Understanding Motion in Sign Language: A New Structured Translation Dataset

Jefferson Rodríguez[1,2][0000−0003−2394−5683], Juan Chacón[1,2], Edgar Rangel[1,2], Luis Guayacán[1,2], Claudia Hernández[1], Luisa Hernández[1], and Fabio Martínez[1,2,*]

[1] Universidad Industrial de Santander (UIS)
[2] Biomedical Imaging, Vision and Learning Laboratory (BivL^{2}ab)
Bucaramanga, Colombia
{edgar.rangel,claudia2198723,lfherval}@correo.uis.edu.co
{jefferson.rodriguez2,juan.chacon1,luis.guayacan,famarcar}@saber.uis.edu.co

Abstract. Sign languages are the main mechanism of communication and interaction in the Deaf community. These languages are highly variable in communication with divergences between gloss representation, sign configuration, and multiple variants, among others, due to cultural and regional aspects. Current methods for automatic and continuous sign translation include robust and deep-learning models that encode the visual signs representation. Despite the significant progress, the convergence of such models requires huge amounts of data to exploit sign representation, resulting in very complex models. This fact is associated to the highest variability but also to the shortage exploration of many language components that support communication. For instance, gesture motion and grammatical structure are fundamental components in communication, which can deal with visual and geometrical sign misinterpretations during video analysis. This work introduces a new Colombian sign language translation dataset (CoL-SLTD), that focuses on motion and structural information, and could be a significant resource to determine the contribution of several language components. Additionally, an encoder-decoder deep strategy is herein introduced to support automatic translation, including attention modules that capture short, long, and structural kinematic dependencies and their respective relationships with sign recognition. The evaluation in CoL-SLTD proves the relevance of the motion representation, allowing compact deep architectures to represent the translation. Also, the proposed strategy shows promising results in translation, achieving Bleu-4 scores of 35.81 and 4.65 in signer independent and unseen sentences tasks.

1 Introduction

Over 5% of the world's population ($\sim$ 466 million people) have some form of disabling hearing loss. From this group, today, only the 17% use some hearing aid to facilitate communication [1]. Thus, Sign language (SL), a spatial, temporal,

and motion structured set of gestures, constitutes the main channel for interaction and communication of the Deaf community. Like any language, SL around the world reports many variants due to cultural and regional changes, with more than 300 official languages [2]. Even considering methods that focus on a specific regional SL, like any natural language, the problem remains quite challenging due to marked variability of gestures, the richness of glosses, and the multiple modifications that could have any expression during the communication. This fact introduces a huge challenge to the development of assistive devices that allows automatic translation among sign languages and w.r.t spoken languages.

Regarding the automatic SL recognition (SLR), recent advances in computer vision using deep learning strategies have allowed moving from a naive classification of isolated gestures (ISLR) [3, 4] to robust frameworks that allow the continuous recognition (CSLR) and translation of sign languages (SLT) [5, 6]. However, the effectiveness of these strategies depends strongly on large labelled datasets and very complex deep models that must deal with sign variations. Moreover, such approaches only exploit, at least at the first levels, the geometric and spatial relationships with glosses captured from appearance information of video sequences. This would make the models, faced with real scenarios, more complex in order to obtain an adequate sign representation. Therefore, it is necessary to review the main components of SL and try to understand how the interaction of signs is produced and focus on modelling the main components of language. For instance, motion is a fundamental primitive in the development of SL gestures that define much of the relationship among glosses and may even redefine the meaning of many communication segments. In terms of automatic processing, this motion SL component could be the key to deal with variance in gestures, reducing complexity in representation models. However, this motion component is still under-explored in the SL domain, and its use is only implicitly included in semantic and relational processing.

In the literature, both new deep models for SLR and datasets that support these tasks have been proposed, which together have allowed a progression in modelling such challenging tasks. Regarding the SL representation models, nowadays, 2D and 3D convolutional networks are used to extract sign descriptors in images and videos, being the main tool in ISLR [3, 4]. On the other hand, for CSLR and SLT it is common to find, additionally, recurrent neural networks for temporal modeling of signs. Especially for SLT, the sequence to sequence approach with temporal attention models is used to translate the sign sequence into text [6, 7]. Furthermore, some approaches have recently included two-stream approaches to focus on other SL components. For instance, in [8], RGB sequences were modelled together with skeletons to achieve a better representation of the sign communication. In terms of available datasets, there exist different open dataset proposals that record signs from non-intrusive RGB cameras, capturing a significant amount of signs in natural SL communications. These datasets support ISLR [4, 9, 10], CSLR [5, 11] and SLT [12] tasks. Particularly, there are few SLT datasets and those available have long sentences, huge variability of sentences, and words which limit the analysis of additional components of lan-

guage. Hence, proposing new datasets that allow the analysis of others components, such as movement or structure, could be fundamental to understanding how approaches perform sign translation to improve current performance.

