How important is motion in sign language translation?

Jefferson Rodriguez | Fabio Martínez

Biomedical Imaging, Vision and Learning Laboratory (BivL Lab), Universidad Industrial de Santander (UIS), Bucaramanga, Colombia

Correspondence
Fabio Martínez, Biomedical Imaging, Vision and Learning Laboratory (BivL Lab), Universidad Industrial de Santander (UIS), Bucaramanga, Colombia.
Email: famarcar@saber.uis.edu.co

Funding information
Universidad Industrial de Santander

Abstract
More than 70 million people use at least one sign language (SL) as their main channel of communication. Nevertheless, the absence of effective mechanisms to translate massive information among sign, written and spoken languages is the main cause of a negligible inclusion of deaf people into society. Therefore, SL automatic recognition systems have widely proposed to support the characterisation of the sign structure. Today, natural and continuous SL recognition is an open research problem due to multiple spatio-temporal shape variations, challenging visual sign characterisation, as well as the non-linear correlation among signs to express a message. A compact sign is introduced to text architecture that explores motion as an alternative to support sign translation. Such characterisation results are robust to appearance variance with relative support to geometrical variations. The proposed representation focuses on the main spatio-temporal regions to each corresponding word. The proposed architecture was evaluated in a built SL data set (LSCDv1) dedicated to motion study and also in the state-of-the-art RWTH-Phoenix. From the LSCDv1 data set, the best configuration reports a BLEU-4 score of 63.04 in a testing set. Regarding RWTH-Phoenix, the proposed strategy achieved a BLEU-4 score in a test of 4.56, improving the results under similar reduced conditions.

1 | INTRODUCTION

Sign language (SL), a visual-gesture-based system, is the main mechanism of communication for the deaf community. According to the Ethnologue list [1], there are more than 144 official SLs with total independence on grammar and lexicons. In fact, almost each country has its own SL which can vary drastically from one region to another. The lack of knowledge of each specific SL is one of the main reasons for the deficiency in the inclusion of the deaf community in society, leading to many limitations in access to services, which in most cases is almost non-existent. This dramatic fact is mainly related to the absence of interfaces that easily translate from deaf languages to spoken or written languages. Worse still, such limitations are present among the deaf in different regions due to the marked differences between the SLs. Thus, nowadays it is essential, but challenging, to develop technological support for automatic translation between deaf people and the rest of society.

Technically, SL is represented as a set of visual spatio-temporal gestures, which can be represented in their written form through fundamental communication units named glosses. These communication components can represent simple words, expressions, and phrases with complex grammatical structures or even complete concepts [2]. Remarkably, like any natural language, the unlimited lexical richness of SL, together with the high sign variability, constitute the main challenge to automatically understanding the visual-manual utterances of the articulators. Additionally, these glosses can be developed in different video lengths, and entail non-linear time sign relationships. In addition to this, capture setup variations or slight sign geometric changes can affect the representation.

Regarding state-of-the-art, classical approaches that have been dedicated to proposing naive gesture recognition strategies over isolated and non-natural signs, such works have mainly been based on hand-craft features that mainly code the static shape of signs to find isolated and global word correspondences [3–5]. These approaches, however, lose the temporal capability to recognise gestures in more realistic scenarios. Alternatively, approaches based on hidden Markov models (HMMs) [6,7] have been proposed to model sign changes during sequences. These approaches exploit appearance and shape sign observations that together with temporal modelling find a sign-text correspondence [8]. Nevertheless,
these approaches are based on the hypothesis of an almost consecutive temporal dependence on signs, which leads to a false 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 [9–11]. These approaches have represented a significant advance in the introduction of sign language translation (SLT) systems in real-life scenarios. Camgöz et al. introduced a neural encoder–decoder architecture, into the SLT context, with the most promising results [12]. Nevertheless, such approaches require complex hyperparametric schemes due to insufficient 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. Taking into account this dynamic component allows to stratify the communication components, as well as to differentiate among the transition of gestures [13].

The main contribution of this research is an SLT deep encoder–decoder that takes advantage of motion glosses representation at different levels and under different coding schemes, to represent SLs. From motion representation, the proposed approach results are compact (50 million fewer parameters than standard nets of state-of-the-art) and invariant to multiple appearance sign variations. In summary, the specific scientific contributions are:

