A Kinematic Gesture Representation Based on Shape Difference VLAD for Sign Language Recognition

Jefferson Rodríguez[1,2,3] and Fabio Martínez[1,2,3]

[1] Grupo de investigación en ingeniería biomédica, (GIIB)
[2] Motion Analysis and Computer Vision (MACV)
[3] Universidad Industrial de Santander (UIS)
Bucaramanga, Colombia.
{jefferson.rodriguez2,famarcar}@saber.uis.edu.co

Abstract. Automatic Sign language recognition (SLR) is a fundamental task to help with inclusion of deaf community in society, facilitating, noways, many conventional multimedia interactions. In this work is proposed a novel approach to represent gestures in SLR as a shape difference-VLAD mid level coding of kinematic primitives, captured along videos. This representation capture local salient motions together with regional dominant patterns developed by articulators along utterances. Also, the special VLAD representation allows to quantify local motion pattern but also capture shape of motion descriptors, that achieved a proper regional gesture characterization. The proposed approach achieved an average accuracy of 85, 45% in a corpus data of 64 sign words captured in 3200 videos. Additionally, for Boston sign dataset the proposed approach achieve competitive results with 82% of accuracy in average.

Keywords: Motion analysis · sign recognition · mid-level representation · VLAD representation.

1 Introduction

Deaf community and people with some auditive limitation around world is estimated in more than 466 millions according to world health organization (WHO) [3]. This community had achieved to structure different natural sign languages as spatio-temporal gestures, developed by articulated motions of upper limbs that together with facial expressions and trunk postures allows communication and interaction. Sign languages are as rich and complex as any spoken language, being the automatic recognition a tool that allows to include deaf community in society. Nevertheless, the automatic interpretation of deaf languages remains as an open problem because the multiple inter and intra signers variations and also external variations produced by culture, history and particular interpretations according to regions. Such variations implies great challenges to understand and associate semantic language labels to spatio-temporal gestures. Also, typical video processing problems are present such as illumination changes,

perspective of signers, occlusion of articulators during the deaf conversation, among many others. Such related problems limits the use of automatic learning methodologies, the usability of multimedia tools and puts deaf community in disadvantage to explore much of information in multimedia platforms.

The sign recognition has been addressed in literature by multiple approaches that include global shape representations that segment all articulators of language but with natural limitation due to occlusions and dependences of controlled scenarios [17]. Other strategies have developed analysis of gestures from local representations that include the characterization of interest points [11, 16] and the analysis of appearance and geometrical primitives to capture shape of gestures in videos [12]. For instance, Zahedi *et. al.* [19] proposed a SLR by computing descriptors of appearance that together with gradient of first and second order characterize particular signs. Such approach is however dependent of signer appearance and perspective in video sequence. An extension of this work developed the analysis of multi-modal information to recover shape from RGB-D sequences and also compute trajectories from accelerometers to complement sign description [17]. Despite the advantages of 3D analysis the depth sequences are limited to controlled scenarios and the external accelerometers can alter the natural motion of gestures.

Motion characterization have been fundamental to develop strategies to recognize gestures being robust to appearance variance and illumination changes [10,11]. For instance, in [10,16] utterance sequences were characterized from firs order relationships of appearance velocities captured from Lukas-Kanade motion field. This approach is however prone to errors because the flow sensibility to little camera displacements and also the sparse nature of the approach capture few displacement points that difficult any statistical analysis. In same line, Jakub Konecný *et. al.* [10] integrates local shape information with histograms of an optical flow to describe gestures. This approach achieved a frame-level representation but lose local and regional information to represent gestures. Jun Wan *et. al.* [16] proposed a dictionary of sparse words codified from salient SIFT points and complemented with flow descriptors captured around each point. This representation achieve a proper performance of sign recognition but remains limited to cover much of the variability gestures. In [11] was obtained a local motion description at each frame of a particular SL by computing motion trajectories along the utterance. Nevertheless, this approach lost spatial representation of signs because the nature of the bag of words representation.

The main contribution of this work is a regional mid level representation of kinematic primitives that achieve a local description but also a regional coding representation of gestures during utterance sequences. The proposed approach is robust to describe sign from incomplete utterance representations being efficient in on-line applications. For doing so, a set of salient motion patches are characterized with kinematic histogram features like speed and motion boundaries and also regional features such as rotational and divergence were computed. Such volumes are coded in a shape difference VLAD that recover main centroids described by the means and variance of motion. This representation allows to

recover partial gestures and robustly describe different gestures in a particular sign language. Finally the obtained motion descriptor is mapped to a Support vector machine and validated with respect to both LSA64 and Boston public corpus.