This work presents a new structured SLT dataset dedicated to exploring the complementary SL components such as motion and structure and their roles in communication. Despite the importance of pose and geometry in signs, they are visually affected by multiple variations in automatic video analysis. For example, the capture of such language components based on appearance can dramatically affect the representation of signs. As an additional contribution, this paper introduces a novel encoder-decoder SLT strategy that pays attention to temporal structure and motion to demonstrate the ability of these components to support translation. Three main contributions of this work are:

- A new Colombian SLT dataset dedicated to exploring temporal structure and motion information. The set of phrases and glosses were selected to analyze the structure and motion dependencies in the sentences, therefore, signers naturally describe the motion using different articulators during communication. The dataset is open to the scientific community.
- A structured encoder-decoder deep strategy that fully exploits motion information and structural relations in sentences. For doing so, two kinematic attention models are herein introduced to recover short and long kinematic sign correspondences.
- A full validation with a state-of-the-art strategy, based on the deep encoder-decoder architecture. The evaluation is entirely dedicated to exploring the advantages and limitations of motion analysis. Also, how this SL component can reduce complexity in the translation process.

The paper is organized as follows: Section 2 describes the available datasets and the main related approaches focused on SLT, Section 3 introduces the proposed SLT dataset, Section 4, presents the baseline strategy and the proposed method and Section 5 presents a quantitative motion evaluation and the results of our proposed approach.

2 Related Work

Currently, SLT has advanced dramatically due to new gestural representations, translation architectures, and the availability of some datasets that allow more complex and realistic scenarios to be explored. These efforts have allowed the introduction of more difficult problems that require new perspectives and include the analysis of additional linguistic components. The following subsections summarize the state-of-the-art strategies and datasets used today to address SLT.

2.1 SLT Approaches

SLT has been addressed from different approaches, based on strategies that combined convolutional and recurrent networks to try to match an SL with direct

Table 1. Summary of sign language translation datasets.

Dataset	Videos	Sentences	Signers	Lexicon
BOSTON-104	201	113	3	104
RVL-SLLL	140	10	14	104
SIGNUM	780	780	25	450
RWTH-PHOENIX-T	8257	-	9	1066
USTC-ConSents	25000	100	50	178
CoL-SLTD (ours)	1020	39	13	~ 90

correspondence to written languages [6, 7]. These architectures were generally integrated into an encoder-decoder framework forming the approach known as sequence to sequence (seq2seq [13]). These models also include attention modules that perform a weak alignment between the grammatical structures of both languages. In [14, 15] hierarchical attention components were proposed, to encode SL units in video clips. However, clip-level processing limits complex sign recognition and verbal agreements related to the sentence structure, which depends on the entire context. To cover such limitations, in [16] dense temporal convolutions were used to extract short-term relationships and long-term dependencies. Also, in [17] local and global dependencies were learned from a Bidirectional LSTM and temporal correlation modules. These methods, nevertheless, fail in structural modelling due to the use of the CTC (Connectionist Temporal Classification) loss function, typically used for aligned and independent word sequences. A more detailed sign grammatical structure was explored from a multi-classification task that recognizes isolated words in sentences, while a n-gram module classifies subsentences [18]. This approach mitigates the error sentence propagation but the architecture remains limited by the vocabulary size. In a more recent approach, Guo et al. [8] proposed a hierarchical scheme of two streams to describe signs, capture directional and positional verbs, and capture the relationship of motion to the spatial position of articulators in sentences. This approach proves the importance of incorporating a complementary source of sign information by adding skeletons in the decoder module. This hierarchical model merges appearance and positional information reports deficiencies due to misalignment of both information.

2.2 Continuous Sign Language Datasets

Regarding the complexity and diversity of sign languages, there are few datasets that allow exploring SLT tasks. Among these, RVL-SLLL [19] is an American Sign Language (ASL) dataset that allows modelling recognition of connected linguistic contexts on short discourses (10 long sentences), performed under a lexical vocabulary of 104 signs. This dataset has some limitations mainly related to the small number of sentences that difficult the analysis of diverse language components, such as motion information. Similarly, the RWTH-BOSTON-104

Fig. 1. Proposed Col-SLTD: Video sequences were recorded under controlled lighting conditions, on a green background, different clothes and signers with a wide age range. The first two signers (top left) are CODAs (children of deaf adults) and interpreters, the rest of the signers are deaf.