- A 3D convolutional optical flow representation which captures spatial kinematic patterns, including typical sign large displacement features from videos. This motion representation, coded in the first module of the architecture, was integrated and learnt together with the translation architecture. In this way, only sign-related motion primitives are retrieved from the video sequences.
- A second level of motion analysis that finds non-temporal relationships between the main video descriptors using recurrent neural units. These modules are bidirectional in time and process embedding features extracted from a low-level coding. The bidirectional modules allow the modelling of complex grammatical rules, such as the relationship of the subject's gender (i.e. at the beginning of the sentence) with the object located at the end of the sentence. Furthermore, significant non-linear changes are reported if the sentences are in the past or in the future. In such a case, the initial information of the sentences could change with respect to the final signs (signs indicating verbal tense).
- An analysis and evaluation of the gestural attention layer position by determining the main spatio-temporal descriptors that enhance non-trivial correlations with spoken language units.
- The architecture and the multi-level motion representation allows the integration of all components in a low-complexity deep encoder–decoder network with only ~15 million parameters (about 70% less than standard networks in this task). This architecture was trained by synchronising both corresponding languages, that is, gestures and corresponding text. The experimental results show that using the compact motion representation allows reducing the architecture to the use of only two recurrent layers (normally at least four are used), achieving stable and coherent results.

The proposed multi-level motion architecture was evaluated with respect to translation capability, on two different data sets: Our motion-dedicated sign data set from Colombian SL (LSCDv1) and also in the RWTH-Phoenix state-of-the-art data set. The LSCDv1 corresponds to Colombian deaf signs, focussed on simple and well-formed structures that allow a closer analysis of temporal information (motion), with 510 video signs developed by a total of 11 signers. The other method, the RWTH-Phoenix, is a German large SL data set composed of 8257 uncontrolled videos, recorded from a weather TV programme.

Section 2 describes the main related approaches focussed in SLT. Section 3 introduces the proposed end-to-end architecture that fully exploits motion information. Section 4 summarises the experimental setup and main configuration of the proposed architecture. Section 5 presents the results and a quantitative evaluation of the method. Sections 6 and 7 discuss remarkable issues of the proposed strategy with respect to the state-of-the-art methods and present the conclusions and prospective works.

## 2 RELATED WORK

Current advances in comprehensive and dense learning approaches, together with robust convolutional representations, have allowed for going beyond traditional recognition tasks. For instance, the naive detection of isolated signs has evolved to focus on more natural and realistic scenarios, covering challenges such as continuous sign language recognition (CSLR). Initially, from the statistical systems, the sequential modelling of the visual reference points has exploited first-order Markov dependencies among signs to approximate CSLR [6,14]. For instance, Cooper et al. studied several 2D and 3D visual features temporarily modelled with explicit Markov models and a sequential pattern boosting [6]. Also, Koller et al. coded SL as an HMM to model a temporal sign sequence observed from hand features and a set of face reference points [14]. These approaches have been useful to understand the importance of temporal relationships between signs, but their assumptions, most of the time, prove insufficient in modelling high-order dependencies of language.

Recently, robust visual representations obtained from convolutional neural networks (CNNs) have been integrated into sequence models such as HMMs, to achieve a more robust and continuous recognition of SL [7,8,15]. These CNN-HMM approaches improved the visual description due to the discriminatory CNN properties, allowing a better temporal prediction of the corresponding sign sequences. Nevertheless, these approaches underlie Markovian assumptions, modelling only neighbourhood 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 (RNNs) \[9,16\]. These approaches learn long-term dependencies, using for instance LSTM and GRU units, from a large amount of matching information between sign 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 \[17\]. From CTC “sequence-to-sequence” learning strategies were introduced with the main advantage to operate over weakly labelled sign videos \[10,11\]. However, CTC has almost no inference on visual sign modelling and both the structure and grammar of utterances are poorly exploited.

Based on CTC limitations, new advanced strategies have been faced with sign language translation (SLT) by using an encoder–decoder architecture with RNN units between signs and text \[12\]. 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. Similarly, Huang et al. proposed a hierarchical attention-based encoder architecture that characterises video from a two-stream convolutional neural network \[18\]. Also, Guo et al. proposed a hierarchical-LSTM (HLSTM) encoder–decoder model that combines a 3D-CNN video description with multi-layered connected LSTM units to achieve subvisual words, words and complete video recognition \[19\]. A clear limitation of these approaches is the sign description from appearance information, which requires exhaustive and dense learning to extract the main related information. As an alternative, Ko et al. introduced a strategy to model signs as randomly selected human body poses from 124 key-points and using the same encoder–decoder configuration for high processing levels \[20\]. Nevertheless, these poses could be sensitive to spatial changes and the random selection of poses to represent the structure of the language representation. Despite current encoder–decoder advances, the architectures require a huge quantity of parameters to learn sign representation, which results in complex computational approaches. Therefore, motion processing at different scales can reduce architecture complexity and contribute to temporal sign modelling.

3 | 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 \[13,21\]. Motion is a fundamental, but poorly analysed, component of SL, the results of which are communicative by itself and determines transitions between gestures contributing to the grammatical structure of language. For instance, shape-motion units along the video $U = (u_1, u_2, ..., u_K)$ could robustly represent sign transitions in a video sequence. In such a case, the translation of a text sequence, $y = (y_1, y_2, ..., y_M)$ with $M$ words, could be expressed as the conditional probability $p(y|U)$. A compact recurrent encoder–decoder network is presented here that solves this conditional probability by exploiting motion and temporal information at different levels.