2 Proposed approach

In this work is presented an automatic strategy to recognize gestures from a kinematic mid-level representation. The herein proposed approach achieves a robust local description of signs by considering several motion features. Then a classical dictionary approach allows to cluster volumetric patches to represent the signs. The mid-level representation allows to capture local kinematic similarities of patches but also is able to recover the shape of descriptor distribution from a shape difference VLAD [7] representation. The pipeline of the proposed approach is illustrated in Figure1.

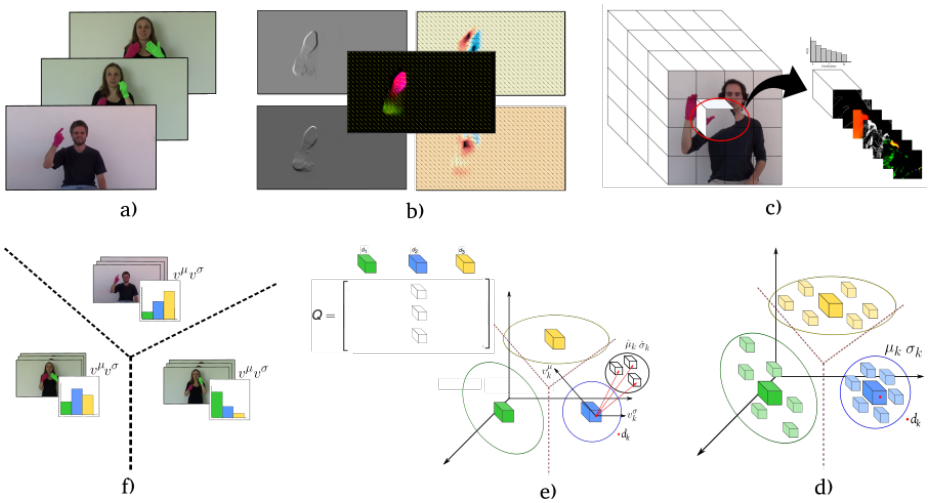

Fig. 1. Pipeline of the proposed approach to SLR. A set of utterance sign samples (a) are codified from a set of motion primitives (b). Then kinematic primitives are computed in utterance sequences (c) and coded as a local volumetric patch description. A dictionary of patches (d) is obtained and finally a coding representation is obtained using Hard assignment and SD-VLAD. The computed descriptor is mapped to a previously trained SVM to recognize the gestures (f)

2.1 Computing kinematic features

The motion characterization had proven to be fundamental to analyzing atomic symbols in gesture recognition applications. A fundamental task in this kind of

applications is to quantify large motion regions developed by independent actuators, such as arm, hands, face or even shoulders. In this approach we compute a set of kinematic primitives to describe gestures independent of appearance. The set of set of computed features are described as follows:

– **Dense flow velocity fields**
A first kinematic primitive herein considered was the dense appearance field produced among consecutive frames. Typical approaches remain limited to quantify large displacements because the assumption of smooth motion in local neighborhoods. To avoid these limitations, herein was implemented a robust optical flow approach able to capture dense flow fields but considering large displacements of gestures [2]. In this approach is considered classical flow assumptions in which color $E_{color}(w)$ and gradient $E_{gradient}(w)$ changes remain constant among consecutive frames. Likewise, it is also included a local criteria of field smoothness, expressed as:

$$E_{smooth}(w) = \int_{\Omega} \Psi(|\nabla u|_{t_{i+1}} + |\nabla v|_{t_i}) d\mathbf{x} \tag{1}$$

where Ψ represents the atypical values that are penalized in a specific neighborhood and Ω is the region analyzed. Finally a non-local criteria is considered in this approach that allows the estimation of coherent large displacements. In this case, a sift matching is carried out among consecutive frames, and then the flow regions of such interest matched regions are measured to find flow similar patterns $\mathbf{f}_{t_i}(\mathbf{x})$, described as:

$$E_{desc}(w_1) = \int_{\Omega} \delta(\mathbf{x})\, \Psi(|\mathbf{f}_{t_{i+1}}(\mathbf{x} + w_1(\mathbf{x})) - \mathbf{f}_{t_i}(\mathbf{x})|^2) d\mathbf{x} \tag{2}$$

with $\delta(\mathbf{x})$ as step function that is active only for regions where exist interest points. The sum of whole restrictions are minimized from a variational Euler-Lagrange approach.