Database [20] has 201 sentences with a vocabulary of 104 signs. Despite the wide range of sentences and structures, this dataset reports a reduced number of videos and signers, which could bias the analysis. In a more linguistically controlled environment, Von et al. [11] proposed a private SIGNUM dataset with 780 pre-defined sentences from German Sign Language, under a lexical vocabulary of 450 signs. The RWTH-PHOENIX-Weather 2014 dataset translation version [6] represents a first large public dataset for SLT with approximately 8000 videos and a vocabulary of 1066 signs and 2887 words. This dataset was built in an uncontrolled scenario but its complexity prevents a detailed linguistic analysis and the language components during communication. Recently, USTC-ConSents is a Chinese language dataset with 5000 videos (wit repetition has 25000 samples) of 100 pre-defined sentences and a lexicon of 178 signs [14]. These datasets represent a huge effort to model sign language but many components of languages remain unexplored because this data limit their analysis. For instance, the analysis and evaluation of kinematic patterns could be associated with verbal agreement and directional and motion verbs. Consequently, capturing data that carefully pays attention in this component could lead to the use of kinematic primitives to help in continuous translation tasks. Table: 1 presents a summary of the above-mentioned datasets.

3 Proposed CoL-SLT Dataset

Sign language, in general, preserves the structural communication Subject-Verb-Object, expressed as a visual combination of hand shapes, articulator locations and movements [21]. The motion shape information is considered the core of the SL, allowing, among others, to differentiate signs related to the pose and also to define the verbal agreement in the sentences [22]. For instance, in American SL, the expression of "I give You" has a similar geometrical description as "You give her", the biggest difference is given by motion direction. Also, while the

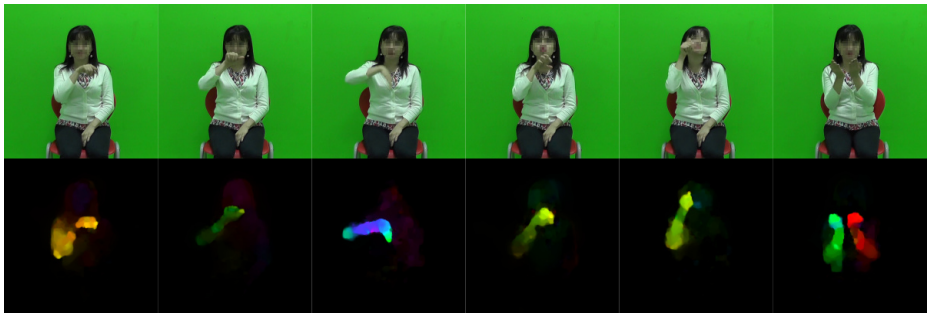

Fig. 2. *Top:* Col-SLTD sign example sequence. *Bottom:* The corresponding optical flow representation. This optical flow allows the accurate tracking and large movements codification, typical of sign language.

hand shapes represent noun classes, the combination with motion patterns could represent associated verbs and complete utterances [23].

This work presents a SLT dataset that focuses efforts on capturing well-formed utterances with structural kinematic dependencies, allowing further analysis of this fundamental linguistic component. To the best of our knowledge, this is the first dataset dedicated to quantify and exploit motion patterns to analyze their correspondence with the sentence structures. The proposed dataset incorporates interrogative, affirmative and negative sentences from Colombian Sign Language. Furthermore, this dataset includes different sentence complexities such as verbal and time signs that define subject and object relationships, such as the phrase: "Mary **tells** John that she will buy a house in the **future**". In this dataset, the videos were pre-processed and interpreted first into written Spanish, as the regional equivalence and then also translated to English equivalence. This dataset also includes signers of different ages to avoid bias in the analysis and to capture a large variability of the same language. This dataset has been approved by an ethics committee of the educational institution. This approval includes informed consent and participant authorization to use this information for the research community.

The proposed SLT dataset, named CoL-SLTD (Colombian Sign Language Translation Dataset), obtains sign expressions using a conventional RGB camera, which facilitates the naturalness of each sign. Each video sequence was recorded under controlled studio conditions using a green chroma key background with lighting conditions, the position of the participants in front of the camera, and the use of clothing of a different color than the background. In CoL-SLTD, there are 39 sentences, divided into 24 affirmative, 4 negative, and 11 interrogative sentences. Each of the sentences has 3 different repetitions, for a total of 1020 sentences, which allows capturing sign motion variability related to specific expressions. Also, the phrases were performed by 13 participants (between 21 to 80 years old), with sentence length between two to nine signs. Figure 1 illustrates the signers of the proposed dataset. All recorded videos were resized to

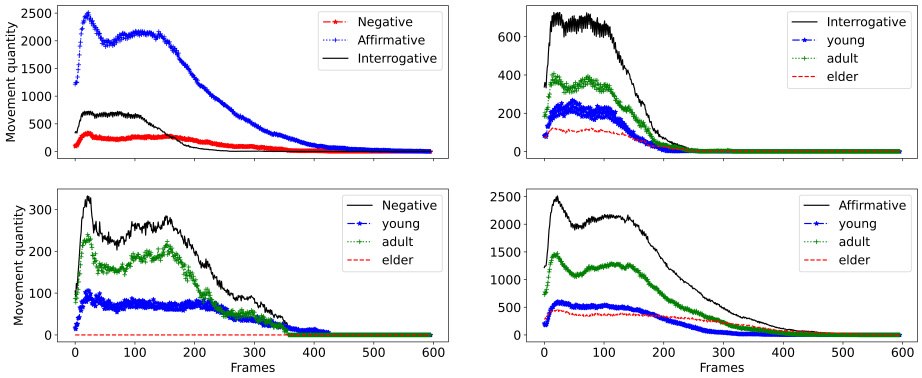

Fig. 3. Motion analysis from optical flow magnitude at frame level: The top left chart compares the quantity of movement present in each frame for the different sentence categories. The remaining three figures analyze the amount of movement performed by signers grouped by age in each sentence type.