First, a low-level representation of signs based on the optical flow allows learning long-term dependencies on the signs followed by a 3D-CNN architecture. Then, this 3D-CNN-movement representation contains robust kinematic components that facilitate sequential learning using an RNN that propagates and represents the global temporal behaviour of the sequence. This resulting RNN vector is decoded to generate word sequences at the network output, that is, translation of the complete video. Additionally, an attention model is herein included to highlight local temporal patterns that mainly contribute to word translation. This mechanism allows finding complex higher order temporal relationships between the sequence modes. Then, the architecture break downs the sequential learning problem into two specialised RNN modules supported by attention models. In Figure 1 is illustrated the general pipeline of the proposed approach.

3.1 | A first motion level: convolutional representation of flow velocity volumes

A low-level model of motion shape is introduced here 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. Two main assumptions were considered to satisfy proper SL modelling: (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.

On the first consideration, a dense optical flow that considers large coherent motion displacements was used herein \[22\]. This optical flow faces typical assumptions of very small displacements to recover proper smoothed fields. The velocity field $u_k := (u_{x_k}, u_{y_k})^T$, for a particular frame $k$ is obtained from a variational Euler–Lagrange minimisation that includes local and non-local restrictions between two consecutive frames: $I(x_k, I(x_{k+1})$. To capture large displacements, a non-local assumption is introduced by matching key-points with similar velocity field patterns. This final assumption could be formalised as: $E_p(u) = | g_{k+1}(x + u(x)) - g_k(x) |^2$ where $p$ is the descriptor vector and $(g_{i(t)}, g_{i(t+1)})$ are the computed velocity patterns in matched non-local regions. The captured flow field volume result is highly described, keeping spatial coherence and aggregating motion information patterns as a low-level representation. 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.
Secondly, a 3D-CNN architecture learns the long-term dependencies present between the flow velocity fields: \( \mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_N) \) with \( N \) being the number of total frames. As proposed by Varol et al., the implemented architecture uses 3D convolutions to deal with temporal dependencies on entire volumes capturing the most relevant features of the sign [23]. A total of six 3D convolutional blocks allow a volumetrical decomposition of the sign sequence. Then, the model learns linear transformations hierarchically, followed by ReLu activations, allowing enhancement of the main local spatio-temporal patterns correlated non-linearly with the glosses. This motion representation progressively computes linear transformations, projecting the final information on a set of \( \mathbf{Z} \) high-level learnt filters. Then, the architecture returns a motion convolved representation matrix \( \mathbf{F} = \{f_1, \ldots, f_Z\}, f_z \in \mathbb{R}^d \), that represents the \( d \)-dimensional \( \mathbf{Z} \) filters in the last convolutional layer. The resulting representation recovers spatio-temporal patterns throughout the video. This module is fully integrated with the encoder-decoder scheme, allowing to learn a discriminative representation in concordance with associated translations. Figure 2 illustrates the 3D-CNN scheme.

3.2 A second motion level: the bidirectional motion encoder

SL sequences have very complex compositions that depend, among other things, on particular grammatical structures, signers’ habits, or regional compositions. For instance, interrogative sentences have a strong non-linear correspondence between the beginning and end of the sentence. A motion encoder is then designed to compute temporal non-linear correlations in both sequential modes: video and text. A recurrent multi-layer architecture was then implemented to propagate the shape-motion representation \( \mathbf{F} \). This architecture allows describing the global temporal behaviour of a particular video sequence. A total of two bidirectional network layers (BiRNN) form the architecture, which together obtains a high-level temporal sign description by computing hidden states in forward and backward directions.

In the first layer, the deep motion features \( \mathbf{F} = \{f_1, f_2, \ldots, f_Z\} \) are sequentially propagated by a set of recurrent units, which compute the states: \( \mathbf{b}_z = q_1(f_z, \mathbf{b}_{z-1}) \). 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: \( [b_{1:Z}, b_{Z:1}] \). An additional second layer was designed 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 propagates temporal information as: \( b_z = q_2(b_z, \mathbf{b}_{z-1}) \). In this layer, the propagation is also performed into a bidirectional scheme, with the resulting concatenated vectors: \( [b_{1:Z}, b_{Z:1}] \) where \( q_1, q_2 \) are the activation functions. The decoder uses the last encoder output vectors \( b_{Z:1} \) as initialisation vectors for each decoder layer \( l \). It should be noted that in this encoder
representation, the decoder receives both layer representations, which enriches the description of motion sign translation and helps in the text generation process.

