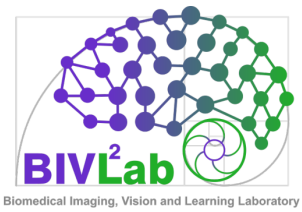


STRUCTURED AND CONTINUOUS VIDEO SIGN LANGUAGE RECOGNITION

JEFFERSON DAVID RODRÍGUEZ CHIVATÁ

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA

2021

STRUCTURED AND CONTINUOUS VIDEO SIGN LANGUAGE RECOGNITION

JEFFERSON DAVID RODRÍGUEZ CHIVATÁ

Research work in partial fulfillment of the requirements for the degree of:
Magíster en Ingeniería de Sistemas e Informática

Advisor:
Fabio Martínez Carrillo
Ph.D in Systems and Computer Engineering

Co-advisor:

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2021

ACKNOWLEDGEMENTS

The author expresses his acknowledgement:

Mainly to Professor Fabio Martínez Carrillo for being a great academic and professional guide, for his patience, dedication, effort, guidance and teamwork. In general, to the research group Biomedical Imaging, Vision and Learning Laboratory for supporting most of the technical and theoretical component needed. Thanks also to all the colleagues who in some way made possible the achievements that today are consolidated in this document.

Special thanks to my couple, family members and parents who have been unconditional and important during this process and deserve to receive all the credits.

Finally, thanks to the Escuela de Ingeniería de Sistemas e Informática (EISI) and the Universidad Industrial de Santander (UIS) and other staff who made this process possible.

CONTENTS

page

LIST OF FIGURES

page

LIST OF TABLES

page

LIST OF APPENDICES

page

ABSTRACT

TITLE: STRUCTURED AND CONTINUOUS VIDEO SIGN LANGUAGE RECOGNITION [*]

AUTHOR: JEFFERSON DAVID RODRÍGUEZ CHIVATÁ [**]

KEYWORDS: SIGN LANGUAGE TRANSLATION, CONTINUOUS SIGN RECOGNITION, SIGN LANGUAGE, SHAPE AND MOTION PATTERNS.

DESCRIPTION: Sign languages are the main mechanism of communication in the deaf community. These languages are highly variable in communication, with divergence in gesture representation, sign configuration and multiple variants due to cultural aspects. Current methods for automatic and continuous sign translation include deep learning models that encode the visual representation of signs. Despite significant advances, the convergence of these models requires huge amounts of data to exploit the sign representation, resulting in very complex models. This fact is associated with increased variability, but also with the limited exploration of many components of language that support communication. For example, gestural movement and grammatical structure are fundamental components in communication, which can address misinterpretations of visual and geometric signs during video analysis. This paper introduces a compact architecture for sign-to-text translation that explores motion as an alternative to support sign translation. Such a characterization is robust to appearance variance with support for geometric variations. In addition, this work proposes two modules that provide robustness to the structural component directly reflected in the translation. The proposed architecture was evaluated on a own Colombian Sign Language dataset built specifically for this task (CoL-SLTD) dedicated to the study of motion and sentence structure, also on a state-of-the-art dataset called RWTH-Phoenix-weather. From the CoL-SLTD dataset, the best configuration reports a BLEU-4 score of 35.81 on the test set. As for the RWTH-Phoenix-weather, the proposed strategy achieved a BLEU-4 score in test set of 4.65 improving the results in similar reduced conditions.

---

[*] Research work

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Advisor: Fabio Martínez Carrillo, Ph.D. Co-advisor:

RESUMEN

TÍTULO: RECONOCIMIENTO ESTRUCTURADO Y CONTINUO DE SIGNOS EN LA LENGUA DE SEÑAS REGISTRADOS EN VIDEO. *

AUTOR: JEFFERSON DAVID RODRIGUEZ CHIVATÁ **

PALABRAS CLAVE: RECONOCIMIENTO CONTINUO DE SIGNOS, LENGUA DE SEÑAS, PATRONES DE FORMA Y MOVIMIENTO, RECONOCIMIENTO ESTRUCTURADO.

DESCRIPCIÓN: Las lenguas de señas son el principal mecanismo de comunicación en la comunidad sorda. Estas lenguas son muy variables en la comunicación, con divergencias entre la representación de los gestos, la configuración de los signos y múltiples variantes debido a aspectos culturales. Los métodos actuales para la traducción automática y continua de signos incluyen modelos de aprendizaje profundo que codifican la representación visual de los signos. A pesar de los importantes avances, la convergencia de estos modelos requiere enormes cantidades de datos para explotar la representación de las señas, lo que da lugar a modelos muy complejos. Este hecho se asocia a la mayor variabilidad, pero también a la escasa exploración de muchos componentes del lenguaje que sustentan la comunicación. Por ejemplo, el movimiento gestual y la estructura gramatical son componentes fundamentales en la comunicación, que pueden hacer frente a interpretaciones erróneas de los signos visuales y geométricos durante el análisis del vídeo. Este trabajo introduce una arquitectura compacta para la traducción de señas a texto que explora el movimiento como alternativa para apoyar la traducción de signos. Dicha caracterización resulta robusta a la varianza de la apariencia con apoyo a las variaciones geométricas. Además, este trabajo propone dos módulos que aportan robustez al componente estructural reflejado directamente en la traducción. La arquitectura propuesta se evaluó en un conjunto de datos propio de lengua de señas colombiana construido específicamente para esta tarea (CoL-SLTD) dedicado al estudio del movimiento y de la estructura de las oraciones, también en un conjunto de datos del estado del arte llamado RWTH-Phoenix-weather. Del conjunto de datos CoL-SLTD, la mejor configuración reporta una puntuación BLEU-4 de 35.81 en el conjunto de pruebas. En cuanto al RWTH-Phoenix-weather, la estrategia propuesta alcanzó una puntuación BLEU-4 en prueba de 4.65 mejorando los resultados en condiciones reducidas similares.

---

* Trabajo de investigación

** Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Fabio Martínez Carrillo, Ph.D. Codirector:

INTRODUCTION

Sign language (SL), a visual-gesture based system, is the main mechanism of communication for the Deaf community, *i.e.,* the primary language alternative for $\sim$ 466 million people with hearing loss, of whom only 17% use a hearing aid [1]. Like any natural language, SL around the world reports many variants due to cultural and regional changes, with more than 300 official languages [2]. Each of these languages has its own grammar, lexicon, and multiple ways to represent words, concepts, and expressions through isolated or composed gestures. Nowadays, the main reason for deficiency in the inclusion of the Deaf community in society is the lack of knowledge of SL, leading to many limitations in access to services, which in most cases is totally null for this population. These complicated facts are mainly related to the absence of interfaces that easily translate from deaf languages to spoken or written languages. Even considering methodologies that focus on a specific regional SL, the learning and modeling of SL remain quite challenging due to marked variability of gestures and the multiple modifications that could have any expression during the communication. Thus, nowadays it is essential, but challenging to develop technological support for this automatic translation.

Technically, SL is represented as a set of visual spatio-temporal gestures structurally connected, known as glosses, which can be represented in their written form through fundamental communication units. These communication components can represent simple words, expressions, and phrases with complex grammatical structures or even complete concepts [3]. Therefore, it is a

---

[1] WHO Media centre. *Deafness and hearing loss*. English. Visited 28-April-2020. World Health Organization, 2020.

[2] WFD Media centre. *Our Work*. English. Visited 28-April-2020. Word Federation of the Deaf (WFD), 2020.

[3] William C Stokoe. "Sign language structure". In: *Annual Review of Anthropology* 9.1 (1980), pp. 365–390.

challenge to automatically understand the visual-manual utterances of the articulators. Additionally, these glosses can be developed in different video lengths, and entail non-linear temporal sign relationships.

In the literature, several automatic SL recognition (SLR) have been proposed ranging from classical naive gesture recognition strategies to more sophisticated frameworks that deal with continuous sign language recognition (CSLR). On the one hand, the classical approaches have mainly been based on hand-craft features that mainly code the appearance of signs to find isolated and global word correspondences over lexical and non-lexical isolated signs (ISLR) [4,5,6]. These approaches, however, lose the temporal capability to recognize gestures in more realistic scenarios. Alternatively, approaches based on Hidden Markov Models (HMMs) [7,8] have been proposed to model sign changes during sequences to continuously recognize signs (CSLR). These approaches exploit appearance and shape sign observations that together with temporal modeling find a sign-text correspondence[9]. Nevertheless, these approaches are based on the hypothesis of an almost consecutive temporal dependence on signs, which leads to a false

---

[4] Morteza Zahedi, Daniel Keysers, and Hermann Ney. "Appearance-based recognition of words in american sign language". In: *Iberian Conference on Pattern Recognition and Image Analysis*. Springer. 2005, pp. 511–519.

[5] Mahmoud M Zaki and Samir I Shaheen. "Sign language recognition using a combination of new vision based features". In: *Pattern Recognition Letters* 32.4 (2011), pp. 572–577.

[6] Jefferson Rodríguez and Fabio Martínez. "A Kinematic Gesture Representation Based on Shape Difference VLAD for Sign Language Recognition". In: *International Conference on Computer Vision and Graphics*. Springer. 2018, pp. 438–449.

[7] Helen Cooper et al. "Sign language recognition using sub-units". In: *Journal of Machine Learning Research* 13.Jul (2012), pp. 2205–2231.

[8] Oscar Koller et al. "Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs". In: *International Journal of Computer Vision* 126 (2018), pp. 1311–1325.

[9] Oscar Koller et al. "Deep sign: hybrid CNN-HMM for continuous sign language recognition". In: *Proceedings of the British Machine Vision Conference 2016*. 2016.

assumption for SL. Currently, more sophisticated learning frameworks have allowed finding complex correlations between raw video volumes and corresponding glosses. In such strategies, deep convolutional features have been used to represent visual signs that, together with recurrent neural networks, exploit more complex temporal relationships [10,11,12]. These approaches have represented a significant advance in the introduction of sign language translation (SLT) systems in real-life scenarios. The neural encoder-decoder architecture introduced in SLT by Camgoz *et. al.*[13] and used in many recent works [14,15] has presented the most promising results. Such architectures, nevertheless, require complex hyper-parametric schemes due to insufficient feature description when using raw appearance and shape information, requiring high computational capabilities. Moreover, these approaches, so far, lose a fundamental component of SL: the sign's motion coherence.

Motion is a fundamental SL primitive that defines 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 poorly

---

[10] Runpeng Cui, Hu Liu, and Changshui Zhang. "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7361–7369.

[11] Necati Cihan Camgöz et al. "SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition". In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 3075–3084.

[12] Shuo Wang et al. "Connectionist Temporal Fusion for Sign Language Translation". In: *2018 ACM Multimedia Conference on Multimedia Conference*. ACM. 2018, pp. 1483–1491.

[13] Necati Cihan Camgoz et al. "Neural Sign Language Translation". In: *CVPR 2018 Proceedings* (2018).

[14] Sang-Ki Ko et al. "Neural Sign Language Translation based on Human Keypoint Estimation". In: *arXiv preprint arXiv:1811.11436* (2018).

[15] Dan Guo et al. "Hierarchical Recurrent Deep Fusion Using Adaptive Clip Summarization for Sign Language Translation". In: *IEEE Transactions on Image Processing* 29 (2019), pp. 1575–1590.

explored in the SL domain, and its use is only implicitly included in semantic and relational processing. Therefore, it is necessary to review this component of the SL and try to understand how the motion interaction of signs is produced.

The main contribution of this research is an in-depth analysis of motion modeling, at different levels and under different coding schemes, to represent sign languages. In summary, the specific contributions are:

- A structured encoder-decoder deep strategy that fully exploits motion information and structural relations in sentences.
- A 3D convolutional optical flow representation that captures relevant kinematic features from video sign. This motion representation, coded in the first module of the encoder-decoder architecture, was integrated and learned together with the translation architecture. In this way, only salient motion primitives related to signs are recovered from video sequences.
- An analysis and evaluation of the gestural attention layers types by determining the main spatio-temporal descriptors correlated with spoken language.
- A new structured SLT dataset dedicated to exploring temporal structure and motion information and their roles in communication. 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.