spatial resolution of 448x448 with temporal resolutions of 30 and 60 FPS. Also, the whole set was centered on the signer removing a lot of background. Videos have an average length of 3.8 ± 1.5 seconds and an average number of frames of 233 ± 90.

To support the analysis of the motion component, a kinematic vector field descriptor was calculated for each video sign. For this purpose, an optical flow approach with the capability to recover large displacements and relative sharp motions was selected to capture motion signs descriptions at low or high temporal resolutions [24]. Such cases are almost present in any sign, which reports different velocity and acceleration profiles but are especially observed in the exclamation marks. The resultant velocity field $\mathbf{u}_k := (u_{x_1}, u_{x_2})^T$, for a particular frame t is obtained from a variational Euler-Lagrange minimization, that includes local and non-local restrictions between two consecutive frames: $I(\mathbf{x})_t$, $I(\mathbf{x})_{t+1}$. To capture large displacements, a non-local assumption is introduced by matching key-points with similar velocity field patterns. This final assumption could be formalized as: $E_p(\mathbf{u}) = |g_{t+1}(\mathbf{x} + \mathbf{u}(x)) - g_t(\mathbf{x})|^2$ where p is the descriptor vector and $(g_{(t)}, g_{(t+1)})$ are the computed velocity patterns in matched non-local regions. The captured flow field volume result is highly described, keeping spatial coherence and aggregating motion information patterns as a low-level representation. In figure 2 an optical flow sequence computed on the RGB images is illustrated. Also, it is interesting to note in Figure 3 how important sentence patterns are discovered from the optical flow quantification (motion vector norm in each pixel). For example, two big kinematic moments allow identification of affirmative sentences (bottom right). While in interrogative sentences (top right) the movement peaks are not so marked and conversely, they tend to be constant which means that there is more expressiveness.

Table 2. Statistics of each split proposed for evaluation

	SPLIT 1		**SPLIT 2**	
	Train	Test	Train	Test
Number of videos	807	213	922	98
Number of signers	10	3	13	13
Number of sentences	24/10/5	24/10/5	22/9/4	2/1/1
Number of signs	~ 90	~ 90	~ 90	~ 90
Number of words	110	110	110	16

3.1 Evaluation Scheme on CoL-SLTD

Two different evaluations are proposed over Col-SLTD. In a first evaluation, a signer independence split aims to evaluate the capability to translate sequences of signers not seen during training. In this split, a total of 10 signers were selected for training and 3 signers with different ages for testing. In a second evaluation, the task should report the capability to generate sentences not seen during training. In this task, a total of 35 sentences were selected in training and 4 sentences in testing. The words in test sentences have the highest occurrence in training and the sentences involve affirmations, negations, and interrogations. Table 2 summarizes the statistics per split.

A total of three metrics are suggested to evaluate model performance, namely: BLEU score [25], ROUGE-L score (F1-score value) [26] and WER error. The BLEU score measures precision to recover a set of consecutive n-grams. The last two calculate sentence level score and error. The ROUGE-L takes into account similarity regarding sentence structure and identifies longest co-occurrence in compared n-grams sequences and WER error provides complementary information to the scoring metrics.

4 Seq2seq Architecture for SLT

Today, most of the common translation approaches use encoder-decoder architectures, transforming one sequence into another (seq2seq), to translate sign language into a particular written or spoken language. These strategies have shown promising results and therefore these networks are used as a baseline to validate and analyze CoL-SLTD. This section introduces the general principles that follow seq2seq architectures, and how the encoder-decoder model allows for sign translation. Additionally, as a second contribution, a new encoder-decoder scheme is presented here to deal with and address the sign structure and motion component.