### 3.3 | 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 \((b_1^2, b_2^2)\), computed at each BiRNN layer. These observed encoder motion vectors describe the kinematic sign history at different time dependence levels. Then, the decoder could be modelled by decomposing the joint probability in ordered conditional probabilities, expressed as follows:

\[
p(y | u) = \prod_{m=1}^{M} p(y_m | \{y_{m-1} : y_1\}, (b_1^2, b_2^2))
\]

where each predicted \( y_m \) word is conditioned by previous predictions \( \{y_{M-1} : y_1\} \) together with observed encoder vectors \((b_1^2, b_2^2)\). This conditional probability is solved by integrating a unidirectional recurrent network with a motion attention mechanism. The network was herein considered with two layers. From such integration, it was possible to relate both language modes. The unidirectional network layers preserve encoder recurrent unit dimensions. In such a way, it is possible to initialise the hidden states \((h_0^1, h_0^2)\) (on decoder) with \(b_2^1\) (on encoder). The attention mechanisms \(\mathcal{A}_m\) were implemented in both decoder layers to associate motion-encoder representations with the associated translation word at different levels of representation. Each motion attention model \(\mathcal{A}_{m}^l\), at a particular layer \(l\), relates all encoder hidden states \(b_{1,2}^l\) to the hidden representations of each word in the decoder through a particular context vector \(c_m^l\). This relationship is defined as the concatenation of current context vectors and hidden states, as \(\mathcal{A}_{m}^l = [c_m^l, h_m^l]\). In this case, the context vector \(c_m^l\) computes the weights of all encoder motion vectors \(b_{1,2}^l\) with respect to each \(m\) translated word, which could be expressed as:

\[
c_m^l = \sum_{z=1}^{Z} \gamma_m^l b_z^l
\]

where \(\gamma_m^l\) are the attention weights, that define the relevance of a particular encoder input descriptor \(f_z\) to generate the \(y_m\) 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 \(h_m^l\) against each encoder output \(b_z^l\) as:

\[
\gamma_m^l = \frac{\exp(h_m^l W b_z^l)}{\sum_{z=1}^{Z} \exp(h_m^l W b_z^l)}
\]

where \(h_m^l W b_z^l\) is the scoring function and \(W\) are the learnt parameters used to match the temporal descriptors with the generated words. The hidden states of the decoder are then computed as follows:

\[
h_m^{(1)} = q_3(e_m, h_m^{(1)})
\]

\[
h_m^{(2)} = q_4(A_m^{(1)}, h_m^{(2)})
\]

where \(e_m = \text{Embedding}(y_m)\) and \((q_3, q_4)\) are the activation functions. The final prediction word \(y_m\) is given by \(y_m = \text{softmax}(W_A A_m^{(2)})\), where \(W_A\) are the learnt parameters of the fully connected layer. The embedding layer transforms the one hot encoding words of the written language into a dense representation which relates words with semantic components.

For each predicted word \(y_m\), the decoder uses the previous word and hidden states \(\{y_{m-1}, h_m^{(1)}, h_m^{(2)}\}\) to update the next hidden states \(h_m^{(2)}\). Then, to start the sentence generation process, the \(y_0\) word is the special token \(<\text{bos}>\) 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 the main patterns that contribute to a specific word translation. The code is available in this repository:\(^1\)

\(^1\)https://github.com/bivl2ab/research/2018-jefferson-slt/how-important-is-motion-in-sign-language-translation
4 | EXPERIMENTAL SETUP

The proposed approach was evaluated in two different data sets dedicated to SLT, described as follows.

4.1 | LSCDv1: Colombian sign language data set

The LSCDv1 introduced herein focuses on continuous signs with utterances involving motion as key mechanisms of language. This data set is part of the real and common signs of the Colombian SL that has corresponding global Spanish-text translations for complete sequences. A total of 15 well-formed affirmative sentences (subject, object, verb) were recorded by 11 signers (nine deaf and two children of deaf parents, with an average age of 35 ± 17), with three repetitions for each sentence, and an average sentence length of five words. The complete data set has 510 videos, for a total of 143,103 frames, with a vocabulary of 64 words and ~90 different signs. Each video was recorded with a spatial resolution of 224 × 224 and a temporal resolution of 60 fps. In such videos, strong variations were observed among signers due to cultural aspects, education level, age differences, and the alternative expressions specific to each language. Figure 3 illustrates some sign examples of captured sequences. This study was approved by the Ethics Committee of Universidad Industrial de Santander in Bucaramanga, Colombia, with number D19-13353. Written informed consent was obtained for every participant. The data set is available on this website².

To validate the proposed approach, the data set was split into training and testing sets. The authors chose 464 videos for training and 46 videos for testing. The splits were made to ensure that all sentences were evaluated and there were no repetitions of the same signer in training and testing for a particular sentence.