The proposed approach was fully evaluated w.r.t. translation capability, on two different datasets: Our motion-dedicated sign dataset from Colombian SL (Col-SLTD) and also in the RWTH-Phoenix state-of-the-art dataset. Also, the proposed approach was compared with a state-of-the-art strategy, based on the deep encoder-decoder architecture.

1. SIGN LANGUAGE RECOGNITION (SLR)

Sign language recognition is a research area responsible for studying, analyzing, processing, and modeling sign language to build systems to aid the interaction and communication of the Deaf community. SLR has been addressed from many approaches in the literature, which can be grouped according to the complexity. A primary group has been dedicated to the isolated sign language recognition (ISLR) that includes the alphabet, finger, and word individual characterization. Other groups have a focus on continuous sign language recognition (CSLR) that includes more complex sentence interactions to give a structured translation of a particular set of words. More recently new learning architectures have allowed approaching continuous sign language translation (SLT) that take into account more historical sentence relationships but also with a non-linear nature, and even can predict and generate feature signs.

The proposed methods in each of these areas generally try to exploit the main components of language such as manual signs, the shape, the position and the movement of hands, but also non-manual gestures such as facial expression and body posture, which can modify and complement the meaning from manual signs. Regarding the methods, a timeline can be followed, starting with strategies based on *"hand-craft"* features that provide a specific solution to each sign modeling problem. For instance, seminal ISLR works were proposed from features extracted from appearance [4,5] but with notable limitation on sign description, being also sensible to illumination changes. Other works characterized the signs as a decomposition of fundamental kinematic component,[16], or by describing the signs as a composed representation obtained from optical flow[6]. These approaches, however, lose the temporal capability to recognize gestures in

[16] Konstantinos G Derpanis, Richard P Wildes, and John K Tsotsos. "Definition and recovery of kinematic features for recognition of American sign language movements". In: *Image and Vision Computing* 26.12 (2008), pp. 1650–1662.

more realistic scenarios. Alternatively, approaches based on Hidden Markov Models (HMMs) [7,17,18] have been proposed to model sign changes during sequences to continuously recognize signs. These approaches have been useful to understand the importance of temporal relationships between signs, but their assumptions, most of the time, prove insufficiency in modeling long-term temporal dependencies of language.

Current advances in comprehensive and deep learning approaches, together with robust convolutional representations, have allowed for going beyond traditional recognition tasks. For instance, robust visual representations obtained from convolutional neural networks (CNN) have been integrated into HMMs, to achieve a more robust and continuous recognition of SL[9,19,8]. These CNN-HMM approaches improved the visual description due to the discriminatory CNN properties, allowing a better temporal prediction of the corresponding sign sequences. However, these approaches are still based on Markov's assumptions, modeling only neighborhood sentence units, which have been proved to be insufficient to capture whole temporal connections between gestures. As a consequence, some approaches have dedicated their efforts to model non-consecutive sign relationships, by implementing recurrent neural networks (RNN)[10,20]. These approaches learn long-term dependencies, using for instance Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), from a large amount of matching information between sign

---

[17] Oscar Koller, Jens Forster, and Hermann Ney. "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers". In: *Computer Vision and Image Understanding* 141 (2015), pp. 108–125.

[18] Dan Guo et al. "Online early-late fusion based on adaptive HMM for sign language recognition". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14.1 (2017), pp. 1–18.

[19] Oscar Koller, Sepehr Zargaran, and Hermann Ney. "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4297–4305.

[20] Yuancheng Ye et al. "Recognizing American Sign Language Gestures from within Continuous Videos". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 2064–2073.

and text. These approaches, however, require sophisticated alignment strategies between signs and corresponding glosses. To overcome these limitations, a Connectionist Temporal Classification loss function (CTC) was proposed to find the best dependencies of text sentences with visual sequences, being independent of the spatial sign distribution[21]. From CTC "Sequence-to-sequence" learning strategies were introduced with the main advantage to operate over weakly labeled sign videos[11,12]. However, CTC has almost no inference on visual sign modeling and both the structure and grammar of utterances are poorly exploited. Based on CTC limitations, advanced strategies have been faced with SLT by using an encoder-decoder architecture with RNN units between signs and text[13]. This scheme includes a CNN video representation that, together with temporal attention mechanisms, align both modes of language, achieving translations with structural and grammatical coherence. Similar works have proposed a hierarchical attention and a hierarchical LSTM Encoder module that combines a 3D-CNN video description to achieve sub visual words, words, and video clips translation[22,23]. 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, Guo *et al.*[24] used dense temporal convolutions to extract short-term relationships and long-term dependencies. Also, Song *et al.* [25] proposed to learn global and local dependencies from a Bidirectional LSTM

---

[21] Alex Graves et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". In: *Proceedings of the 23rd international conference on Machine learning*. ACM. 2006, pp. 369–376.

[22] Jie Huang et al. "Video-based sign language recognition without temporal segmentation". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[23] Dan Guo et al. "Hierarchical lstm for sign language translation". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[24] Dan Guo et al. "Dense temporal convolution network for sign language translation". In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press. 2019, pp. 744–750.

[25] Peipei Song et al. "Parallel Temporal Encoder For Sign Language Translation". In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pp. 1915–1919.

and temporal correlation modules. These methods, nevertheless, fail in structural modeling due to the use of the CTC loss function, typically used for independently aligned word sequences. A more detailed sign grammatical structure was explored from a multi-classification task that recognizes isolated words in sentences, while an n-gram module classifies sub-sentences[26]. This approach mitigates the error sentence propagation but the architecture remains limited by the vocabulary size. As an alternative to these appearance-based architectures, Ko *et al.*[14] introduced a strategy to model signs as randomly selected human body poses from 124 key-points and Guo *et al.*[15] proposed a hierarchical scheme of two different streams of information to describe signs and capture directional and positional verbs. These approaches prove the importance of incorporating a complementary source of sign information by adding skeletons as input to encoder and decoder modules, respectively. Despite current encode-decoder advances, the architectures require a huge quantity of parameters to learn sign representation, which results in complex computational approaches. As an alternative, motion processing can reduce architecture complexity and contribute to temporal sign modeling.

[26] Chengcheng Wei et al. "Deep Grammatical Multi-classifier for Continuous Sign Language Recognition". In: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE. 2019, pp. 435–442.

2. RESEARCH PROBLEM

The deaf community, today, faces great challenges of inclusion, mainly related to the absence of mechanisms or interfaces that allow the direct translation of sign language into written and/or spoken languages. This problem is even more critical, considering that each country has its own sign language, and that even regionally these languages can vary drastically. Therefore, it is essential to create tools that support translation by decoding signs into sentences of a given spoken language. However, sign languages, like any natural language, have a high variability in their linguistic units and a richness of multiple signs to represent similar concepts. Also, the multiple grammatical, syntactic and lexical rules contain complex spatial and temporal correlations, which also admit variations during communication. Therefore, the coding and modeling of the linguistic structures represented in a set of signs entails a challenge for the development of systems that support automatic translation. Specifically, there are still challenges for the characterization of visual and motion primitives that allow the compaction of sign representation. On the other hand, there are still reported limitations for the structured and continuous recognition of language, so it is necessary a better coding and modeling of implicit temporality during communication.

Research Question

How contribute temporal, shape and movement patterns to the structured and continuous recognition of signs in sign language recorded on video?

3. OBJECTIVES

General Objective

To propose a computational method for the structured and continuous recognition of signs in sign language using temporal shape and motion patterns on video.

Specific Objectives

- To select a statistically significant set of videos containing continuous signs of a specific sign language.
- To encode visual features of shape and motion in temporal patterns that represent the signs of sign language.
- To develop a translation architecture that approaches the continuous and structured recognition of signs in sign language.
- To evaluate the strategy developed in the selected set of video signs.

4. PROPOSED APPROACH

In recent years, the emerging neural architectures have allowed the analysis and processing of sequential information, providing new representation to very complex tasks, such as the translation of sign language, from video to text. Today, the most remarkable approaches for translation use "Sequence to Sequence" architectures[27]. The main objective of these architectures is to transform the input image sequence (sign representation) into a text sequence translation. Let the sign video $x = (x_1, x_2, \ldots, x_t)$, with t frames, and representing target sequence as $y = (y_1, y_2, \ldots, y_m) \in \sum_{tgr}^{M}$, over a vocabulary of M words, the base probabilistic model of seq2seq is solved, as:

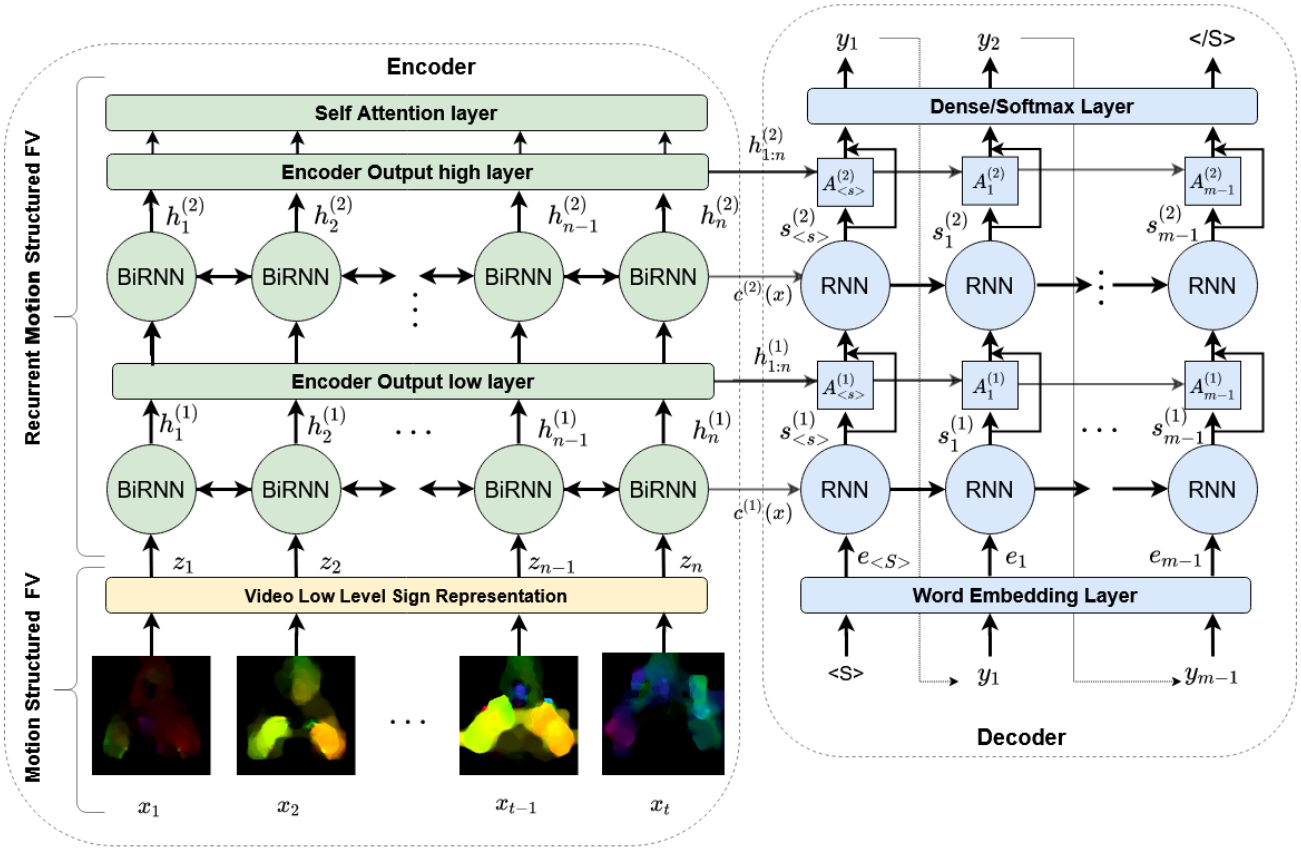
$$P(y|x) \overset{\text{chain rule}}{=} \prod_{j=1}^{m} P(y_j|y_{j-1}, x), \tag{1}$$

So, the the grammatical and structural dependencies of the translation depends on the conditional probability $P(y_j|y_{j-1}, x)$ and the chain rule statement refers to the recurrent relationship present in the language model. To solve this statement, this work introduces a based encoder-decoder model dedicated to extracting and correlating motion patterns with the grammatical structure of signs. A general pipeline of the proposed approach is illustrated in figure 1. The main concepts and components of the architecture are explained in the following subsections. *Part of the content of this chapter is in the second round of evaluation in the international journal* ***IET Computer Vision*** *(see Appendix A).*

---

[27] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems*. 2014, pp. 3104–3112.