4.1 Encoder - Decoder Model

In translation, commonly, the encoder-decoder is composed of two synchronized recurrent neural networks that estimate the conditional probability $p(y_{\{1:m\}}|x_{\{1:t\}})$,

where $(x_1, \ldots, x_t)$ is the sequence of t frames and $(y_1, \ldots, y_m)$ is the corresponding target sequence of m words [13]. On one hand, the encoder codes the inputs in a latent feature space to obtain the state vector h_t at time t. On the other hand, the decoder receives as input the vector h_t to decode and relate with target sequence. The decoder decomposes the conditional probability $p(y_{\{1:m\}}|x_{\{1:t\}})$ into ordered conditional probabilities:

$$p(y_{\{1:m\}}|x_{(1:t)}) = \prod_{m=1}^{M} p(\hat{y}_m|\hat{y}_{\{m-1:1\}}, h_t)\,, \tag{1}$$

where $p(\hat{y}_m|\hat{y}_{\{m-1:1\}}, h_t)$ is the predicted distribution over all m words in the vocabulary. From recurrent methodology, the decoder learns to predict the next most likely word $\hat{y}_m$, conditioned by sign language encoder representation in h_t and previous predicted words $\hat{y}_{\{m-1:1\}}$. These conditional probabilities are solved from stacked RNN (LSTM and GRU) modules that compute the hidden states through the sequence [6, 7, 15]. The error in such models is calculated using word-level cross entropy, described as:

$$\ell = 1 - \prod_{m=1}^{M} \sum_{d=1}^{D} p(y_m^d) p(\hat{y}_m^d)\,, \tag{2}$$

where $p(y_m^d)$ represents the ground truth probability of word y^d at decoding step m and D is the target language vocabulary size.

Baseline Architecture: Herein the NSLT approach was selected to analyze translation on CoL-SLTD [6]. This model uses a pretrained AlexNet 2D-CNN to capture spatial features in each frame x_t. The encoder and the decoder are composed of 4 recurrent layers with GRU units and 1000 neurons in each layer, respectively. This model includes a temporal attention module that provides additional information to the decoding phase by reinforcing the long-term dependencies. Furthermore, this module avoids the vanishing gradients, during the training, as well as the bottleneck caused by the fixed representation of the whole video (very large information) in a fixed embedding vector. The attention module computes a context vector at each decoding step m, as:

$$c_m = \sum_{t=1}^{T} \gamma_t^m h_t\,, \tag{3}$$

where γ_t^m represents the relevance weight of an encoder input x_t to generate the word y_m. These weights are calculated by comparing the decoder hidden state $\hbar$ at step m, with all encoder hidden states h_t, through a scoring function, as:

$$\gamma_t^m = \frac{exp(\hbar_m^\top W h_t)}{\sum_{t'=1}^{T} exp(\hbar_m^\top W h_{t'})}\,, \tag{4}$$

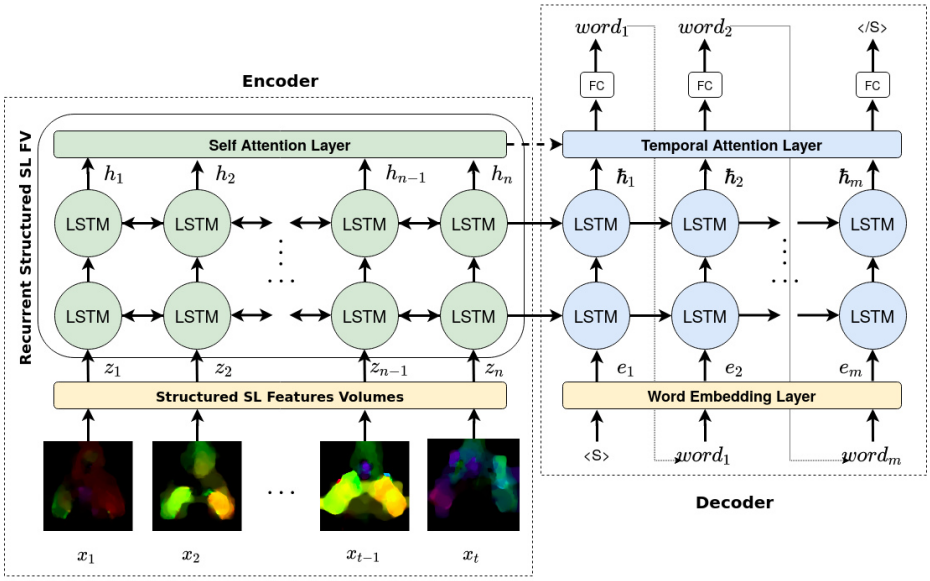

Fig. 4. Proposed structured SLT architecture: Optical flow video is the input to the network. The encoder extracts, at low level, structured kinematic descriptors. Then, at a higher level, the Encoder sequentially processes the descriptors. Finally, they are passed to the Decoder to generate the translation.

where W is a learned parameter. In such sense, the equation 4.1 can be rewritten as follows:

$$p(y_{\{1:m\}}|x_{(1:t)}) = \prod_{m=1}^{M} p(\hat{y}_m|\hat{y}_{\{m-1:1\}}, c_{m-1}, h_t)\,. \tag{5}$$

4.2 Focus on SL Structure

We introduce a new encoder-decoder architecture to robustly include motion modeling and structure in SLT to obtain a very compact representation of the language. At a low level, we use a 3D-CNN network that recovers multiple spatio-temporal features volumes, with relevant short-term kinematic information. Hence, long-term dependencies are captured from an attention module that includes structural and temporal relationships. Also, in the encoder, kinematic descriptors are processed from a stack of bi-directional LSTM modules. These hidden states are refined with a self-attention layer that complements the structural information [27]. Figure 4 illustrates the proposed architecture.