4.1.1 | Data augmentation:

To enrich data representation during the training process, data augmentation was carried out over LSCDv1 by following the strategies of frame-sampling and horizontal flipping to emulate changes on the signer’s dominant hand. Over time, the frame sub-sampling was performed to cover whole message representation, using a dynamic step rate and fixing the same length for the whole data set. For videos with an original length less than the fixed parameter, oversampling was defined by introducing new frames as a result of averaging a pair of consecutive images. Using these strategies, the data set was augmented from 3.4 to 17.7 times, resulting in a data set with 4000 video clips sequences.

https://bivl2abgitlab.io/research/2020-juan-colombiansld/lscd/

4.2 | RWTH-PHOENIX-weather 2014T

To exhaustively evaluate the proposed approach, the state-of-the-art RWTH-PHOENIX-Weather 2014T data set was also considered in this work. This data set records sequences that correspond to German SL with a total of nine signers that explain weather news on local TV. The vocabulary in such a data set is composed of 1066 signs that correspond to 2887 words in the German spoken language. The data set 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 data set has been widely used on current SLT strategies because of the rich video information together with challenges of variability of signs recorded at each sequence [7,11,12].

4.3 | Configuration of proposed architecture

The main goal of this work is to analyse the motion contribution on SLT and how the proposed deep architecture is able to recover such patterns and use them to translate into the respective spoken languages. Taking into account this fact, extensive validation of proposed architecture, together with an exhaustive parameter search was performed for LSCDv1. Thereafter, the best-achieved configuration of the proposed
architecture was selected to evaluate the RWTH-Phoenix data set. A complete description of the architecture components is as follows:

- Motion CNN representation: This first motion layer is composed of an architecture of six space–time convolutional layers followed by three successive fully connected layers. Every filter is obtained from a (3 × 3 × 3) kernel with a stride of 1 for all dimensions. Also, batch normalisation, ReLU activation, and max pooling operation with a kernel size of (2 × 2 × 2) with a stride of 2 are applied to the volumes resulting from the space–time convolution. In Figure 2 is shown a complete description of the implemented architecture. In contrast to the original 3D-CNN architecture, the authors implement a batch normalisation through temporal channels dimension, which allows accelerating the learning process from a higher learning rate.
- Encoder architecture: For this second motion level, as mentioned above, the authors only use two layers of BiRNNs with tanh recurrent activation functions. Then, to keep the encoder and decoder fully connected, the first layer has a total of N = 128 neurons, while for the second layer the total number of neurons results from the relation: 2N^{lid}.
- Decoder architecture: Unlike the encoder module, the decoder does not use bidirectional networks but is also composed of only two recurrent layers 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, the authors use general attention modules [24]. 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.

4.3.1 Learning parameters

A stochastic gradient descent was herein implemented (with mini-batches of size 1 because of GPU limitations) and a weighted categorical cross-entropy loss function (with zero weight for the padding token). For each experiment, a learning rate of 0.001 was used with a learning rate decay of 0.1 and a dropout of 0.4 for faster convergence. The momentum was set to 0.99 and the convolutional weight decay was set to 0.0005. The gradient clipping was also used with a threshold of 5 and 10 epochs used for LSCDV1 and 20 for RWTH-Phoenix data set.

4.3.2 Metrics for evaluation

Three classical metrics for translation problems were herein used to validate the proposed approach: ROUGE-L (Longest Common Subsequence) [25], BLEU score [26] and METEOR score [27]. The ROUGE-L takes into account similarity regarding sentence structure and identifies longest co-occurrence in compared n-grams sequences. The BLEU score measures precision to recover a set of consecutive n-grams. The METEOR is related to the harmonic mean of unigram precision and recall.

5 EXPERIMENTAL EVALUATION

An exhaustive evaluation of the proposed SLT architecture was carried out over the LSCDV1. Each of the motion coding and translate components was evaluated to analyse their contribution to the challenging sign translation task. Thereafter, the best configuration of the proposed architecture was evaluated over the RWTH-Phoenix. In such a sense, the generality and compactness of the approach were evaluated in this challenging data set, highlighting the value of motion representation in the SLT task.

Firstly, the flow motion shape representation 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 red, green, blue (RGB) raw video information. A total of K = 128 frames were used as input to the network and Z = 128 filters were fixed on the last 3D CNN layer. For this first analysis, the authors selected a recurrent network with GRU units for second motion processing. Also, attention units were positioned at the top and middle of the decoder to obtain different single configurations. In addition, a double attention was configured using the top and middle attention modules at the same time. In Table 1 the obtained results are summarised for both sequences: flow and RGB, either for training or test sets. As expected, the motion-shape descriptor improves translations compared to RGB descriptors with a remarkable difference in all experiments. Specifically, in tests, 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.