Figure 1. Proposed structured SLT architecture: Optical flow video is the input to the network. The encoder processes at low level extracting structured kinematic descriptors. Then, at a higher level, the Encoder sequentially processes the descriptors. Finally, the descriptors are passed to the Decoder to generate the translation.



4.1. An Attention-based Encoder-Decoder for Sequential Motion Learning

Signs are articulated motions, drawn on a spatio-temporal canvas, that follow a temporal coherence to communicate an idea[28,29]. Motion is a fundamental, but poorly analyzed, a component

[28] Scott K Liddell and Robert E Johnson. "American sign language: The phonological base". In: *Sign language studies* 64.1 (1989), pp. 195–277.

[29] Wendy Sandler and Diane Lillo-Martin. "Natural sign languages". In: *The handbook of linguistics* (2001), pp. 533–562.

of sign language, which results communicative by itself and determines transitions between gestures contributing to the grammatical structure of language. Therefore, we have designed an architecture based on the encoder-decoder model to include the kinematic component in the structural and temporal modeling of the signs. Initially, motion fields (optical flow) were calculated on the RGB sequence to process the temporal motion dependencies directly. Then, the first part of the Encoder generates latent feature motion vectors Z_n from a low level 3D-CNN architecture that computes robust kinematic responses. Also, a self-attentional module was designed to include long-term structural relations that facilitate the sequential learning in the second part using a 2-layer RNN that propagate and represent the global temporal behavior of the sequence. In this stage, was also used standard self-attention modules to refine the temporal RNN descriptions of the signs. Finally, the Decoder uses these representations to estimate the recurrent term in equation 4, by associating the motion sign representation vectors $c^{(l)}(x)$ with the language translation model in multiple iteration steps. Namely, a set of 2 recurrent neural networks (RNNs) are used to process such sequences. In detail, for each step j, the Decoder estimates the probability $P(y|x)$ in 4 as:

$$\prod_{j=1}^{m} P(y_j|y_{j-1}, x) = g(y_j|s_{j-1}, y_{j-1}, c^{(l)}(x)), \tag{2}$$

where s_{j-1} is the final hidden state of the RNN Decoder, $g(.)$ is a Fully Connected Deep Network (FCNN) with softmax activation and $l = \{1, 2\}$ is the RNN index Decoder layer. The $g(y_j|s_{j-1}, y_{j-1}, c^{(l)}(x))$ is the estimated word distribution over all M words in the vocabulary. From recurrent methodology, the decoder learns to predict the next most likely word y_j, conditioned by sign language encoder motion representations $c^{(l)}(x)$ and the previous estimated words $y_{\{j-1:1\}}$. The Decoder can be codified from different alternatives such as LSTM and GRU

modules, that compute the hidden states through the sequence[30,14]. Additionally, an attention temporal model is herein included in the Decoder to highlight local temporal patterns that mainly contribute to word translation. This mechanism allows finding complex higher-order temporal relationships between the sequence modes.

4.1.1. A First Motion and Structural Level: Motion Structured Feature Volumes (SFV). A low-level structural motion shape modeling is introduced to code visual sign sequences from a 3D-CNN representation of flow velocity volumes. This representation works as a low-level motion processing, capturing the main dynamic sign patterns without losing the spatial representation. Furthermore, to model the structure, we designed a self-attention module that operates on each CNN response and introduces structural relations present in each filter. In this work, two main assumptions were considered to satisfy proper SL modeling: 1) a motion representation able to capture exaggerated and abrupt sign motions, typically found in daily language and 2) the capability to code dynamic patterns with long-term dependencies during the sequence. As a base, a dense optical flow that considers large coherent motion displacements was herein used[31]. This optical flow facing typical assumptions of very small displacements to recover proper smoothed fields. The captured flow field volume result highly described, keeping spatial coherence and aggregating motion information patterns as a low-level representation. Large displacement in sign representation is very valuable because some exclamation marks are represented by sharp motions and almost all signs have different velocities and accelerations. In detail, this module manly codes short-term relations by processing the optical flow sequences with successive 3D convolutions, capturing the most relevant features of the sign. This hierarchical scheme obtains a volume $V_r \in \mathbb{N}^{t' \times h' \times w' \times f'}$ with reduced t', h', w'

---

[30] Necati Cihan Camgoz et al. "Neural sign language translation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7784–7793.

[31] Thomas Brox and Jitendra Malik. "Large displacement optical flow: descriptor matching in variational motion estimation". In: *IEEE transactions on pattern analysis and machine intelligence* 33.3 (2010), pp. 500–513.

input dimensions and multiple non-linear kinematic responses (f'). The V_r volume describes the motion information at a low-level from different independent operations. Also, for modeling the long-term dependencies and capture the complete context of the sign, we apply, on each kinematic response f', self-attention[32], along the time axis t', in an independent and parallel way. As a result, we obtain the square matrix $M'_b \in \mathbb{N}^{t' \times t'}$ which codes the correlation among frames in the same feature filter f'_b. 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:

$$M'_b = softmax(\frac{Q_{V_r} K_{V_r}^{\top}}{\sqrt{p}}) \,. \tag{3}$$

The scaling factor $\frac{1}{\sqrt{p}} = 8$ for $p = 64$ avoids small gradients in softmax [32] and the projections are parameter matrices $W^{Q_{V_r}}$ and $W^{K_{V_r}} \in \mathbb{N}^{h'w' \times p}$. To include this structural information we apply frame feature context, defined for each step t'_i of the filter f'_b as:

$$f'_{bt'_i} = \sum_{l=1}^{t'} f'_{bt'_i} M'^{li}_b \,. \tag{4}$$

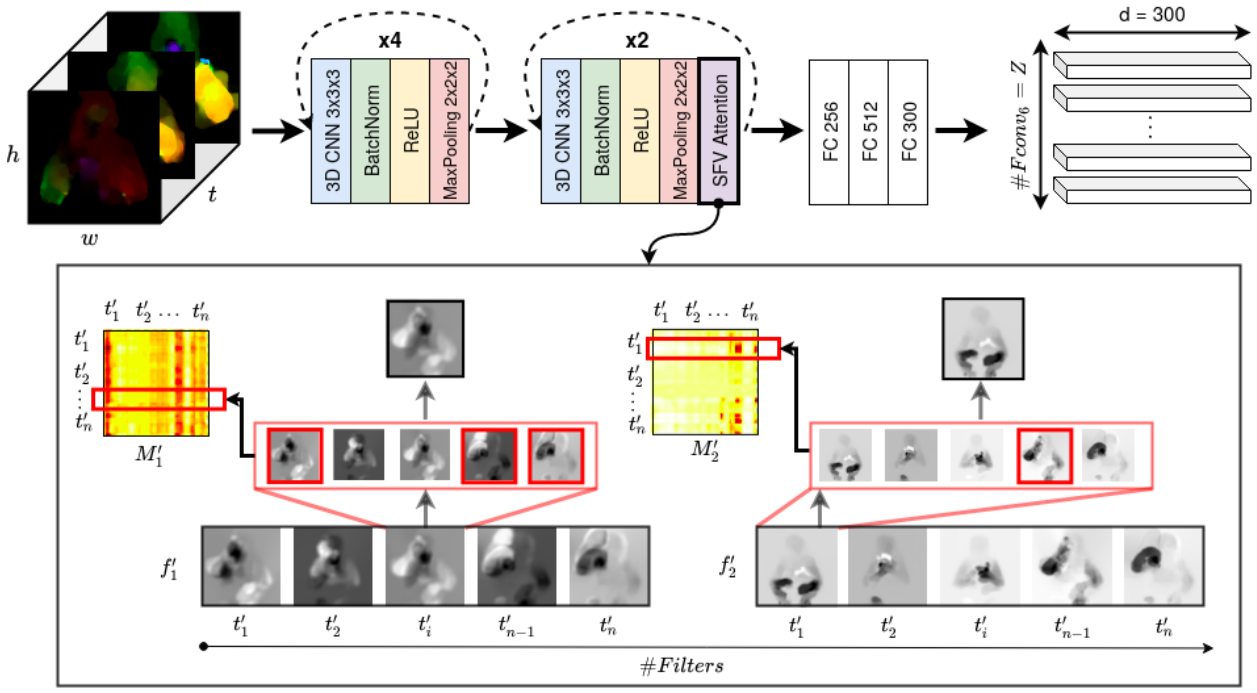
This frame feature context weights each slice $f'_{bt'_i} \in \mathbb{N}^{h' \times w'}$ to include its structural relationship with other slices in the filter. This motion structural representation progressively computes linear transformations, projecting the final information on a set of $d-$ dimensional Z_n high level descriptors, where n indicates the number of filters in the last convolutional layer. Figure 2 shows in detail the module with some additional normalization and activation layers.

4.1.2. A Second Motion and Structural Level: Recurrent Motion Structured Feature Vectors (RSFV). Sign language sequences have very complex compositions that depend, among other things, on particular grammatical compositions, signers' habits, or regional

---

[32] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

Figure 2. 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.



compositions. For instance, interrogative sentences have a strong non-linear correspondence between the beginning and the end of the sentence. A motion encoder is then designed to compute temporal non-linear correlations among the computed motion descriptions. A recurrent multi-layer architecture was then implemented to propagate the structural shape-motion representation (Z_n). This architecture allows describing the global temporal behavior of a particular video sequence. A total of two bi-directional network layers (BiRNN) form the architecture, which together obtains a high-level temporal sign description by computing hidden states in forward and backward directions. At the top of this representation, a single standard self-attention layer was introduced to refine the captured structural relationships. Detailed, in first layer, the deep motion features $Z_n = \{z_1, z_2 \dots z_n\}$, are sequentially propagated by a set of recurrent

units, which compute the states: $\overrightarrow{h}_n^{(1)} = q_1(z_n, \overrightarrow{h}_{n-1}^{(1)})$. In this layer, a resulting mid-level representation captures first temporal dependencies, computing the recurrent units in forward and backward directions, with the resulting concatenated vectors described as: $[\overrightarrow{h}_{1:n}^{(1)}; \overleftarrow{h}_{n:1}^{(1)}]$. An additional second layer was used to recover higher level propagation, taking as input the set of resultant vectors from the first layer. Specifically, a second layer of BiRNN in our proposed approach is designed to capture complex temporal correspondences for more consistent translations. Then, each recurrent unit propagate temporal information as: $\overrightarrow{h}_n^{(2)} = q_2(\overrightarrow{h}_n^{(1)}, \overrightarrow{h}_{n-1}^{(2)})$. In this layer, the propagation is also performed into a bidirectional scheme, with the resulting concatenated vectors: $[\overrightarrow{h}_{1:n}^{(2)}; \overleftarrow{h}_{n:1}^{(2)}]$ where (q_1, q_2) are the activation functions. The Decoder uses the last Encoder output vectors $h_n^{(l)} = c^{(l)}(x)$ as initialization vectors for each Decoder layer l. It should be noted that in our Encoder representation, the Decoder receives both layer representation, which enriches the description of motion sign translation and helps to the text generation process. Also, to update the final higher hidden states $h_{1:n}^{(2)}$, 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:

$$h_{1:n}^{(2)} = softmax(\frac{Q_h K_h^\top}{\sqrt{p_n}})V_h\,, \tag{5}$$

where the dimension of the latent space p_n is the same as the hidden vectors $h_n^{(2)} \in R^{512}$ and the projections are parameter matrices W^{Q_h}, W^{K_h} and $W^{V_h} \in \mathbb{N}^{p_n \times p_n}$. For this self-attention the V_h (values) matrix is the result of a third projection of the hidden vectors.