Structured SL Features Volumes Extractor (SFV): Short-term modeling is achieved here by processing optical flow sequences with successive 3D convolu-

tions. This hierarchical scheme obtains multiple non-linear kinematic responses, which describe the motion information, at low levels. Long-term dependencies are modelled on each kinematic response that fully captures the context of the sign. Hence, we apply self-attention [27], along the time axis t' for the complete feature volume $V_r \in \mathbb{N}^{t' \times h' \times w' \times f'}$ for each f' filter responses, in an independent and parallel way. As result, we obtain the square matrix $M'_j \in \mathbb{N}^{t' \times t'}$ which codes the correlation among frames in the same feature filter f'_j. The self-attention computes the weights matrix through the independent projections K (keys) and Q (queries) of the volume Vr in a latent space of dimension p as follows:

$$M'_j = softmax(\frac{Q_{V_r} K_{V_r}^\top}{\sqrt{p}}) \,. \tag{6}$$

The scaling factor $\frac{1}{\sqrt{p}} = 8$ for $p = 64$ avoids small gradients in softmax [27] and the projections are parameter matrices $W^{Q_{V_r}}$ and $W^{K_{V_r}} \in \mathbb{N}^{hw \times p}$.

To include this structural information we apply the frame feature context, defined for each step t'_i of the filter f'_j as:

$$f'_{jt'_i} = \sum_{l=1}^{t'} f'_{jt'_i} M'^{li}_j \tag{7}$$

This frame feature context weights each slice $f'_{jt'_i} \in \mathbb{N}^{h' \times w'}$ to include its structural relationship with other slices in the filter. Figure 5 shows the module in detail with some common normalisation, reduction and fully connected layers.

Recurrent Structured SL Features Vectors (RSFV): High-level sequential dependencies are captured here by using a stack of recurrent bi-directional LSTM layers that receive as a input the computed kinematic volume transformed into a matrix Z with n motion descriptors through the last dense layer. Thus, the final hidden state for each Z_k descriptor, where $k = 1 : n$, is the concatenation of each hidden state from both directions: $h_k = [\overrightarrow{h}_k; \overleftarrow{h}_k]$. To update the hidden states h_k, we propose to include a self-attention layer to refine the relationships between these resulting recurrent vectors. Therefore, the new hidden states are calculated by the following matrix way operation:

$$h_{1:n} = softmax(\frac{Q_h K_h^\top}{\sqrt{p_k}}) V_h \,, \tag{8}$$

where the dimension of the latent space p_k is the same as the hidden vectors $h_k \in R^{512}$ and the projections are parameter matrices W^{Q_h}, W^{K_h} and $W^{V_h} \in \mathbb{N}^{p_k \times p_k}$. For this self-attention the V_h (values) matrix is the result of a third projection of the hidden vectors.

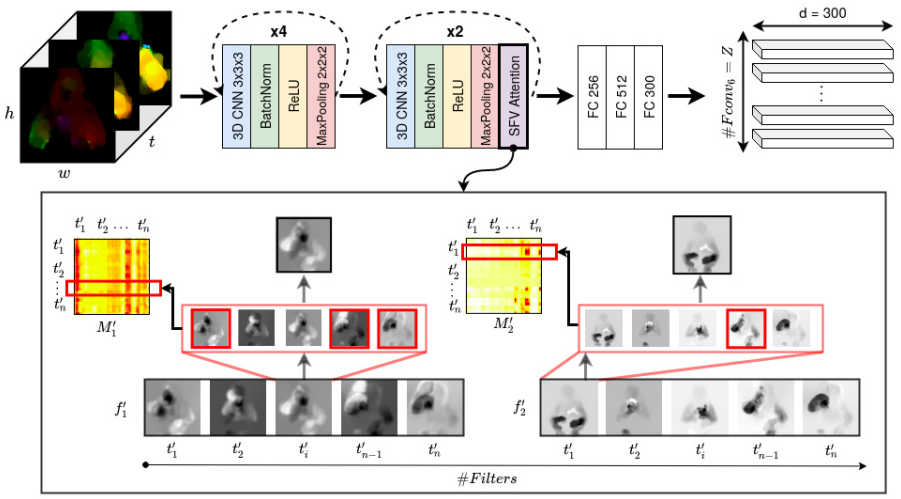

Fig. 5. Structured SL Features Volumes extractor: This proposed module extracts low-level spatio-temporal features volumes through successive 3D-CNN. The SFV attention module takes each resulting convolution and applies self-attention on the whole volume by calculating an attention matrix for each filter in an independent and parallel way. Each feature frame is then related in different proportions to the other frames according to the temporal relationship between them.