To analyse temporal embedding patterns, the recurrent units for motion representation were also modified by LSTM units. The LSTM scheme, fitted with feedback connections, allows coding of motion patterns with separate temporal relationships. Table 2 reports the results obtained during this experiment for training and test sets. In such a configuration, the flow sequences also achieved the best performance, compared to RGB input sequences. A remarkable result of such configuration is the contribution of LSTM architecture to the second level of motion. In such a case, this architecture is 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 at the middle of the architecture obtains the best result. In general, a gain of 10.78% was then obtained for the bleu-4 score in a test compared to RGB input sequences. A remarkable result of this configuration was selected for more comprehensive and exhaustive element-architecture evaluations. Figure 4 shows the performance (Blue-4 in the y-axis) of this architecture for a different number of input frames (x-axis), but varying the number of filters (stylish lines) and attention modules (coloured lines), for training (left panel) and testing sets (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 the other best results obtained for 60-frame and 30-frame video inputs, respectively. This result was also obtained using top attention and 60 filters. A compact representation was obtained from 60 filters compared to representations generated from 128 and 256 CNN filters, representing gains of 8.7 and 10.86, respectively. Clearly, the 60 filters and 128 frames configuration allows the analysis of a larger amount of coded information from a more compact group of filters, reducing noise from non-zero activation responses. In addition, the top attention has a gain of 10.87 and 4.35 with respect to middle and double attention, respectively.

This result could be associated with the temporal dependence of words at the top of the architecture. Also, at a low-level there is a major variability that is difficult for a proper correspondence with words at the semantic level.

A further evaluation of the proposed approach was carried out also over the RWTH-Phoenix data set, which includes less controlled scenarios and involves both static and motion linguistic expressions obtained from a German television channel. This data set has been evaluated with a very large architecture that considers more than 63 million parameters for training and achieve a BLEU-4 score in a testing set of 9.56 [12] (S2T). Specifically, the S2T architecture has four RNN layers, 1000 neurons in each layer, while the proposed approach only considers two recurrent layers and less than 300 neurons. For this experiment, the authors wanted to evaluate the ability of both strategies to incorporate motion into the translation process. For a fair comparison due to GPU limitations, they reduced the parameters of the S2T approach to the same number of parameters as the proposed strategy. Table 3 summarises the obtained result for both strategies using as input flow sequences. The authors’ approach reports a competitive result on this data set, achieving a bleu 4 score in the testing set of 4.56 considering the compactness of architecture with 15 million parameters. The reported bleu score resides on the complexity of this data set expressed on high variability of the signs motion and the inclusion of static gestures with representative semantic knowledge. However, the main issue is analysing the contribution of motion representation for SL, which result is relevant but a modelling of sign geometry will be required.

6 | DISCUSSION

A fully SL motion modelling was herein introduced by using a deep end-to-end low-complexity neural architecture (~15 million of parameters). This architecture was able to model, analyse and measure the motion patterns capability to support automatic text translation of video sign sequences. The motion was here designed at different levels and analysed hierarchically up to the dynamic text spoken translation. The proposed approach was firstly evaluated over a Colombian SL data set (LSCDv1), dedicated to model well-formed phrases where motion is determinant on gestural information. Also, the RWTH-Phoenix state-of-the-art data set was here used to validate and compare the proposed strategy.

The proposed network is composed of two main modules. The first module (Encoder) was designed to extract and code spatio-temporal 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 characterise the limb movements in conventional cameras. The resulting sequences serve as input to a deep 3D-convolutional network to extract the main 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 organised 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, at a local level. To achieve optimal training convergence, tolerance to high learning rates, and better feature discrimination, batch normalisation is followed in each layer followed by ReLu activations [28,29]. During the evaluation, these kinds of sequences outperformed experiments regarding raw RGB sequences, with a strong difference of 43.88 on BLEU-4 score average, on LSCDv1. Afterwards, a bidirectional recurrent neural network was then used to identify the highly recurrent temporal relationships present in these extracted patterns. Hence, from the decoder (second module) is received as input of these recurrent motion descriptors and correlated with individual words through gestural attention units. The motion descriptors are decoded using a unidirectional recurrent network. For the entire network, the authors analyse 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. Likewise, temporal attention modules were evaluated at different levels of decoder architecture to correlate encoder descriptors with the intermediate and higher words vectors. They found that the attention module at top-level reports the best performance in their scheme with a gain bleu-4 of 10.87 with respect 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 for some sub-task such as continuous independent sign recognition [7,10,11,30]. Some of these approaches have tried to model temporal sign associations using a connectionist temporal CTC [17]. However, the structure and grammar of the utterances are poorly modelled. Despite the demonstrated importance of the motion patterns in language to include temporal connections, almost all approaches are fully based on the characterisation of isolated gestures or on pure appearance information. For instance, one of the most salient works implemented a four-layer recurrent encoder–decoder network using raw RGB input sequences, in an architecture with 1000 neurons in each layer and more than 63 million of parameters (S2T) [12]. In that work, motion patterns are present in the descriptors obtained 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 may be due to the feature extraction process at frame level using 2D convolutions. In contrast, this work introduces an encoder–decoder framework dedicated to analysing the contribution of motion in SL, demonstrating the powerful and compact representation over a motion-dedicated LSCDv1 data set. An additional advantage of the motion representation is the compact design of the architecture, which allows the use of fewer parameters and reliable results in real applications. Indeed, the authors developed an additional experiment reducing the S2T architecture to obtain the same number of layers and approximately the same number of neurons and parameters as this strategy. Then, these architectures were compared over an RWTH-Phoenix data set using flow images.
to analyse motion propagation. As expected, the motion input was sufficient for this approach to achieving a similar performance with respect to S2T. Interestingly enough, the best configuration of the proposed architectures improves bleu-2,3,4 scores in tests with a gain of 0.8, 0.31 and 0.06 with respect to the reduced S2T method. In such a line, it can be argued that motion inclusion provides salient clues to improve the translation of two, three and four consecutive words (n-grams). Differences between results achieved in LSCDv1 and RWTH-Phoenix data sets can be attributed to dictionary size, that is, the number of words, that for LSCDv1 has 90 signs corresponding to a vocabulary of 64 words, while in the RWTH-Phoenix data set there are 1066 signs and a vocabulary of 2887 words.