– **Divergence fields**
Additionally to velocity field description, in this work was also considered the physical pattern of divergence over the field. The feature result from the derivative of flow components (u, v) at each point x along spatial directions (x, y), described as:

$$div(p_t) = \frac{\partial u(p_t)}{\partial x} + \frac{\partial v(p_t)}{\partial y} \tag{3}$$

This feature capture a local expansion of field, and result useful to characterize independent body actuators along a sign description.

– **Rotational fields**
The rotational measures of flow field was also herein considered. From each local point of estimated field is measured the rotation around of a perpendicular axis [1, 8]. This rotational patterns stand out circular gestures, commonly reported in sign languages. Also, this measure estimate the flow

rigidity, useful to distinguish articulated motions. The rotation of field can be expressed as:

$$curl(p_t) = \frac{\partial v(p_t)}{\partial x} - \frac{\partial u(p_t)}{\partial y} \tag{4}$$

– **Motion limits**
First spatial derivative in flow components are also estimated [6] as kinematic information of signs, coding the relative motion among pixels. The gradient of flow remove constant motion information, while remain the changes of velocity. This primitive also highlight main articulator motions.

2.2 Coding motion gesture patches

A main drawback of global gesture characterization is the sensibility to occlusion of articulators, and scene perturbations while the sign is described. The herein proposed approach is based on a local gesture representation, from which, a set of local motion patches can represent the temporal description of a sign gesture. Because the proposed representation is mainly based in motion patches, a first step was to remove background patches with poor motion information. For doing so, we compute the average background of the video as: $B(\widehat{x}, y) = \frac{1}{t}\sum_{t=1}^{t} f_t(x, y)$ for each t frames and then foreground pixels are get by a simple subtraction w.r.t the background $|f_t(x, y) - B(\widehat{x}, y)| > \tau$. Differences larger than τ (experimentally obtained value) are considered static pixels and removed. For on-line purposes, the average background can be built from a recursive mean estimator. To remove relative static patches also improve the computational efficiency of the approach (see in Figure 2).

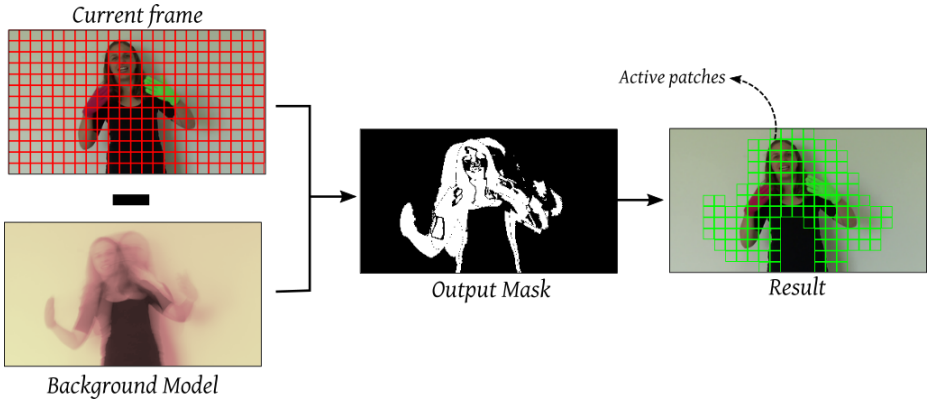

Fig. 2. An efficient kinematic patch representation is achieved by only considering patches with relevant motion information. To remove static pixels is herein considered a simple but efficient background model.

2.3 Kinematic patch description

In this work, a particular sign is defined as a set of n spatio-temporal patches $S = \{p_{1...n}^{(c,j)} : j \in [t_1 - t_2]; c \in [x_1, x_2]\}$ bounded in a temporal interval j and spatially distributed in a c region. Each of these volumetric patches are described using the motion local information, coded as kinematic histograms. Then, for every kinematic primitive considered in the proposed approach a local histogram representation was considered, as:

$$h(p) = \sum_{\mathbf{x} \in p} R_b(\mathbf{x}) W(\mathbf{x}), b = \left\{1, 2, \cdots, \frac{2\pi}{\Delta\theta}\right\}$$
$$R_b(x, y) = \begin{cases} 1 \;\; if \;\; (b-1)\Delta\theta \leq \theta(\mathbf{x}) < b\Delta\theta \\ \\ 0 \;\; \text{elsewhere} \end{cases} \tag{5}$$

where $R_b(\mathbf{x})$ is an activation function that determine the particular bin that code the local kinematic feature, while the $W(\mathbf{x})$ corresponds to the particular weight that sum in the histogram bin. Particularly, for orientation flow histograms (**HOOF**) the bins b correspond to orientations, while the $W(\mathbf{x})$ is defined by the norm of each vector, as proposed in [5]. Likewise, the motion limits are codified as MBH histograms, quantified for each x, y components as proposed in [6]. For divergence and curl the primitives are statistically cumulated by defining the bins as: $\{\max, \frac{\max}{2}, 0, \frac{\min}{2}, \min\}$. In such case the curl histogram (**HCURL**) quantify the main motion around perpendicular axis, while divergence histogram (**HDIV**) summarize the main moments of divergence present around each spatio-temporal patch. For divergence a simple occurrence counting is carried out while for rotational the occurrence is weighted according to angular speed. The final descriptor for each patch is formed as the concatenation of all histogram herein considered.

2.4 Mid level gesture representation

Gesture dictionaries Each volume is coded with kinematic histograms that represent different local and regional primitives. The relationships among such volumes along the video highlight predominant patterns to describe gestures. Such patterns are the local representation to compute SLR from motion characteristics. To build a volumetric dictionary, a set K representative volumes: $D = [d_1, d_2, ...d_k] \in \mathbb{R}^{\mathbf{d} \times \mathbf{K}}$ are recovered from a set of N volumetric patches described by a *d-dimensional* descriptor $X = [x_1, x_2, ...x_N] \in \mathbb{R}^{\mathbf{d} \times \mathbf{N}}$ using a classic *k-means* algorithm, where $K << N$. Under the assumption of patch density, we consider that a set of K patches are sufficient to represent particular gestures and that each articulator is form by a set of these mean patches.

Gesture coding The computed dictionary is used as reference to code a global representation of registered sign along the video. The codification strategies are

group in three different classes: (1) voting based strategies that associated each descriptor volume to a specific word in the dictionary, (2) the reconstruction based coding that built a local descriptor sample from inputs and (3) super-vectors base coding that achieve a high dimensional representation by adding statistical information about descriptor shape [13]. To preserve independence of local description, the proposed approach implements a hard assignment HA of each computed kinematic patch w.r.t the dictionary of gestures, as:

$$HA(x) = \begin{cases} 1\ if \quad i = \text{argmin}_j\ ||x - d_j||_2 \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

where each kinematic volume vote for the most similar pattern in the dictionary. This kind of assignment allows to stand out main spatio-temporal regions associated with salient learned patches in the dictionary. Eventually, such representation can border similar gestures in regional salient details recovered.

Shape difference VLAD Classical Bag of Words (BoW) codify occurrences using simple occurrence strategies that lost information about descriptor and particular details of gestures, which can be dramatical in SLR [16]. Currently, the codification Vector of Locally Aggregated Descriptors (VLAD) have shown advantages w.r.t mid level representations by considering statistics of first order about computed cluster descriptors [9]. In such representation, the difference among local descriptors and predominant patterns are cumulated in a local characteristic vector. The whole difference vectors for each clusters form the video descriptor, denoted as: $v_k^{\mu} = \sum_{j=1}^{n_k}(x_j - D_k)$ with dimensionality of $K \times d$. This particular strategy achieves gesture description from dominant kinematic patterns by capturing similarities sign motions w.r.t centroid dictionaries and adding variance informations. In clusters with low variability, the resultant vectors has mainly zero values. Such fact result interesting to differentiate kinematic patterns with dynamic and spatial similarities along utterances. Nevertheless,this strategy is limited to capture the local distribution of the motion descriptor and is variant w.r.t. symmetric motions. For instance, same features vectors can result from different kinematic gestures. Hence, the standard deviation of each cluster can be aggregated to complement statistical information of signs, recovering regional relationships of patches that form a cluster [7]. To achieve this variance cluster representation, firstly, the characteristic vectors of each cluster proposed in [9] are weighted by their respective standard deviations and normalized by the number of descriptors as:

$$v_k^{\mu} = \frac{1}{n_k} \sum_{j=1}^{n_k} \frac{(x_j - D_k)}{\sigma_k} \tag{7}$$

where the normalization n_k is a pooling carried out to VLAD descriptors. To highlight the calculation of the variance descriptor, a new cluster $\hat{D}_k$ is estimated with projected samples of particular test utterances what are assigned to the

pattern D_k. Then, the variance of the means is defined as the difference between the new $\hat{D}_k$ estimated centroid and the dictionary centroid D_k, as:

$$\begin{aligned} \bar{v}_k &= \frac{1}{n_k}\sum_{j=1}^{n_k}(x_j - D_k) = \frac{1}{n_k}(\sum_{j=1}^{n_k}(x_j) - n_k D_k) \\ &= \frac{1}{n_k}\sum_{j=1}^{n_k}(x_j) - D_k = \hat{D}_k - D_k \end{aligned} \tag{8}$$

From same analysis, a new representation is added to descriptor by computing differences among standard deviation, such as:

$$v_k^{\sigma} = \hat{\sigma_k} - \sigma_k = \left(\frac{1}{n_k}\sum_{j=1}^{n_k}(x_j - D_k)^2\right)^{\frac{1}{2}} - \sigma_k \tag{9}$$

where $\hat{\sigma_k}$ is the standard deviation of assigned local descriptors in VLAD and σ_k is the standard deviation of assigned descriptors D_k. Such difference recover shape information of descriptor. The SD-VLAD descriptor is form by the concatenation of vectors v^{μ} and v^{σ}. Finally is applied a normalization at each dimension of the descriptor as suggested in [14] as: $f(\mathbf{p}) = sign(\mathbf{p})|\mathbf{p}|^{\frac{1}{2}}$

2.5 SVM sign recognition

The recognition of each potential sign is carried out by a Support Vector Machine (SVM) [4] classifier since this constitutes a proper balance between accuracy and low computational cost. The present approach was implemented using a *One against one SVM multi-class classification* with a Radial Basis Function (RBF) kernel. Here, the classes represent the particular signs coded as SD-VLAD descriptors and optimal hyperplanes separate them by a classical max-margin formulation. For m motion classes, a majority voting strategy is applied on the outputs of the $\frac{m(m-1)}{2}$ binary classifiers. A (γ, C)-parameter sensitivity analysis was performed with a grid-search using a cross-validation scheme and selecting the parameters with the largest number of true positives.

3 Evaluation and results

A public corpus of a sign language LSA64 [15] was herein used to evaluate the proposed approach. Such corpus describe a total of 64 signs that correspond to the Argentinian Sign language performed by 10 non-expert signers. Each sign is developed 5 times by each signer by a total of 3200 utterance videos. The spatial resolution is 1920×1080 at 60 frames per second. The selected signs involve articular motions, the use of one or both hands, and evident displacements in space and time. The corpus was captured in different scenarios, with some illumination changes. Several challenges are present in some different gestures with dynamic

and geometric similarities during the sequences except in some localized spatio-temporal regions. For experimental evaluation, the dataset was spatially resized to 346×194, since proposed approach is mainly based in kinematic features captured from temporal correlations. The whole experiments were computed with volumetric patches of $15 \times 15 \times 5$ with kinematic histograms of 7 bins for HOOF and 14 bins for MBH (both directions) and 5 bins for HDIV y HROT features. A total of 31 scalar values constitutes the dimension for each considered patch. Regarding, the ASL64 corpus the recognition strategy was validated by using a k-fold cross validation with a $k = 10$. At each iteration, a signer was tested while other 9 signers were used for training model.

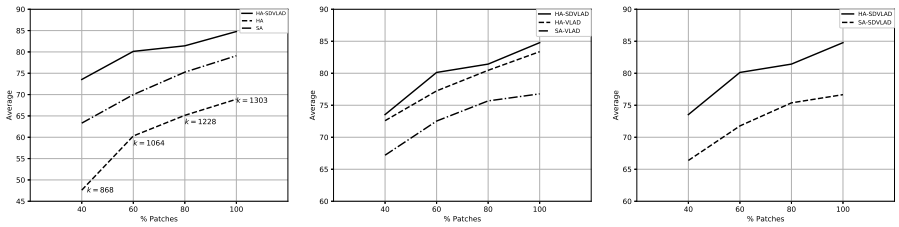

Fig. 3. On the left is illustrated the average performance using the common approaches Hard and Soft assignment from different dictionaries computed from $K = (\frac{patches}{2})^{\frac{1}{2}}$. In middle is illustrated the performance of VLAD version while on the right is illustrated the SD-VLAD computed over Hard and Soft assignments with $K = 64$. For all experiments we used partial information, from 40% to 100%.