5 Experimental Results

The evaluation was designed to determine sign kinematic relevance in CoL-SLTD sentences. Firstly, we analyzed performance with NSLT, trained with 20 epochs, in both RGB and optical flow sequences. This architecture has around 65 Million training parameters (without the AlexNet backbone parameters). Table 3 shows the results obtained for both defined tasks in CoL-SLTD. For Signer Independence evaluation (split 1), the translations generated using the optical flow report around 43% less word error than sentences from the RGB model. The Bleu-4, obtained from flow sequences, also highlights the consistency of the translation with a 46% margin over RGB. These results prove the relevance of the motion in sign recognition and translation.

Regarding the second CoL-SLTD task, to generate unseen sentences (split 2), the table 3 summarizes the obtained results by using the NSLT approach. This task is much more challenging, which could be associated with a poor local representation of the gesture and a language model bias. Nevertheless, even in this case, the motion shape information shows remarkable results w.r.t RGB sequences.

Taking advantage of the resulting motion representation made it possible to address the problem of the complexity. In this sense, a second experiment was designed to compact the network complexity by reducing the recurrent layers and units. In a first experiment, the NSLT was reduced to approximately 25%

Table 3. Translation results for RGB and Flow images in both splits. *Top* of table: results for split 1, *Bottom:* split 2. The experiments were performed with the complete base architecture and then reduced in different proportions.

SPLIT 1	**Data**	**WER**	**Rouge-l**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**
Baseline	RGB	77.41	31.83	30.56	19.78	16.24	14.50
	Flow	34.94	69.91	68.53	63.73	61.42	60.24
Reduction to 75%	Flow	**30.87**	**72.95**	**71.58**	**67.27**	**65.23**	**64.23**
Reduction to 50%	Flow	44.00	62.82	57.09	50.34	47.45	46.07
Reduction to 25%	Flow	62.67	43.92	37.89	26.21	19.67	15.56
SPLIT 2	**Data**	**WER**	**Rouge-l**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**
Baseline	RGB	77.55	23.43	21.68	7.01	2.91	1.74
	Flow	78.33	**36.96**	**39.67**	**18.94**	**12.17**	**8.69**
Reduction to 75%	Flow	78.96	36.17	38.28	15.94	9.28	6.34
Reduction to 50%	Flow	**77.08**	33.73	32.91	11.41	7.18	5.17
Reduction to 25%	Flow	80.06	24.90	26.61	8.05	0.0	0.0

(around 50M parameters less), using only one recurrent layer with 250 neurons. Surprisingly, this compact network achieves even better results than the original RGB representation. Secondly, the best results obtained were when the architecture was reduced to 75%, using 3 layers of 750 neurons (approximately 25M parameters less), demonstrating again the potential use of motion components of language to support sign representation.

From CoL-SLTD, the motion component takes on a relevant role in translation, which could be further utilized in specialized architectures that focus on the attention and model kinematic patterns for a better structural language understanding. In this work, a new seq2seq network was also introduced that exploits mainly motion patterns, the main advantage being the robustness on sign representation (around 10 M parameters). This architecture is composed by the feature extractor module (see figure 5) and a RNN stack with a total of 256 LSTM units in the first layer and 512 in the second layer for encoder and decoder modules[3]. Table 4 summarizes the achieved result in both CoL-SLTD tasks. In the first task, split 1, the architecture with the RSFV self-service module achieves the best results due to the effectiveness to complement the temporal structure initially learned by the LSTM. Interestingly enough, the combination of RSFV y SFV modules improves the Bleu-1 and rouge-l scores, incorporating relevant short-term dependencies captured in SFV. Similarly, for the second task (using split 2), the proposed network achieves similar performance highlighting the relevance of coding, short and structural motion dependencies and their

[3] For training, Adam optimizer was selected with a learning rate of 0.0001 and decay of 0.1 every 10 epochs. Also, batches of 1 sample and a dropout of 0.2 in dense and recurrent layers were herein configured. The convolutional weight decay was set to 0.0005 and gradient clipping with a threshold of 5 was also used.

Table 4. Obtained results using the proposed modules in both splits.

SPLIT 1	**WER**	**Rouge-l**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**
Vanilla approach	64.12	44.39	42.16	33.43	29.91	27.96
SFV module	63.88	45.01	45.90	36.65	32.85	31.02
RSFV module	**58.33**	48.39	47.80	**40.44**	**37.39**	**35.81**
SFV+RSFV modules	59.33	**49.45**	**48.98**	39.98	35.88	33.81
SPLIT 2	**WER**	**Rouge-l**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**
Vanilla approach	90.42	25.59	26.12	10.89	5.21	2.77
SFV module	**88.85**	**30.59**	**30.05**	**12.86**	**7.09**	**4.65**
RSFV module	89.95	24.63	26.08	9.15	4.07	2.41
SFV+RSFV modules	88.85	26.56	27.45	8.94	3.20	1.70

relationships with sign recognition. These relevant kinematic and structural relationships are principal attributed to SVF (short-term dependencies) allowing the achievement of the best performance over the vanilla approach in this task.