An additional state-of-the-art approach uses pose estimation in the input, coded as normalised feature vectors, and modelled into an encoder–decoder architecture [20]. This architecture was used in an overall two-layer recurrent encoder–decoder network of 256 GRU units. This architecture result is interesting because of the sparse representation obtained from poses that help with variability among poses. Nevertheless, the poses are dependent on spatial transformations that could be prohibitive in real scenarios. This approach was validated on a non-public Korean sign language data set with 11,578 videos, 105 sentences and a vocabulary of 419 words. The validation of this approach was only done with simple a unigram (1-gram) metrics that limits the analysis of more complex relationships.

The proposed approach is able to exploit motion representation, from non-controlled video sequences, and translate sentences of a natural SL. Motion representation 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 on the large data set. The main hypothesis of that limitation is the very compact low-level representation, learnt by the 3D-CNN, designed with only 60 filters in the last layer. In Figure 4 is illustrated such a distribution. In this case it was necessary to find a trade-off between the density in the representation and the set of videos available to learn motion information. From the current version, the proposed approach will work properly in a specific domain but will require a larger 3D-CNN architecture to a more general domain. For example, Figure 5 illustrates how attention is distributed over the filters for each word, but some of the words point to the same motion-activation filters. This fact suggests that a richer representation of SL obtained from a more discriminating representation base will be necessary.

7 | CONCLUSIONS

Advantage of a hierarchical encoder–decoder deep representation is made herein, with the introduction of special modules to model motion shape patterns in SL for continuous translation of video sequences. A total of three encoder–decoder levels were implemented to analyse motion from the instantaneous spatio-temporal descriptor to the long non-linear history of sign sequences. First, a low motion representation was obtained from a 3D convolutional strategy using optical flow volumes that support special large displacements as an input. The motion embeddings, from the 3D architecture, were able to capture kinematic patterns, invariant in appearance, and represent robustly the signs for translation. Then, a set of recurrent bidirectional units propagate motion embeddings over time, capturing history on long-term temporal scales. From these two levels of representation, a visual attention module, on a higher level of representation, allows the text to be correlated 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 representation. The proposed approach was evaluated on a dedicated motion Colombian SL data set, from which the optical flow embeddings outperform raw sequences. Also, the strategy was evaluated on a more general data set 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.

ACKNOWLEDGEMENTS
This work was partially funded by the Universidad Industrial de Santander. The authors acknowledge the Vicerrectoria de Investigación y Extensión (VIE) of the Universidad Industrial de Santander for supporting this research registered by the project: Identificación de Características Asociadas con Pleomorfismo Nuclear en Imágenes Histológicas de CÁncer de Mama Utilizando Algoritmos de Aprendizaje Profundo, with SIVIE code 2692. They also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

ORCID
Jefferson Rodriguez https://orcid.org/0000-0003-2394-5683
Fabio Martínez https://orcid.org/0000-0001-7353-049X