In a first experiment, were evaluated different mid-level representation with different coding strategies over the LSA64 corpus. In whole experiments was considered only patches that have motion information, i.e., patches computed mainly from foreground signers. In order to evaluate the capability of representations to recognize partial and incomplete utterances, several incremental sub-set of input patches were considered to built the dictionary. In Figure 3 is illustrated the performance of different representation and codings. In Figure 3-left is shown the proposed HA-SDVLAD w.r.t classical bag of words representation using hard and soft assignment representation. As reported, the SD-VLAD representation achieved a better description of utterances and its able to represent partial signs with 80% of accuracy using only the 60% of information. In Figure 3-middle is reported several configuration for VLAD representation. As expected, the best performance is achieved by the hard assignment (HA) in SDVLAD because the proper codification of salient sign patterns in both local and regional characterization. At Figure 3-right is observe the contribution for hard assignment (HA) coding in the proposed representation achieving 85% of accuracy for complete utterances. Also, it should be mentioned that the proposed strategy is able to capture partial utterance only using the 60% of information and even with the 40% of information achieved plus of 70% of accuracy.

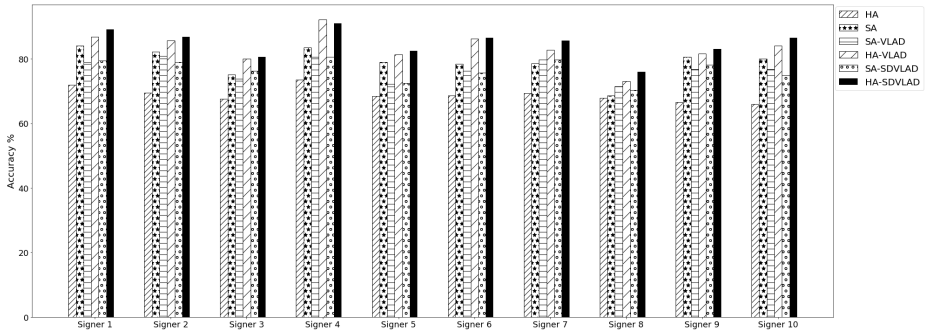

Fig. 4. A individual signer analysis is carried out from different mid-level representations for LSA64.

In a second experiment (see in Figure 4) was evaluated the performance of proposed approach in individual signers using different representations. In whole signers performance the HA-SDVLAD approach achieve the best representation of signs, thanks to shape representation of kinematic clusters. Particularly for signer 8 there exist some limitations because strong noise variation in temporal recording such as hair movement. Other noise artifacts occupy more than 25% of spatial video area.

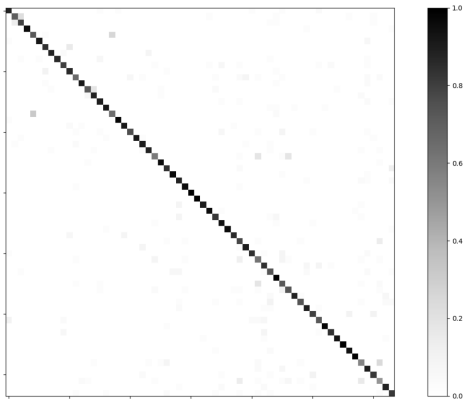

Fig. 5. The confusion matrix obtained with the LS64 dataset. The proposed approach achieves an average score close to 85% for the multi-class recognition

A complete analysis of analyzed gestures is reported in confusion matrix of Figure 5. As expected the proposed approach is able to recognize almost perfectly whole considered sign gestures with some limitations in particular gestures that share more than the 90% of dynamic gesture information. For instance the particular sign "realize" is misclassified w.r.t to sign "buy" because in almost

whole utterance the gestures are the same except in a local finger action at the end of gestures.