4.1.3. A Motion Attention-based Decoder: Finally, the Decoder module predicts sequentially a set of $\{y_1, y_2, \ldots, y_m\}$ words given the set of signs recorded in a video sequence. At this level, sign motion units are represented by the encoder outputs $(h_n^{(1)}, h_n^{(2)})$, computed at each BiRNN layer. These observed encoder motion vectors describe kinematic sign history at different time dependence levels. Then, the decoder could be modeled by decomposing the joint probability $P(y|x)$ in sequential conditional probabilities as in 4.1. This conditional prob-

ability is solved by integrating an unidirectional recurrent network with a motion attention mechanism. The network herein considered has a total of two layers. From such integration, it was possible to relate both language modes, *i.e.,* video signs and text. The unidirectional networks layers preserve encoder recurrent unit dimensions, allowing to initialize the hidden states $(s_0^{(1)}, s_0^{(2)})$ with $(c^{(1)}(x), c^{(2)}(x))$. Also, the attention mechanisms acts as a complementary information, at each step j, with only dependency with final descriptors $c^{(l)}(x)$. Then, the $A^{(l)}$ modules associate motion-encoder representations with the associated input word at different representation levels. Each motion attention units $A_{j-1}^{(l)}$, at a particular layer l and step $j-1$, relates all Encoder hidden states $h_{1:n}^{(l)}$ to the hidden representations of each input word in the Decoder through a new context vector $c_{j-1}^{(l)}(x)$. This relationship is defined as the concatenation of context vectors and hidden states, as $A_{j-1}^{(l)} = [c_{j-1}^{(l)}(x); s_{j-1}^{(l)}]$. In this case, the context vector $c_{j-1}^{(l)}(x)$ computes the weights of all encoder motion vectors $h_{1:n}^{(l)}$ w.r.t each $j-1$ input word, which could be expressed as:

$$c_{j-1}^{(l)}(x) = \sum_{i=1}^{n} \gamma_{j-1,i}^{l} h_i^l \tag{6}$$

where $\gamma_{j-1,i}^{l}$ are the attention weights, that define the relevance of a particular encoder input descriptor Z_i to generate the y_j word. From this mechanism, it is possible to capture global and sequential motion patterns rather than isolated information based on hidden states. These weights are calculated by comparing the decoder hidden state s_{j-1}^{l} against each encoder output h_i^l as:

$$\gamma_{j-1,i}^{l} = \frac{exp(s_{j-1}^{l\top} W h_i^l)}{\sum_{i'=1}^{n} exp(s_{j-1}^{l\top} W h_{i'}^l)} \tag{7}$$

where $s_{j-1}^{l\top} W h_i^l$ is the general scoring function[33] and W are the learned parameters used to match the temporal descriptors with the generated words. The hidden states of the Decoder are then computed as follows:

$$s_{j-1}^{(1)} = q_3(e_{j-1}, s_{j-2}^{(1)}) \tag{8}$$

$$s_{j-1}^{(2)} = q_4(A_{j-1}^{(1)}, s_{j-2}^{(2)}) \tag{9}$$

where $e_{j-1} = \text{Embedding}(y_{j-1})$ and (q_3, q_4) are the activation functions. The final prediction word y_j is given by $y_j = \text{softmax}(W_A A_{j-1}^{(2)})$, with W_A as the learned parameters of the fully connected layer. The embedding layer transforms the one hot encoding words of the written language into a dense representation, which allows to relate words with semantic components. For each predicted word y_j, the decoder uses the previous word and hidden states $(y_{j-1}, s_{j-1}^{l_{(1,2)}})$ to update the next hidden states $s_j^{l_{(1,2)}}$. Then, to start the sentence generation process, the y_0 word is the special token $< s >$ that indicates the beginning of the sentence. Finally, this decoder, based on two motion attention mechanisms, enables analysis of overall shape motion representation and highlights main patterns that contribute to a specific word translation. Finally, the equation 4.1 can be rewritten as follows:

$$\prod_{j=1}^{m} P(y_j | y_{j-1}, x) = g(y_j | s_{j-1}, y_{j-1}, c_{j-1}^l(x)) \tag{10}$$

---

[33] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. "Effective approaches to attention-based neural machine translation". In: *arXiv preprint arXiv:1508.04025* (2015).

5. CoL-SLTD: A New Structured Translation Dataset

In the literature, both new deep models for SLR and datasets that support these tasks have been proposed, which together have allowed progress in modeling such challenging tasks. Particularly, there are few SLT datasets and those available have long sentences, huge variability of sentences, and words that limit the analysis of additional components of language. Hence, proposing new datasets that allow the analysis of other components, such as movement or structure, could be fundamental to understanding how approaches perform sign translation to improve current performance. This chapter introduces a new sign language translation dataset (CoL-SLTD), that focus on motion and structural information, and could be a significant resource to determine the contribution of several language components. This new CoL-SLTD dataset is 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. *The content of this chapter has been published in* ***Asian Conference on Computer Vision (ACCV-2020)*** *see Appendix A).*

5.1. Sign Language Translation Datasets

In the state of the art, the few datasets public available are limited to carried out an easy analysis of the grammatical sign structure. For example, RVL-SLLL[34] is an American Sign Language (ASL) dataset that allows to model the recognition of connected linguistic contexts on short discourses (10 long sentences). This dataset has some limitations mainly related to the small number of sentences that difficult the structural analysis of diverse expressions. Sim-

[34] Aleix M Martínez et al. "Purdue RVL-SLLL ASL database for automatic recognition of American Sign Language". In: *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. IEEE. 2002, pp. 167–172.

ilarly, the RWTH-BOSTON-104 Database[35] has 201 sentences with a wide range of sentences and structures. However, this dataset reports a reduced number of videos and signers, which could biases the analysis. In a more linguistically controlled environment, Von et al.[36] proposed a private SIGNUM dataset with 780 pre-defined sentences from German Sign Language. This dataset could be interesting because the sentences were built under strict linguistic rules, but the private nature limit its exploration on scientific community. The RWTH-PHOENIX-Weather 2014 dataset translation version[30] 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[22]. The main disadvantage is that its structure varies considerably, making grammatical analysis difficult. The proposed CoL-SLTD try to preserve the Subject-Verb-Object structure, expressed as a visual combination of hand shapes, articulator locations and movements[3], with the main goal to study the interdependencies between signs in negative, interrogative and affirmative utterances. Table 1 shows a quantitative description of the above-mentioned datasets.

5.1.1. CoL-SLTD Description: Sign language, in general, preserves the structural communication Subject-Verb-Object, expressed as a visual combination of hand shapes, articulator locations, and movements[3]. 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

---

[35] Philippe Dreuw et al. "Speech Recognition Techniques for a Sign Language Recognition System". In: *Interspeech*. ISCA best student paper award Interspeech 2007. Antwerp, Belgium, Aug. 2007, pp. 2513–2516.

[36] Ulrich Von Agris and Karl-Friedrich Kraiss. "Towards a video corpus for signer-independent continuous sign language recognition". In: *Gesture in Human-Computer Interaction and Simulation, Lisbon, Portugal, May* (2007).

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

agreement in the sentences[37]. For instance, in American SL, the expression of "I give You" has a similar geometrical description that "You give her", the biggest difference is given by motion direction. Also, while the handshapes represent noun classes, the combination with motion patterns could represent associated verbs and complete utterances[38].

This work also presents an 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

---

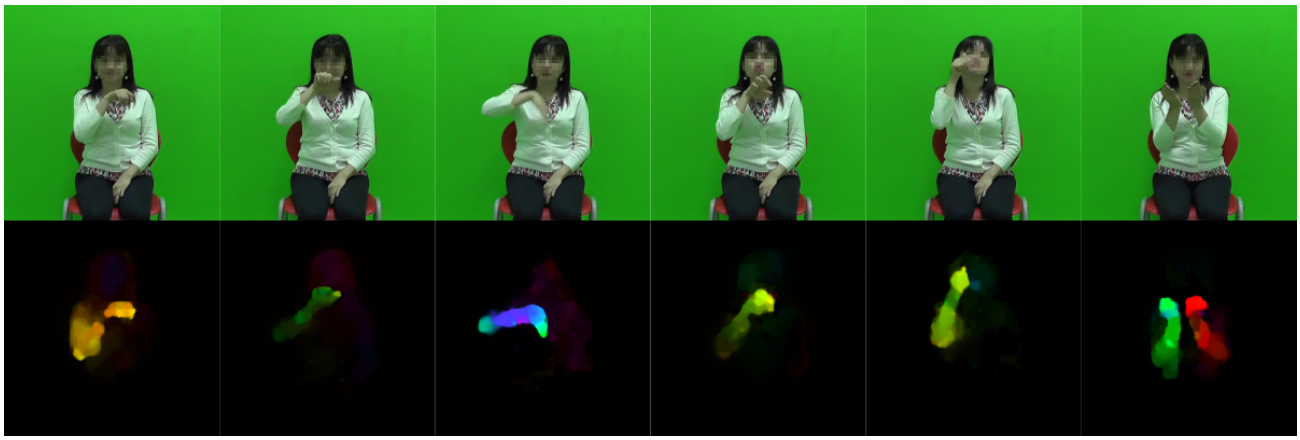
[37] Wendy Sandler. "The phonological organization of sign languages". In: *Language and linguistics compass* 6.3 (2012), pp. 162–182.

[38] Ted Supalla. "The classifier system in American sign language". In: *Noun classes and categorization* 7 (1986), pp. 181–214.

Figure 3. 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.

signers of different ages to avoid bias in the analysis and to capture a large variability for the same language. This dataset has been approved by an ethics committee of the Universidad Industrial de Santander in Bucaramanga – Colombia (see Appendix B for a scanned copy with the assent of the ethics committee). This approval includes informed consent and participants authorize the use of this information for the research community. The proposed SLT dataset, named CoL-SLTD (Colombian Sign Language Translation Dataset), obtains sign expressions from a markerless strategy 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 affirmatives, 4 negatives, 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 3 illustrates the signers of the proposed dataset. All recorded videos were resized to a 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

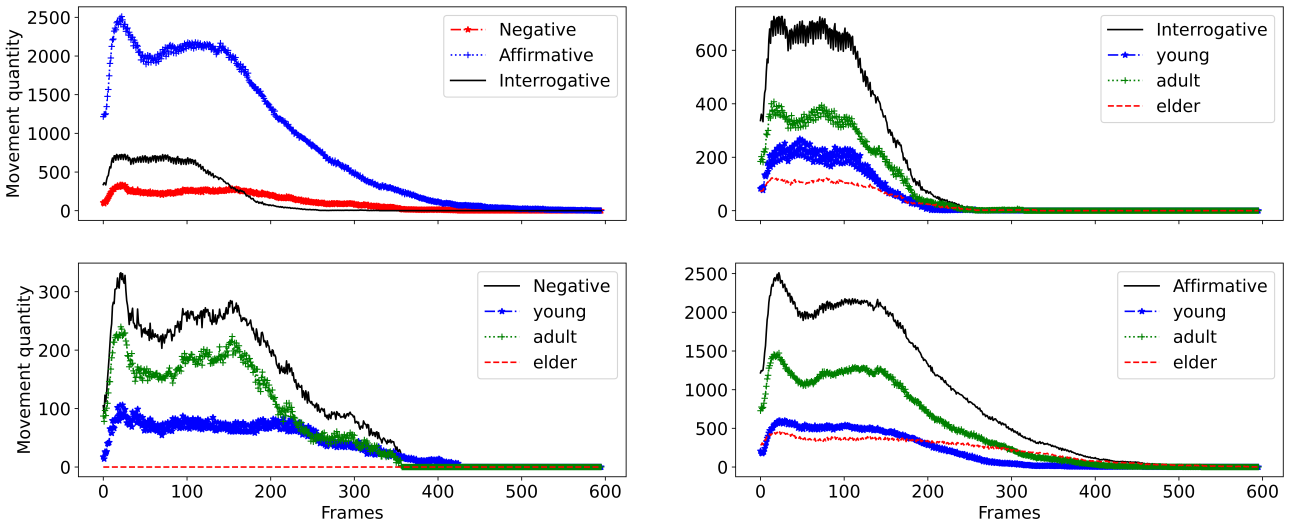
Figure 4. *Top:* CoL-SLTD sign example sequence. *Bottom:* The corresponding optical flow representation. This optical flow allows the accurate tracking and large movements codification, typical of sign language.



