (2016)
7. Aran, O., Ari, I., Akarun, L., Sankur, B., Benoit, A., Caplier, A., Campr, P., Carrillo, A.H., Fanard, F.-X.: SignTutor: an interactive system for sign language tutoring. IEEE Multimed. **16**(1), 81–93 (2009)
8. Zafrulla, Z., Brashear, H., Presti, P., Hamilton, H., Starner, T.: CopyCat: an American Sign Language game for deaf children. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG) (2011)
9. Ellis, K., Fisher, J., Willoughby, L., Barca, J.C.: A design science exploration of a visual-spatial learning system with feedback. In: Proceedings of the Australasian Conference on Information Systems (ACIS), pp. 1–13 (2015)
10. Chuan, C.-H., Guardino, C.: Designing SmartSignPlay: An interactive and intelligent American Sign Language app for children who are deaf or hard of hearing and their families. In: Proceedings of the International Conference on Intelligent User Interfaces (IUI), pp. 45–48 (2016)
11. Ebling, S., Camgöz, N.C., Braem, P.B., Tissi, K., Sidler-Miserez, S., Stoll, S., Hadfield, S., Haug, T., Bowden, R., Tornay, S., Razavi, M., Magimai-Doss, M.: SMILE Swiss German Sign Language dataset. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC), pp. 4221–4229 (2018)
12. Joy, J., Balakrishnan, K., Sreeraj, M.: SignQuiz: A quiz based tool for learning fingerspelled signs in Indian sign language using ASLR. IEEE Access **7**, 28363–28371 (2019)
13. Potamianos, G., Papadimitriou, K., Efthimiou, E., Fotinea, S.E., Sapountzaki, G., Maragos, P.: SL-ReDu: Greek sign language recognition for educational applications. Project description and early results. In: Proceedings of the PErvasive Technologies Related to Assistive Environments Conference (PETRA) (2020)
14. Papadimitriou, K., Potamianos, G., Sapountzaki, G., Goulas, T., Efthimiou, E., Fotinea, S.-E., Maragos, P.: Greek sign language recognition for the SL-ReDu learning platform. In: Proceedings of the Language Resources and Evaluation Conference Workshop on Sign Language Translation and Avatar Technology (LREC-SLTAT), pp. 79–84 (2022)
15. Sapountzaki, G., Efthimiou, E., Fotinea, S.E., Papadimitriou, K., Potamianos, G.: Educational material organization in a platform for Greek sign language self monitoring and assessment. In: Proceedings of the International Conference on Education and New Learning Technologies (EDULEARN), pp. 3322–3331 (2021)
16. Efthimiou, E., Fotinea, S.-E., Flouda, C., Goulas, T., Ametoglou, G., Sapountzaki, G., Papadimitriou, K., Potamianos, G.: The SL-ReDu environment for self-monitoring and objective learner assessment in Greek sign language. In: Proceedings of the Conference on Universal Access in Human-Computer Interaction. Access to Media, Learning and Assistive Environments (HCII), vol. LNCS-12769, pp. 72–81 (2021)
17. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6450–6459 (2018)
18. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv: 1704.04861 (2017)
19. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**(11), 2673–2681 (1997)
20. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5693–5703 (2019)
21. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M., Lee, J., Chang, W.-T., Hua, W., Georg, M., Grundmann, M.: MediaPipe: A framework for perceiving and processing reality. In: Proceedings of the Workshop on Computer Vision for AR/VR at IEEE CVPR (2019)
22. Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., Black, M.J.: Monocular expressive body regression through body-driven attention. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 20–40 (2020)
23. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5987–5995 (2017)
24. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3D residual networks. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 5534–5542 (2017)
25. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 4489–4497 (2015)
26. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2720–2729 (2017)
27. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1647–1655 (2017)
28. Sun, D., Yang, X., Liu, M.-Y., Kautz, J.: PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8934–8943 (2018)
29. Yu, C., Tianrui, L., Zhen, J., Chengfeng, Y.: BGRU: a new method of Chinese text sentiment analysis. J. Phys. Conf. Ser. **13**(06), 973–981 (2019)
30. Potamianos, G., Papadimitriou, K., Efthimiou, E., Fotinea, S.-E., Goulas, T., Flouda, C., Ametoglou, G.: SL-ReDu deliverable

D5.2: Technical specifications and system architecture definition. Technical report (2021). [Online:] https://sl-redu.ece.uth.gr/deliverable/Deliverable_D5.2.pdf
31. Adaloglou, N., Chatzis, T., Papastratis, I., Stergioulas, A., Papadopoulos, G., Zacharopoulou, V., Xydopoulos, G.J., Atzakas, K., Papazachariou, D., Daras, P.: A comprehensive study on sign language recognition methods. IEEE Trans. Multimed. **24**, 1750–1762 (2022)
32. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255 (2009)
33. Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., Zisserman, A.: A short note about Kinetics-600. arXiv:1808.01340 (2018)
34. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016)
35. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
36. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: Proceedings NIPS-W (2017)
37. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 369–376 (2006)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.