REFERENCES
1. Eberhard, D.M., Simons, G.F., Fennig, C.D.: ‘Ethnologue: Languages of the World’. Ethnologue. https://www.ethnologue.com/subgroups/sign-language (2019). Accessed 13 August 2019
2. Stokoe, W.C.: Sign language structure. Annu. Rev. Anthropol. 9(1), 365–390 (1980)
3. Zabedh, M., Keysers, D., Ney, H.: Appearance-based recognition of words in American sign language. In: Proceedings of Iberian Conference on Pattern Recognition and Image Analysis, 511–519 Springer, Estoril (2005)
4. Zaki, M.M., Shaheen, S.I.: Sign language recognition using a combination of new vision based features. Pattern Recogn. Lett. 32(4), 572–577 (2011)
5. Rodríguez, J., Martínez, F.: A kinematic gesture representation based on shape differences vlad for sign language recognition. In: International Conference on Computer Vision and Graphics, 438–449 Springer, Warsaw (2018)
6. Cooper, H., et al.: Sign language recognition using sub-units. J. Mach. Learn. Res. 13(Jul), 2205–2231 (2012)
7. Koller, O., et al.: Deep sign: enabling robust statistical continuous sign language recognition via hybrid cem-hmms. Int. J. Comput. Vis. 126, 1311–1325 (2018)
8. Koller, O., et al.: Deep sign: hybrid cem-hmms for continuous sign language recognition. In: Proceeding of the British Machine Vision Conference 2016, pp. 438–449 BMVA Press, York (2016)
9. Cui, R., Liu, H., Zhang, C.: Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7361–7369. IEEE/CVF, Honolulu (2017)
10. Necati, C.C., et al.: Subunets: end-to-end hand shape and continuous sign language recognition. In: Proceeding of IEEE International Conference on Computer Vision, pp. 3075–3084 IEEE/CVF, Venice (2017)
11. Wang, S., et al.: Connectionist temporal fusion for sign language translation. In: Proceeding of Multimedia Conference on Multimedia Conference, pp. 1483–1491. ACM, Seoul (2018)
12. Carrero, N.C., et al.: Neural sign language translation. In: Proceeding of Conference on Computer Vision and Pattern Recognition, pp. 7784–7793 IEEE/CVF, Salt Lake City (2018)
13. Liddell, S.K., Johnson, R.E.: American sign language: the phonological base. Sign Lang. Stud. 64(1), 195–277 (1989)
14. Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: towards large vocabulary statistical recognition systems handling multiple signers. Comput. Vis. Image Understand. 141, 108–125 (2015)
15. Koller, O., Zargaran, S., Ney, H.: Re-sign: re-aligned end-to-end sequence modelling with deep recurrent cem-hmms. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4297–4305. IEEE/CVF, Honolulu (2017)
16. Ye, Y., et al.: Recognising American sign language gestures from within continuous videos. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2064–2073 IEEE/CVF, Salt Lake City (2018)
17. Graves, A., et al.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceeding of the 23rd International Conference on Machine Learning, pp. 369–376 ACM, New York (2006)
18. Huang, J., et al.: Video-based sign language recognition without temporal segmentation. In: Proceeding of Conference on Artificial Intelligence, pp. 17137. AAAI, New Orleans (2018)
19. Guo, D., et al.: Hierarchical lstm for sign language translation. In: Proceeding of Conference on Artificial Intelligence, p. 16177. AAAI, New Orleans (2018)
20. Ko, S., et al.: Neural sign language translation based on human keypoint estimation. arXiv preprint arXiv:181111436 (2018)
21. Sandler, W., Lillo.Martin, D.: Natural sign languages. In: The handbook of linguistics, pp. 533–562 (Blackwell, Oxford 2001)
22. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. IEEE Trans. Pattern Anal. Mach. Intell. 33(3), 500–513 (2010)
23. Valol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40(6), 1510–1517 (2017)
24. Luong, M., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceeding of Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421. Association for Computational Linguistics, Lisbon (2015)
25. Lin, C.: ROUGE: a package for automatic evaluation of summaries’. In: Text summarisation branches Out, pp. 74–81. Association for Computational Linguistics, Barcelona (2004). https://www.aclweb.org/anthology/W04-1013
26. Papineni, K., et al: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Association for Computational Linguistics, pp. 311–318 (Philadelphia, Pennsylvania, USA 2002)
27. Banerjee, S., Lavie, A.: Meteor: an automatic metric for Mt evaluation with improved correlation with human judgements, pp. 65–72 (2005)
28. Thakkar, V., Tewary, S., Chakraborty, C.: ‘Batch normalisation in convolutional neural networks—a comparative study with cifar-10 data’. IEEE, pp. 1–5 (2018)
29. Ioffe, S., Szegedy, C.: Batch normalisation: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:150203167 (2015)
30. Pu, J., Zhou, W., Li, H.: Dilated convolutional network with iterative optimization for continuous sign language recognition. In: Proceeding of International Joint Conference on Artificial Intelligence, pp. 885–891. IJCAI, Estocolmo (2018)

How to cite this article: Rodríguez J, Martínez F. How important is motion in sign language translation? IET Comput. Vis. 2021;15:224–234. https://doi.org/10.1049/cvi2.12037