3.8 ± 1.5 seconds and an average number of frames of 233 ± 90.

CoL-SLTD: The Sign Motion Component: 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[31]. 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 $\mathbf{u} := (u_{x_1}, u_{x_2})^T$, for a particular frame t is obtained from a variational Euler-Lagrange minimization, that include 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: $E_a(\mathbf{u}) = |g_{t+1}(\mathbf{x} + \mathbf{u}(x)) - g_t(\mathbf{x})|^2$ where a 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 highly described, keeping spatial coherence and aggregating motion information patterns as a low-level representation. In figure 4 is illustrated an optical flow sequence computed on the

Figure 5. 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.



RGB images. Also, it is interesting to note in Figure 5 how important sentence patterns are discovered from the optical flow quantification (motion vector norm in each pixel). For example, two big kinematic moments allow identifying 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.

6. EXPERIMENTAL SETUP

6.1. Sign Language Datasets and Evaluation Schemes

For the proposed approach evaluation, the CoL-SLTD was divided into two different splits, which evaluate two different tasks. Additionally, the validation was performed over a public dataset named RWTH-PHOENIX-Weather 2014T. The next subsection details the configuration of the main components during the evaluation.

6.1.1. 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 no 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.

6.1.2. RWTH-PHOENIX-Weather 2014T Dataset: To exhaustively evaluated the proposed approach, the state-of-the-art *RWTH-PHOENIX-Weather 2014T dataset* and the proposed NSLT[13] (S2T) architecture was also considered in this work. This dataset record sequences that correspond to German sign language with a total of 9 signers that explain weather news on local TV. The vocabulary in such a dataset is composed of 1066 signs that correspond to 2887 words on German spoken language. The dataset is composed of 8257 videos, and the authors suggest a subset of training with 7096 videos, a dev set with 519 videos, and a test set with 642 sequences. This dataset has widely used on current sign language translation strategies because of the rich video information together with challenges of variability of signs recorded at each

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

sequence. This work uses the validation schemes proposed by the authors.

6.2. Configuration of the proposed architecture

The main goal of this work is to analyze the motion contribution on sign language translation and how the proposed deep architecture can recover structural and grammar patterns. Some preliminary experiments were carried out with a first version of CoL-SLTD and a simplified "vanilla" version of our architecture. The first version of the CoL-SLTD dataset only has affirmative sentences that correspond to approximately 50% of the dataset. Regarding the vanilla approach, it works without self-attention modules. From such experiments, the encoder-decoder approach was tuned together with hyperparameters to will be extended on complete experiments. A complete description of the architecture components are described as follows:

- **SFV Module:** The SFV module is composed of six space-time convolutional layers followed by three successive fully connected layers. Every filter is obtained from a $(3\times3\times3)$ kernel with a stride of 1 for all dimensions. Also, Batch normalization, ReLU activation, and max pooling operation with a kernel size of $(2 \times 2 \times 2)$ with a stride of 2 are applied to the volumes resulting from the spatio-temporal convolution. The number of kernels used were $\{32, 32, 64, 64, 128\}$ and Z_n filters, where Z_n is a validation parameter. The structural attention modules were applied only to the last two CNN layers with a dense layer of 64 units.

- **RSFV Module:** For this second module, as mentioned above, we only use two layers of bidirectional RNNs with *tanh* recurrent activation functions. Then, to keep the encoder and decoder fully connected, the first layer has a total of 128 neurons for initial experiments and 256 for finals while for the second layer the total number of neurons is double respectively. The self-attention layer has the same number of neurons used in the second RNN layer.

- **Decoder architecture:** Unlike the Encoder module, the Decoder uses two recurrent layers in only one direction with *tanh* activation functions. Each layer has twice the number of neurons used in its corresponding layer in the Encoder. The input is a 64-dimensional sparse vector, which is subsequently transformed into a 300-dimensional dense representation vector by the embedding layer, with masked padding tokens. Also, for all experiments, we use general attention modules[33]. Each attention module has a single dense layer with the same number of neurons defined in each layer of the Decoder ensuring a fairly compact network.

Model training and learning parameters: The cost function used for training is the following word-level cross entropy:

$$\ell = 1 - \prod_{j=1}^{m} \sum_{d=1}^{M} p(y_j^d) p(\hat{y}_j^d) \,, \tag{11}$$

where $p(y_j^d)$ represents the ground truth probability of word y^d at decoding step j and M is the target language vocabulary size. The scheme used to train the architecture was the Teaching forcing algorithm[39] while the optimizer selected was the stochastic gradient descent (with mini-batches of size 1 because of GPU limitations). For each experiment, a learning rate of 0.0001 was used with a learning rate decay of 0.1 and a dropout of 0.2. The convolutional weight

---

[39] Ronald J Williams and David Zipser. "A learning algorithm for continually running fully recurrent neural networks". In: *Neural computation* 1.2 (1989), pp. 270–280.

decay was set to 0.0005. The gradient clipping was also used. In initial experiments was fixed a gradient clipping with a threshold of 5 and 10 epochs, while for the final experiment was fixed to 20 epochs.

Metrics for evaluation: A total of three metrics are used to evaluate model performance, namely: BLEU score[40], ROUGE-L score (F1-score value[41] 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 the longest co-occurrence in compared n-grams sequences and WER error provides complementary information to the scoring metrics.

---

[40] Kishore Papineni et al. "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, pp. 311–318.

[41] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81.

7. EVALUATION AND RESULTS

The experiments herein proposed were conducted to evaluate and validate motion performance on the proposed SLT deep architecture at different levels of learning and processing. To highlight the contribution of motion patterns on sign representation, the validation of proposed architecture was carried out over two different datasets: the proposed CoL-SLTD, dedicated to motion analysis and the RWTH-PHOENIX Weather dataset, a state-of-the-art large continuous sign database. Next subsections summarize an ablation study over a reduced CoL-SLTD version, as well as, the results obtained on the two complete datasets.

7.1. Results on CoL-SLTD

Firstly, an ablation study of the proposed SLT architecture was carried out over the first compact version of our CoL-SLTD. To explore each principal component of our approach, a "vanilla" version was herein implemented, *i.e.,* without using the proposed self-attention components. Regarding dataset splits, 70% of the participants were used for training, and the remaining 30% for evaluation (signer independent). The contribution of each motion and translation component was analyzed according to the sign translation task. After general net adjustment and validation, the best configuration was evaluated over the complete CoL-SLTD and the RWTH-Phoenix dataset. Additional experiments were also included to compare NLST architecture and self-attention modules.

Parameter searching and adjustment: The shape representation from motion was evaluated as the first stage of the proposed architecture. Then, two different inputs were used in the architecture: velocity frame fields for video sequences (Flow) and RGB raw video information. Sequences were fixed at 128 frames and the last convolutional layer was also fixed with an output of $Z_n = 128$. For the second motion processing, the GRU units were selected as

Table 3. Comparison of translation performance (blue-4 score) using GRU units, video clips of 128 frames and $Z_n = 128$ filters on CoL-SLTD.

	Train			Test		
Type	**Rouge-l**	**Meteor**	**Bleu-4**	**Rouge-l**	**Meteor**	**Bleu-4**
Double Attention (Top and Middle)						
Flow	70.01	73.18	72.05	51.00	49.55	44.28
RGB	15.67	11.63	7.80	6.96	9.79	6.74
Single Attention (Top)						
Flow	**80.91**	**79.83**	**77.90**	**53.00**	**51.37**	43.53
RGB	14.74	8.84	9.75	16.14	8.87	8.78
Single Attention (Middle)						
Flow	63.60	61.88	59.97	52.12	51.05	**44.87**
RGB	29.94	29.94	29.09	19.63	18.99	17.22

recurrent layers. Also, the attention units were positioned at the top and middle of the decoder to obtain single configurations. Also, double attention was configured using the top and middle attention modules, at the same time. In table 3 is summarized the obtained results for both sequences: flow and RGB, regarding training and test sets. As expected, the motion-shape descriptor improves translations compared to RGB descriptors with a remarkable difference in all experiments. Specifically, in test, the proposed motion-shape representation achieves on average a bleu-4 score of 44.22, while the RGB only achieved an average score of 11.00 computed on the Decoder predictions. This fact could be attributed to invariant representation of motion patterns with spatial shape description on convolutional vectors.

Secondly, the impact of recurrent networks was analyzed by also applying LSTM units, that allows encoding motion with non-consecutive temporal relationships. Table 4 reports the obtained results from which the flow sequences also achieve the best performance. A remarkable result, was the LSTM contribution on the second motion level, being able to recover the non-linear dependencies of motion by capturing important patterns along the sequence. It should also be noted that a single attention module in the middle of the architecture obtains the best result. In general, a gain of 10.78% was then obtained for bleu-4 score in test compared to GRU units using the motion shape representation, obtained from the flow velocity patterns. This

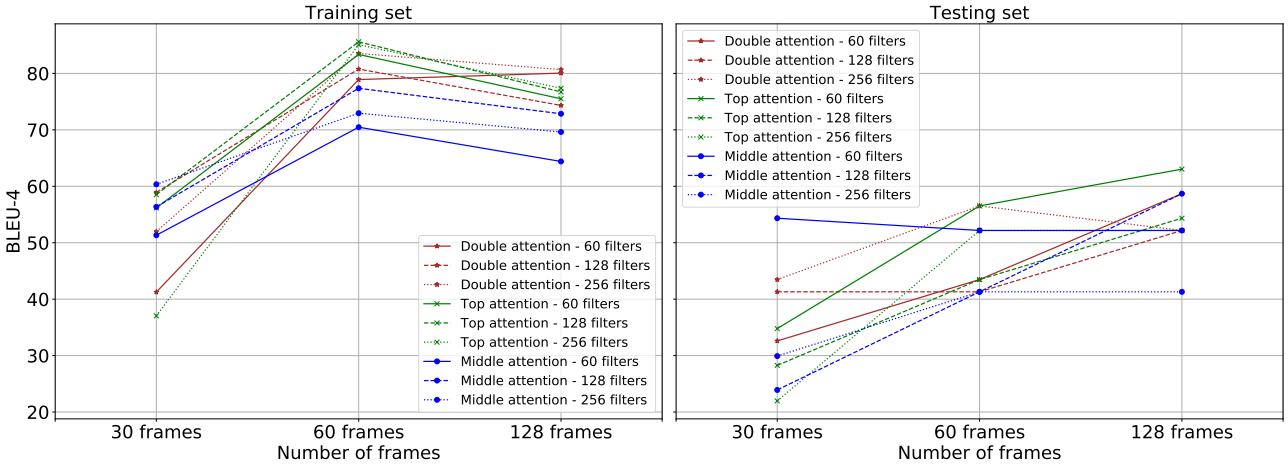
Table 4. Comparison of translation performance (blue-4 score) using LSTM units, video clips of 128 frames and $Z_n = 128$ filters on CoL-SLTD.