Additionally, the proposed motion descriptor was evaluated in the classical RWTH-BOSTON50 [18]. This dataset has widely used to test strategies in realistic scenarios, containing more than 50 words and three signers. The frames were recorded at 30 frames per second and frame size is 195 $\times$ 165 pixels. To compare obtained results, it was selected a set of utterance like proposed in [18, 20]. The considered words (with the respective number of utterances) are: *can* (19), *something*(12), *house* (12), *Possession* (12), *ix far* (12), *woman* (8), *break-down* (5) , *new* (7) , *not*(7), *like*(6). The recognition strategy was validated by using a k-fold cross validation with a $k = 10$. The proposed approach achieved outperform conventional approaches obtaining an average accuracy of 82%.

4 Conclusions

In this work was proposed a novel approach to describe gestures from a shape difference VLAD representation of kinematic volumes. The proposed approach is relative invariant to appearance, robust to occlusion and recover salient regions of the gestures coded in the computed dictionary. From a corpus with more than 3000 utterance videos and five signers, the proposed method achieved 85% of accuracy in average. Also the proposed approach was evaluated over a classical Boston dataset obtaining a 82% of accuracy in average. Future works include continuous representations of gestures to achieve on-line translation in SLR and frame-level evaluation to built a grammatically more complex models.

Acknowledgments

The authors acknowledge the Vicerrectoría de Investigación y Extensión of the Universidad Industrial de Santander for supporting this research registered by the project: Análisis de movimientos salientes en espacios comprimidos para la caracterización eficiente de videos multiespectrales, with VIE code 2347"

References

1. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and multiple instance learning. IEEE transactions on pattern analysis and machine intelligence **32**(2), 288–303 (2010)
2. Brox, T., Bregler, C., Malik, J.: Large displacement optical flow. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 41–48. IEEE (2009)
3. centre, W.M.: Deafness and hearing loss (Mar 2018), `http://www.who.int/mediacentre/factsheets/fs300/en/`
4. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST) **2**(3), 27 (2011)

5. Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R.: Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 1932–1939. IEEE (2009)
6. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: European conference on computer vision. pp. 428–441. Springer (2006)
7. Duta, I.C., Uijlings, J.R., Ionescu, B., Aizawa, K., Hauptmann, A.G., Sebe, N.: Efficient human action recognition using histograms of motion gradients and vlad with descriptor shape information. Multimedia Tools and Applications pp. 1–28
8. Jain, M., Jegou, H., Bouthemy, P.: Better exploiting motion for better action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2555–2562 (2013)
9. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 3304–3311. IEEE (2010)
10. Konecnỳ, J., Hagara, M.: One-shot-learning gesture recognition using hog-hof. Journal of Machine Learning Research **15**, 2513–2532 (2014)
11. Martínez, F., Manzanera, A., Gouiffès, M., Braffort, A.: A gaussian mixture representation of gesture kinematics for on-line sign language video annotation. In: International Symposium on Visual Computing. pp. 293–303. Springer (2015)
12. Paulraj, M., Yaacob, S., Desa, H., Hema, C., Ridzuan, W.M., Ab Majid, W.: Extraction of head and hand gesture features for recognition of sign language. In: Electronic Design, 2008. ICED 2008. International Conference on. pp. 1–6. IEEE (2008)
13. Peng, X., Wang, L., Wang, X., Qiao, Y.: Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision and Image Understanding **150**, 109–125 (2016)
14. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. Computer Vision–ECCV 2010 pp. 143–156 (2010)
15. Ronchetti, F., Quiroga, F., Estrebou, C.A., Lanzarini, L.C., Rosete, A.: Lsa64: An argentinian sign language dataset. In: XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016). (2016)
16. Wan, J., Ruan, Q., Li, W., Deng, S.: One-shot learning gesture recognition from rgb-d data using bag of features. The Journal of Machine Learning Research **14**(1), 2549–2582 (2013)
17. Zafrulla, Z., Brashear, H., Starner, T., Hamilton, H., Presti, P.: American sign language recognition with the kinect. In: Proceedings of the 13th international conference on multimodal interfaces. pp. 279–286. ACM (2011)
18. Zahedi, M., Keysers, D., Deselaers, T., Ney, H.: Combination of tangent distance and an image distortion model for appearance-based sign language recognition. In: Joint Pattern Recognition Symposium. pp. 401–408. Springer (2005)
19. Zahedi, M., Keysers, D., Ney, H.: Appearance-based recognition of words in american sign language. Pattern recognition and image analysis pp. 373–384 (2005)
20. Zaki, M.M., Shaheen, S.I.: Sign language recognition using a combination of new vision based features. Pattern Recognition Letters **32**(4), 572–577 (2011)