6 Conclusions

This work introduced a new sign language translation dataset (CoL-SLTD) that allows the exploration and analysis of motion shape, being one of the fundamental components of language. Through taking advantage of such information, a very compact seq2seq approach was also proposed here to address structure in sign language translation tasks, with remarkable results. The results obtained on the CoL-SLTD prove the relevance of kinematic information to complement sign structure and representation, allowing the design of more compact architectures that could be efficient in real-life conditions. Future works include the continuous growth of this motion dedicated dataset and structural approach, bringing to the scientific community an invaluable source of information to explore new components of sign language.

Acknowledgments

This work was partially funded by the Universidad Industrial de Santander. The authors acknowledge the Vicerrectoriá de Investigación y Extensión (VIE) of the Universidad Industrial de Santander for supporting this research registered by the project: *Reconocimiento continuo de expresiones cortas del lenguaje de señas*, with SIVIE code 1293. Also, we gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

1. centre, W.M.: Deafness and hearing loss (2020) Visited 28-April-2020.

2. centre, W.M.: Our work (2020) Visited 28-April-2020.
3. Joze, H.R.V., Koller, O.: Ms-asl: A large-scale data set and benchmark for understanding american sign language. arXiv preprint arXiv:1812.01053 (2018)
4. Li, D., Rodriguez, C., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: The IEEE Winter Conference on Applications of Computer Vision. (2020) 1459–1469
5. Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding **141** (2015) 108–125
6. Cihan Camgoz, N., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7784–7793
7. Ko, S.K., Kim, C.J., Jung, H., Cho, C.: Neural sign language translation based on human keypoint estimation. arXiv preprint arXiv:1811.11436 (2018)
8. Guo, D., Zhou, W., Li, A., Li, H., Wang, M.: Hierarchical recurrent deep fusion using adaptive clip summarization for sign language translation. IEEE Transactions on Image Processing **29** (2019) 1575–1590
9. Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Yuan, Q., Thangali, A.: The american sign language lexicon video dataset. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE (2008) 1–8
10. Ronchetti, F., Quiroga, F., Estrebou, C., Lanzarini, L., Rosete, A.: Lsa64: a dataset of argentinian sign language. In: XX II Congreso Argentino de Ciencias de la Computación (CACIC). (2016)
11. Von Agris, U., Kraiss, K.F.: Towards a video corpus for signer-independent continuous sign language recognition. Gesture in Human-Computer Interaction and Simulation, Lisbon, Portugal, May (2007)
12. Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J.H., Ney, H.: Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In: LREC. Volume 9. (2012) 3785–3789
13. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. (2014) 3104–3112
14. Huang, J., Zhou, W., Zhang, Q., Li, H., Li, W.: Video-based sign language recognition without temporal segmentation. In: Thirty-Second AAAI Conference on Artificial Intelligence. (2018)
15. Guo, D., Zhou, W., Li, H., Wang, M.: Hierarchical lstm for sign language translation. In: Thirty-Second AAAI Conference on Artificial Intelligence. (2018)
16. Guo, D., Wang, S., Tian, Q., Wang, M.: Dense temporal convolution network for sign language translation. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press (2019) 744–750
17. Song, P., Guo, D., Xin, H., Wang, M.: Parallel temporal encoder for sign language translation. In: 2019 IEEE International Conference on Image Processing (ICIP), IEEE (2019) 1915–1919
18. Wei, C., Zhou, W., Pu, J., Li, H.: Deep grammatical multi-classifier for continuous sign language recognition. In: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), IEEE (2019) 435–442
19. Martínez, A.M., Wilbur, R.B., Shay, R., Kak, A.C.: Purdue rvl-slll asl database for automatic recognition of american sign language. In: Proceedings. Fourth IEEE International Conference on Multimodal Interfaces, IEEE (2002) 167–172

20. Dreuw, P., Rybach, D., Deselaers, T., Zahedi, M., Ney, H.: Speech recognition techniques for a sign language recognition system. In: Interspeech, Antwerp, Belgium (2007) 2513–2516 ISCA best student paper award Interspeech 2007.
21. Stokoe, W.C.: Sign language structure. Annual Review of Anthropology **9** (1980) 365–390
22. Sandler, W.: The phonological organization of sign languages. Language and linguistics compass **6** (2012) 162–182
23. Supalla, T.: The classifier system in american sign language. Noun classes and categorization **7** (1986) 181–214
24. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. IEEE transactions on pattern analysis and machine intelligence **33** (2010) 500–513
25. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics (2002) 311–318
26. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, Barcelona, Spain, Association for Computational Linguistics (2004) 74–81
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. (2017) 5998–6008