	Train			Test		
Type	**Rouge-l**	**Meteor**	**Bleu-4**	**Rouge-l**	**Meteor**	**Bleu-4**
Double Attention (Top and Middle)						
Flow	76.44	75.43	74.33	55.85	53.77	52.17
RGB	20.79	19.06	16.04	24.10	21.82	19.56
Single Attention (Top)						
Flow	**78.46**	**77.40**	**76.72**	**57.70**	**55.70**	54.34
RGB	28.74	25.80	21.64	21.14	18.15	13.96
Single Attention (Middle)						
Flow	75.14	73.91	72.85	60.88	59.58	**58.69**
RGB	14.94	11.29	8.29	6.63	3.27	0.0

configuration was selected for more comprehensive and exhaustive element-architecture evaluations. Figure 6 shows the performance (Blue-4 in the y-axis) for different number of input frames (x-axis), but varying the number of filters (stylish lines) and attention modules (colored lines), for training (left panel) and testing set (right panel). As expected, a higher number of frames in the input (128) improves the representation of the sign, obtaining the best result for the SLT task (63.04), with a gain of 6.52 and 10.87 with respect to results obtained for 60-frame and 30-frame video inputs, respectively. The same quantitative result was also obtained using a configuration of a top attention layer with 60 filters, with a gain of 8.7 and 10.86, w.r.t 128 and 256 CNN filters, respectively. Clearly, the configuration with 60 filters and 128 frames allows an effective coding and compact representation, reducing noise from non-zero activation responses. In addition, the top attention has a gain of 10.87 and 4.35 w.r.t. middle and double attention, respectively. This result could be associated with the temporal dependence of words at the top of the architecture. Also, at the low-level, there exist a major variability that difficult a proper correspondence with words at the semantic level.

Structural component evaluation: From the best-obtained configuration (60 filters, LSTM units, and top attention) we proceeded to validate the SVF and RSVF components in the com-

Figure 6. Results obtained from the architecture parameter validation: A complete validation was performed by changing the frame number of the video, the amount of descriptors or filters used to represent the cinematic patterns and the attention module location to perform a better correlation between information sequences.



plete CoL-SLTD dataset for the two tasks. This complete set includes 510 videos of interrogative and negative sentences. To cover the whole temporal video range, we selected 250 frames, according to the results display in figure 5. The complete model considered in this evaluation has only around 12M of parameters. Table 5 summarizes the achieved result for each independent and combined component in both CoL-SLTD splits. Overall, translation results decrease by about 23% in bleu-4 score for the signer independent task. This fact could be attributed to the complexity of interrogative and negative sentences and the variability introduced by all signers. In particular, for the first task, split 1, the architecture with the RSFV self-attention module effectively complements 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 relationships with sign recognition. These relevant kinematic and structural relationships are principally attributed to SVF (short-term dependencies) allowing the achievement of the best performance regarding

Table 5. Obtained results using the proposed modules in both splits on CoL-SLTD. The architecture was initially evaluated without the proposed modules ("Vanilla"), but they were progressively added to quantify the contribution to the translation process.

SPLIT 1	**WER**	**Rouge-l**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**# Params**
Vanilla approach	64.12	44.39	42.16	33.43	29.91	27.96	11M
SFV module	63.88	45.01	45.90	36.65	32.85	31.02	11.1M
RSFV module	**58.33**	48.39	47.80	**40.44**	**37.39**	**35.81**	11.9M
SFV+RSFV modules	59.33	**49.45**	**48.98**	39.98	35.88	33.81	12M
SPLIT 2	**WER**	**Rouge-l**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**# Params**
Vanilla approach	90.42	25.59	26.12	10.89	5.21	2.77	11M
SFV module	**88.85**	**30.59**	**30.05**	**12.86**	**7.09**	**4.65**	11.1M
RSFV module	89.95	24.63	26.08	9.15	4.07	2.41	11.9M
SFV+RSFV modules	88.85	26.56	27.45	8.94	3.20	1.70	12M

the vanilla approach. Remarkably, there exists a significant performance difference between the two translation tasks, which may be attributed to a possible biased of the model to the most frequent words in sentences.

7.2. Baseline Comparison

In this section was adopted two evaluation baseline schemes: 1) Regarding state-of-the-art approaches on the proposed CoL-SLTD dataset and 2) projecting the proposed approach on a public dataset. Next subsection detail the results obtained in each validation scheme.

7.2.1. Evaluation and results over CoL-SLTD: In this baseline scheme validation, the proposed approach was compared with the state-of-the-art NSLT approach[13] on CoL-SLTD. To validate motion coding capability, the NSLT architecture was also adapted to encode the optical flow inputs. This NSLT strategy is a very large architecture that considers more than 65 million parameters for training. Specifically, the architecture has 4 RNN layers, 1000 neurons in each layer, and uses the AlexNet as a convolutional feature extraction backbone. Table 6 shows the results achieved by NLST, trained with 20 epochs, in both RGB and optical flow

Table 6. Translation results for RGB and Flow images in both splits on CoL-SLTD. *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**
NSLT 100%	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**
NSLT reduced to 50%	Flow	**44.00**	**62.82**	**57.09**	**50.34**	**47.45**	**46.07**
NSLT reduced to 25%	Flow	62.67	43.92	37.89	26.21	19.67	15.56
OURS (RSFV)	Flow	58.33	48.39	47.80	40.44	37.39	35.81
SPLIT 2	**Data**	**WER**	**Rouge-l**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**
NSLT 100%	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**
NSLT reduced to 50%	Flow	**77.08**	**33.73**	**32.91**	11.41	**7.18**	**5.17**
NSLT reduced to 25%	Flow	80.06	24.90	26.61	8.05	0.0	0.0
OURS (SFV)	Flow	88.85	30.59	30.05	**12.86**	7.09	4.65

sequences. 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 (first row). The Bleu-4, obtained from flow sequences, also highlights the translation consistency 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 6 summarizes the obtained results over NSLT. Despite poor local representation and language model bias, the motion shape information shows remarkable results in w.r.t RGB sequences. Secondly, the NSLT network complexity was reduced to compare with the proposed architecture, in similar conditions, and number of parameters. Then, the NLST was reduced to 50% and 25% (around 35M and 50M parameters less respectively) regarding the original configuration. For both splits, we can see how the best results obtained by the proposed approach outperforms the NSLT net reduced to 25% and are close to the NSLT reduced to 50% using approximately 15 million fewer parameters. Surprisingly, the reduced versions of the NSLT network obtain better results than the original RGB representation demonstrating again the potential use of motion components of language to support sign translation.

Table 7. Evaluation of the proposed approach on RWTH-PHOENIX Dataset. The first row compared the performance of the NSLT architecture using optical flow against RGB sequences. The following rows compared our proposed architecture with reduced versions of NSLT using optical flow.

		DEV				TEST			
	Data	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**
NSLT 100%	RGB	31.87	19.11	13.16	9.94	32.24	19.03	12.83	9.58
	Flow	**32.81**	**20.50**	**14.59**	**11.31**	**33.70**	**21.01**	**14.70**	**11.30**
NSLT reduced to 25%	Flow	16.43	6.91	4.98	4.09	17.39	7.32	5.26	4.30
OURS (Vanilla)	Flow	14.25	7.48	5.05	3.81	14.75	8.41	5.87	4.56
OURS (SFV + RSFV)	Flow	**20.89**	**7.65**	**5.35**	**4.70**	**20.19**	**8.08**	**6.01**	**4.82**

7.2.2. Evaluation over RWTH-Phoenix Dataset: To reinforce motion analysis on real scenarios, the proposed strategy was also evaluated over RWTH-Phoenix dataset and compared with NSLT net. The RWTH-Phoenix Dataset (used to validated NSLT), is a large dataset for SLT (about 8000 videos in total), which includes less controlled scenarios and involve both static and motion linguistic expressions obtained from a German television channel[13]. To follow the main purpose of this work, the optical flow sequences were obtained for this dataset. Table 7 summarizes results w.r.t. NSLT approach and for both: RGB and flow sequences. Despite the non-controlled sequences, larger set, and more variable linguistic expressions, the flow representation shows major capability to translation task. For instance, the Bleu-4 score of the NSLT architecture with flow was 1.72 higher than RGB information, a fact associated with better discrimination when geometrical signs are closer. Secondly, the proposed approach was compared with the reduced and compact version of NSLT (25% of parameters, *i.e.,* a total of 15 million parameters with only one recurrent layer of 250 neurons). Table-bottom 7 shows the results obtained using our vanilla version and the version with the proposed modules (SFV and RSFV). Remarkably, the proposed approach achieves better results than the obtained with the reduced NSLT version, being even better the vanilla version in some particular n-grams. Such results can justify the special design of the architecture to recover and properly model motion patterns, enhancing translation on both datasets. Interestingly, the best bleu-1 obtained improves 4.46 to the reduced NSLT, a fact that could be associated with the capability of SFV modules to

extract more important features on SLT.

8. DISCUSSION

An SL motion modeling was herein introduced by using a deep end-to-end low-complexity neural architecture ($\sim$ 12 million of parameters). This architecture was able to model, analyze, and measure the motion patterns capability to support automatic text translation of video sign sequences. The motion was here modeled at different levels and analyzed hierarchically up to the dynamic text spoken translation. As a second contribution in this work, it was proposed Colombian Sign Language Dataset (CoL-SLTD), dedicated to model well-formed phrases where motion is determinant on gestural information. To the best of our knowledge, this is the first dataset dedicated to quantify and exploit kinematic patterns to analyze their correspondence with the sentence structures. The proposed dataset incorporates interrogative, affirmative, and negative sentences from Colombian Sign Language, incorporating verbal and time signs, capturing also large variability from different signers. The proposed approach was firstly evaluated in both: the CoL-SLTD and the RWTH-Phoenix state-of-the-art dataset.

The proposed network, in general, is composed of two main modules. The first module (Encoder) was designed to extract and code spatio-temporal and structural motion patterns, starting from a low-level layer and then capturing recurrent dependencies at a high-level. The Encoder receives the dense optical flow sequences that admit large displacements, which can be useful to characterize the limb movements in conventional cameras. Also, the flow sequences serve as input to a deep 3D-convolutional network to extract the main structural local motion patterns related to signs during the translation process. The implemented 3D CNN is capable of decomposing the volumetric information by operating with 3D learning kernels, progressively organized in a total of six layers. Each layer then retrieves a set of kernels that highlight non-linear motion patterns, which mainly explain particular glosses. Also, to include sign structural information at a local level, we designed an attentional module (SFV) that operates on each feature map resulting from the CNN convolutional layer. This module introduces structural information of

the sentence by finding the temporal dependencies among the convoluted frames. To achieve optimal training convergence, tolerance to high learning rates, and better feature discrimination in this stage, batch normalization was applied on each layer followed by ReLu activations. During the initial evaluation, these flow sequences outperformed experiments regarding raw RGB sequences, with a strong difference of 43.88 on BLEU-4 score average on CoL-SLTD. For SFV evaluation, we found that the module increases the performance and reduces the error in the set of sentences not seen in training (split 2). This could indicate that it improves feature extraction by decreasing translation bias compared to the other configurations. However, the performance is much lower compared to split 1 (gap difference of 31.16 for blue-4). Afterward, a bidirectional recurrent neural network was then used to identify the highly recurrent temporal relationships present in these extracted patterns. At this level, a standard and simple self-attention module (RSFV) was used to refine the recurrent Encoder vectors. Hence, these recurrent motion descriptors are received by the Decoder (second network) as input to be correlated with individual words through gestural temporal attention units. The motion descriptors are decoded by using a unidirectional recurrent network. For the entire network, we analyze the performance of the LSTM and GRU units. As a result, the LSTM units showed superior performance with a bleu-4 gain of 10.78. For RSFV it was evident a good performance in split 1 reaching a bleu-4 of 35.81. This indicates that the structural dependencies learned by RNN are effectively improved at a high level. Likewise, temporal attention modules were evaluated at different levels of decoder architecture to correlate encoder descriptors with the intermediate and higher words vectors. We found that the attention module at top-level reports the best performance in our scheme with a gain bleu-4 of 10.87 w.r.t to double attention modules and 4.35 bleu-4 score for single middle attention. Such a fact could be associated with a high-level description of encoder-decoder embedding.

Many computational approaches have been proposed in the literature to support SLT or to

solve some related sub-task such as continuous independent sign recognition[11,8,42,12]. Some of these approaches have tried to model temporal sign associations using a Connectionist Temporal Classification loss function[21]. However, the structure and grammar of the utterances are poorly modeled. Despite the demonstrated importance of the motion patterns in language to include temporal connections, almost all approaches are fully based on the characterization of isolated gestures or on pure appearance information. For instance, one of the most salient works implemented a 4-layer recurrent encoder-decoder network using raw RGB input sequences, in an architecture with 1000 neurons in each layer and more than 65 million of parameters (S2T)[13]. This work coarsely obtains temporal motion patterns from the recurrent units, but there is no analysis of the contribution of these patterns. In fact, the authors only argue for the use of GRU units instead of LSTM units because of over-fitting issues on the testing set. This fact may associate with the feature extraction process at frame level using 2D convolutions. In contrast, our work introduces an encoder-decoder framework dedicated to analyzing the contribution of motion and structure in SL, demonstrating the powerful and compact representation over the CoL-SLTD dataset. An additional advantage of the motion representation is the compact design of the architecture which allows the use of fewer parameters and results reliable in real applications. Indeed, we developed an additional experiment reducing the S2T architecture to obtain the same number of parameters as our strategy. Then, these architectures were compared over RWTH-Phoenix dataset using flow images to analyze motion propagation. Initially, as expected, the motion input was sufficient to improve the state of the art RGB results using the S2T network. Particularly, the proposed approach achieved a similar performance w.r.t reduced S2T. Interestingly enough, the best configuration of the proposed architectures improves all scores in dev and test sets with respect to the reduced S2T method to 25% and was very close to the results with the S2T architecture reduced to 50%.

---

[42] Junfu Pu, Wengang Zhou, and Houqiang Li. "Dilated Convolutional Network with Iterative Optimization for Continuous Sign Language Recognition." In: *IJCAI*. 2018, pp. 885–891.

In summary, the proposed approach is able to exploit motion and structural representation, from non-controlled video sequences, allowing the translation of sentences of a natural SL. Motion and structural information allows the design of a compact representation and could be useful as a complementary module in more sophisticated architectures based on appearance and geometry. Some limitations were found in recognizing a wide range of sign variations and sentence structures in split 2. The main assumption is that many more samples are needed to capture overall sentence structure and thus decrease translation bias. Similar limitations were also observed in large datasets, from which proposed architecture could be insufficient and very compact regarding the low-level representation, learned by the 3D-CNN, designed with only 60 filters. This last layer could be insufficient to represent properly many variations of the signs. From the current version, the proposed approach will work properly in a specific domain without much computing power.

9. CONCLUSIONS AND FUTURE WORK

This work introduced a novel strategy to SLT based on a deep-learning representation that fully codes velocity apparent pattern inputs. This work takes advantage of a hierarchical encoder-decoder deep representation and introduces modules to code structural patterns in sign language for continuous translation of video sequences. A total of two encoder levels were here implemented to analyze motion and include structural information from the spatio-temporal scene. First, a low motion representation was obtained from a 3D convolutional strategy using optical flow volumes that support special large displacements as an input. These convoluted volumes are refined with self-attention modules to capture the structural relationships in each feature volume. The resulting motion and structural embeddings were able to capture kinematic patterns, invariant in appearance, and represent robustly the structure of the sign for translation. Then, a set of recurrent bi-directional units propagate motion embeddings over time, capturing history on long-term temporal scales. These recurring vectors are refined to highlight those time relationships captured and find patterns not seen by the LSTM. From these two levels of representation, a visual attention module, on a higher level of representation, allowing the text correlation with the embedded sign vectors. The proposed work showed the relevance to include motion sign patterns on automatic translation tasks, resulting in an invariant and compact sign structural representation. The proposed approach was evaluated on a dedicated motion proposed Colombian sign language dataset (CoL-SLTD). Also, the strategy was evaluated on a more general dataset showing promising results from a very compact architecture. Future works include the development of additional attention modules that properly capture motion patterns and deal with complexity in SLT, but remaining compact on deep architecture.

BIBLIOGRAPHY

Brox, Thomas and Jitendra Malik. "Large displacement optical flow: descriptor matching in variational motion estimation". In: *IEEE transactions on pattern analysis and machine intelligence* 33.3 (2010), pp. 500–513 (cit. on pp. 24, 34).

Camgoz, Necati Cihan et al. "Neural Sign Language Translation". In: *CVPR 2018 Proceedings* (2018) (cit. on pp. 13, 17, 36, 44, 46, 50).

Camgöz, Necati Cihan et al. "SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition". In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 3075–3084 (cit. on pp. 13, 17, 50).

centre, WFD Media. *Our Work*. English. Visited 28-April-2020. Word Federation of the Deaf (WFD), 2020 (cit. on p. 11).

centre, WHO Media. *Deafness and hearing loss*. English. Visited 28-April-2020. World Health Organization, 2020 (cit. on p. 11).

Cihan Camgoz, Necati et al. "Neural sign language translation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7784–7793 (cit. on pp. 24, 31).

Cooper, Helen et al. "Sign language recognition using sub-units". In: *Journal of Machine Learning Research* 13.Jul (2012), pp. 2205–2231 (cit. on pp. 12, 16).

Cui, Runpeng, Hu Liu, and Changshui Zhang. "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization". In: *Proceedings of the IEEE Con-*

ference on Computer Vision and Pattern Recognition. 2017, pp. 7361–7369 (cit. on pp. 13, 16).

Derpanis, Konstantinos G, Richard P Wildes, and John K Tsotsos. "Definition and recovery of kinematic features for recognition of American sign language movements". In: *Image and Vision Computing* 26.12 (2008), pp. 1650–1662 (cit. on p. 15).

Dreuw, Philippe et al. "Speech Recognition Techniques for a Sign Language Recognition System". In: *Interspeech*. ISCA best student paper award Interspeech 2007. Antwerp, Belgium, Aug. 2007, pp. 2513–2516 (cit. on p. 31).

Graves, Alex et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". In: *Proceedings of the 23rd international conference on Machine learning*. ACM. 2006, pp. 369–376 (cit. on pp. 17, 50).

Guo, Dan et al. "Dense temporal convolution network for sign language translation". In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press. 2019, pp. 744–750 (cit. on p. 17).

Guo, Dan et al. "Hierarchical lstm for sign language translation". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018 (cit. on p. 17).

Guo, Dan et al. "Hierarchical Recurrent Deep Fusion Using Adaptive Clip Summarization for Sign Language Translation". In: *IEEE Transactions on Image Processing* 29 (2019), pp. 1575–1590 (cit. on pp. 13, 18).

Guo, Dan et al. "Online early-late fusion based on adaptive HMM for sign language recognition". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14.1 (2017), pp. 1–18 (cit. on p. 16).

Huang, Jie et al. “Video-based sign language recognition without temporal segmentation”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018 (cit. on pp. 17, 31).

Ko, Sang-Ki et al. “Neural Sign Language Translation based on Human Keypoint Estimation”. In: *arXiv preprint arXiv:1811.11436* (2018) (cit. on pp. 13, 18, 24).

Koller, Oscar, Jens Forster, and Hermann Ney. “Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers”. In: *Computer Vision and Image Understanding* 141 (2015), pp. 108–125 (cit. on p. 16).

Koller, Oscar, Sepehr Zargaran, and Hermann Ney. “Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent cnn-hmms”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4297–4305 (cit. on p. 16).

Koller, Oscar et al. “Deep Sign: Enabling Robust Statistical Continuous Sign Language Recognition via Hybrid CNN-HMMs”. In: *International Journal of Computer Vision* 126 (2018), pp. 1311–1325 (cit. on pp. 12, 16, 50).

Koller, Oscar et al. “Deep sign: hybrid CNN-HMM for continuous sign language recognition”. In: *Proceedings of the British Machine Vision Conference 2016*. 2016 (cit. on pp. 12, 16).

Liddell, Scott K and Robert E Johnson. “American sign language: The phonological base”. In: *Sign language studies* 64.1 (1989), pp. 195–277 (cit. on p. 22).

Lin, Chin-Yew. “ROUGE: A Package for Automatic Evaluation of Summaries”. In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81 (cit. on p. 39).

Luong, Minh-Thang, Hieu Pham, and Christopher D Manning. "Effective approaches to attention-based neural machine translation". In: *arXiv preprint arXiv:1508.04025* (2015) (cit. on pp. 29, 38).

Martínez, Aleix M et al. "Purdue RVL-SLLL ASL database for automatic recognition of American Sign Language". In: *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. IEEE. 2002, pp. 167–172 (cit. on p. 30).

Papineni, Kishore et al. "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, pp. 311–318 (cit. on p. 39).

Pu, Junfu, Wengang Zhou, and Houqiang Li. "Dilated Convolutional Network with Iterative Optimization for Continuous Sign Language Recognition." In: *IJCAI*. 2018, pp. 885–891 (cit. on p. 50).

Rodríguez, Jefferson and Fabio Martínez. "A Kinematic Gesture Representation Based on Shape Difference VLAD for Sign Language Recognition". In: *International Conference on Computer Vision and Graphics*. Springer. 2018, pp. 438–449 (cit. on pp. 12, 15).

Sandler, Wendy. "The phonological organization of sign languages". In: *Language and linguistics compass* 6.3 (2012), pp. 162–182 (cit. on p. 32).

Sandler, Wendy and Diane Lillo-Martin. "Natural sign languages". In: *The handbook of linguistics* (2001), pp. 533–562 (cit. on p. 22).

Song, Peipei et al. "Parallel Temporal Encoder For Sign Language Translation". In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2019, pp. 1915–1919 (cit. on p. 17).

Stokoe, William C. “Sign language structure”. In: *Annual Review of Anthropology* 9.1 (1980), pp. 365–390 (cit. on pp. 11, 31).

Supalla, Ted. “The classifier system in American sign language”. In: *Noun classes and categorization* 7 (1986), pp. 181–214 (cit. on p. 32).

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. “Sequence to sequence learning with neural networks”. In: *Advances in neural information processing systems*. 2014, pp. 3104–3112 (cit. on p. 21).

Vaswani, Ashish et al. “Attention is all you need”. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008 (cit. on p. 25).

Von Agris, Ulrich and Karl-Friedrich Kraiss. “Towards a video corpus for signer-independent continuous sign language recognition”. In: *Gesture in Human-Computer Interaction and Simulation, Lisbon, Portugal, May* (2007) (cit. on p. 31).

Wang, Shuo et al. “Connectionist Temporal Fusion for Sign Language Translation”. In: *2018 ACM Multimedia Conference on Multimedia Conference*. ACM. 2018, pp. 1483–1491 (cit. on pp. 13, 17, 50).

Wei, Chengcheng et al. “Deep Grammatical Multi-classifier for Continuous Sign Language Recognition”. In: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE. 2019, pp. 435–442 (cit. on p. 18).

Williams, Ronald J and David Zipser. “A learning algorithm for continually running fully recurrent neural networks”. In: *Neural computation* 1.2 (1989), pp. 270–280 (cit. on p. 38).

Ye, Yuancheng et al. “Recognizing American Sign Language Gestures from within Continuous Videos”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 2064–2073 (cit. on p. 16).

Zahedi, Morteza, Daniel Keysers, and Hermann Ney. “Appearance-based recognition of words in american sign language”. In: *Iberian Conference on Pattern Recognition and Image Analysis*. Springer. 2005, pp. 511–519 (cit. on pp. 12, 15).

Zaki, Mahmoud M and Samir I Shaheen. “Sign language recognition using a combination of new vision based features”. In: *Pattern Recognition Letters* 32.4 (2011), pp. 572–577 (cit. on pp. 12, 15).

APPENDICES

Anexo A. Academic Products

Posters

- J. Rodríguez, F. Martínez. "Sign Language Translation using Motion Filters and Attention Models". Latin American Meeting In Artificial Intelligence (KHIPU 2019).
  Status: Presented.

Journals

- J. Rodríguez, F. Martínez. "How important is motion in sign language translation?". IET Computer Vision, 2020. United Kingdom.
  Status: Pre accepted, Second round: Major revisions.

Conference papers

- J. Rodríguez, J. Chacón, E. Rangel, L. Guayacán, C. Hernández, L. Hernández, F. Martínez. "Understanding Motion in Sign Language: A New Structured Translation Dataset". Asian Conference on Computer Vision (ACCV 2020).
  Status: Accepted.

Collaborations

- A. Moreno, J. Rodríguez, F. Martínez. "Regional Multiscale Motion Representation for Cardiac Disease Prediction". XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA 2019). In IEEE Xplore Digital Library. 06 June 2019, IEEE. DOI: 10.1109/STSIVA.2019.8730231.
  Status: Published.

- A. Moreno, J. Rodríguez, F. Martínez. "Kinematic Motion Representation in Cine-MRI to support Cardiac Disease Classification". Computer Methods in Biomechanics and Biomedical Engineering 2020.
  Status: Submitted.

Anexo B. Informed Consent

ESCUELA DE INGENIERÍA DE SISTEMAS
UNIVERSIDAD INDUSTRIAL DE SANTANDER - Laboratorios Vive Digital
CONSENTIMIENTO INFORMADO

Proyecto: Grabación y cuantificación de un conjunto de signos gestuales con significado semántico en la lengua de señas colombiana.
Responsables: Fabio Martínez Carrillo, Juan Felipe Chacón López, Jefferson David Rodriguez Chivatá.

Con base en los reglamentos establecidos en la Resolución Nº 008430 del 4 de octubre de 1993 por la cual se establecen las normas científicas, técnicas y administrativas para la investigación en salud en Colombia y según el artículo 15 relacionado con el Consentimiento Informado usted deberá conocer de forma completa y clara los aspectos de la investigación que se desarrollará. Usted ha sido convocado para este proyecto por cumplir con los requisitos de inclusión para la grabación de un conjunto de datos de la lengua de señas Colombiana. Por tal motivo se le invita formalmente a que participe del estudio teniendo en cuenta los siguientes criterios de inclusión:

- Ser mayor de edad.
- Tener la capacidad de sentarse en una postura cómoda y relajada que le permita centrar su atención en una cámara.
- Saber la lengua de señas colombiana.

De acuerdo con lo anterior y en cumplimiento de estos criterios, por favor indique con una X en una de las siguientes opciones qué tipo de participante es usted:

___ Persona con discapacidad auditiva o sordo.
___ Oyente, hijo de padres sordos (CODA).
___ Oyente, hijo de padres oyentes.

Tenga en cuenta que su participación en este proyecto es ***absolutamente voluntaria***. Por favor lea con cuidado el documento y haga todas las preguntas que desee hasta su total comprensión.

JUSTIFICACIÓN

Usted está invitado a participar en este ejercicio que busca crear un conjunto de datos en video de gestos específicos pertenecientes la lengua de señas local (Bucaramanga-Colombia) para investigaciones futuras sobre estudio, análisis, procesamiento, caracterización, clasificación y reconocimiento automático de las mismas. El conjunto de datos registrados será de uso exclusivamente académico y científico. Este conjunto de datos serán capturados con una cámara convencional.

OBJETIVO

Registrar un conjunto de vídeos que permitan desarrollar métodos para la detección automática de las señas.

DESCRIPCIÓN

Para la realización del estudio se cuenta con un laboratorio de grabación en el **Punto Plus Vive Digital Tecnológico** ubicado en la calle 10 # 28 - 77. Durante la sesión de captura adicionalmente de los investigadores estará un intérprete de la lengua de señas colombiana quien facilitará la comunicación entre los investigadores y el participante. Cada participante, en presencia del intérprete que acompañe la sesión recibirá este consentimiento para su lectura y aclarar dudas, si decide participar, podrá proceder a firmarlo.

Una vez firmado el consentimiento se le registran unos datos personales como mano hábil, condición de naturaleza por la cual sabe comunicarse mediante lengua de señas y se le asignará un ID respectivo para la confidencialidad de los datos . La grabación de los videos tiene un tiempo aproximado de 20 minutos. La indumentaria ideal es camiseta o blusa. (No busos) y preferiblemente sin gafas o cualquier otro tipo de accesorios adicionales. Antes de la grabación el intérprete se le mostrará un listado de frases las cuales serán las que él dirá delante de la cámara. Ante la cámara, se le mostrará cada una de las frases anteriores para que la diga en lengua de señas, como se puede ver en la Figura 1.

Figura 1. Escenario de captura. el participante se sentará de la manera que considera más cómoda para la grabación de las señas.

Al participar en este estudio, usted no recibirá ningún tipo de subvención económica o material ni deberá aportar herramienta alguna para la intervención. Al finalizar la investigación, usted podrá recibir los resultado obtenidos de la captura. Este material será presentado a usted por los investigadores cuando culmine la actividad.

Las inquietudes adicionales que surjan en relación con el desarrollo e implicaciones del proyecto podrán ser aclaradas por Fabio Martínez Carrillo, Profesor de la Escuela de Ingeniería de Sistemas e Informática, a quien puede contactar en el teléfono 3103054041, o mediante correo electrónico dirigido a famarcar@saber.uis.edu.co ; o directamente en su oficina en la Universidad Industrial de Santander (sede principal) ubicada en la Cra. 27 #9 Ciudad Universitaria, Edificio de Laboratorios Pesados, oficina 231; o al teléfono teléfono 577- 6344000 extensión 2110.

RIESGOS

De acuerdo con el Artículo 11 de la Resolución No. 8430 de 4 de octubre de 1993, esta investigación se considera sin riesgo para el participante dado que el estudio únicamente emplea el registro de datos a través de un procedimiento común de captura de vídeos por medio de cámaras ordinarias. De tal forma, ninguno de los métodos utilizados es invasivo o penetra la piel. si durante la captura de los vídeos usted experimenta cualquier tipo de malestar, la grabación será suspendida de inmediato y se le ubicará en estado de reposo. Por cualquier motivo relacionado con esta jornada donde el participante requiere valoración médica inmediata será remitido al servicio de urgencias del Hospital Universitario de Santander o, si es su decisión al servicio de urgencias de la entidad donde se encuentre afiliado al sistema de seguridad social. Durante este proceso será acompañado por el investigador principal.

DERECHO A RETIRARSE

Su participación en este estudio es autónoma y voluntaria, en donde podrá actuar acorde a sus principios personales. Si usted decide no participar, no implicará sanción alguna. Además, usted cuenta con el derecho a negarse a responder a preguntas concretas si así lo desea. También puede optar por retirarse en cualquier momento y toda su información será descartada y eliminada.

CONFIDENCIALIDAD

Los resultados de las pruebas y la información que usted nos ha dado son de carácter absolutamente confidencial, de manera que solamente usted y el investigador principal tendrán acceso a estos datos.

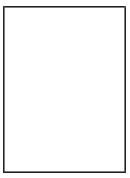
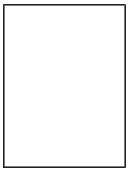
Una copia de los registros con la información de cada participante será archivada por el investigador principal y a cada registro se le asignará un número con el cual se identificará y codificará para su ingreso a la base de datos durante la sistematización de la información. Por lo anterior, los nombres de los participantes no serán divulgados en forma alguna; y cuando los resultados de este estudio sean publicados en revistas o congresos científicos, la información personal de los participantes será debidamente anonimizada previamente.

A menos que Usted dé una autorización específica cuando la ley lo permita, sus resultados personales no estarán disponibles para terceras personas como empleadores, organizaciones gubernamentales, compañías de seguros o instituciones educativas. Esto también aplica a su cónyuge, a otros miembros de su familia. Sin embargo, con el objetivo de realizar un manejo adecuado de los datos, un miembro del Comité de Ética de la Universidad Industrial de Santander podrá consultar sus datos y su registro. Por lo anterior, atentamente se le invita a participar en el estudio y si está de acuerdo, se le solicita su nombre y la firma en las casillas abajo descritas.

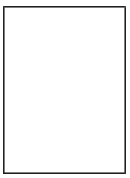
AUTORIZACIÓN PARA EL USO DE LA INFORMACIÓN EN ESTUDIOS FUTUROS

Dentro del equipo de investigación al que pertenecen los investigadores responsables (Grupo de Investigación BIVL^{2}ab - *Biomedical Imaging, Vision and Learning Laboratory*) de la Universidad Industrial de Santander, se espera seguir utilizando la información registrada en este estudio para el desarrollo de estudios futuros y derivados. Por lo tanto, al firmar este consentimiento usted puede autorizar al investigador principal a ceder su información a otros investigadores de su equipo de investigación, con previa aprobación del Comité de Ética de la Universidad Industrial de Santander para realizar los estudios mencionados. Por favor marcar con una X si autoriza o no autoriza, y firmar en caso de si autorizar.

___ Si autorizo
___ No autorizo

_____	
Firma Participante	Huella digital
Nombre:	(En caso que se amerite)
C.C.	
_____	
Firma Investigador Principal	Huella digital
Nombre:	(En caso que se amerite)

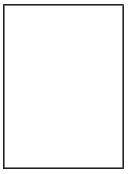
Yo_____, identificado con _____ N°_____ de _____ al firmar este consentimiento el día ___ de _____ del _____, acepto participar de manera voluntaria en el presente estudio y autorizo la grabación de mis vídeos y el uso de mis datos individuales para los análisis requeridos. He leído y entendido la información registrada en este documento y mis dudas fueron aclaradas. Entiendo que soy libre de retirarme del estudio. Por otro lado, se me ha garantizado justicia, equidad, autonomía en la participación y la confidencialidad en el manejo de toda la información recolectada, teniendo en cuenta que los resultados del procesamiento de dicha información podrán ser divulgados con fines científicos, mediante presentaciones en congresos o publicaciones en revistas científicas, con la debida protección de mi identidad



_____

Firma Participante
Nombre:

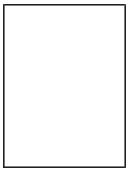
Huella digital
(En caso que se amerite)



_____

Firma Intérprete (Testigo 1)
Nombre:

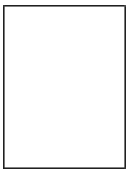
Huella digital
(En caso que se amerite)



_____

Firma Coinvestigador (Testigo 2)
Nombre:

Huella digital
(En caso que se amerite)



_____

Firma Investigador Principal
Nombre:

Huella digital
(En caso que se amerite)

Contacto Comité de Ética: Para preguntas o aclaraciones acerca de los aspectos éticos de ésta investigación pueden comunicarse con el Comité de Ética para la Investigación Científica de la Universidad Industrial de Santander (CEINCI-UIS), o con cualquiera de los

miembros del Comité, al teléfono 6344000 Extensión 3808 ó al correo comitedetica@uis.edu.co.