# Towards On-Line Sign Language Recognition Using Cumulative SD-VLAD Descriptors

Jefferson Rodríguez(✉) and Fabio Martínez

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

**Abstract.** On-line prediction of sign language gestures is nowadays a fundamental task to help and support multimedia interpretation of deaf communities. This work presents a novel approach to recover partial sign language gestures by cumulative coding different intervals of the video sequences. The method starts by computing volumetric patches that contain kinematic information from different appearance flow primitives. Then, several sequential intervals are learned to carry out the task of partial recognition. For each new video, a cumulative shape difference (SD)-VLAD representation is obtained at different intervals of the video. Each SD-VLAD descriptor recovers mean and variance motion information as signature of the computed gesture. Along the video, each partial representation is mapped to a support vector machine model to obtain a gesture recognition, being usable in on-line scenarios. The proposed approach was evaluated in a public dataset with 64 different classes, recorded in 3200 samples. This approach is able to recognize sign gestures using only 20% of the sequence with an average accuracy of 53.8% and with 60% of information, the 80% of accuracy was achieved. For complete sequences the proposed approach achieves 85% on average.

**Keywords:** On-line recognition · Motion analysis
Mid-level representation · Shape difference VLAD

## 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) [2]. Sign languages is the main resource of communication and interaction among deaf people, being rich and complex as any spoken language. This articulated language is composed by coherent and continuous spatio-temporal gestures that summarize the articulated motions of upper limbs, facial expressions and trunk postures. Despite of the importance of automatic interpretation of sign languages, such characterization remains as an open problem because the


J. E. Serrano C. and J. C. Martínez-Santos (Eds.): CCC 2018, CCIS 885, pp. 371–385, 2018.
https://doi.org/10.1007/978-3-319-98998-3_29

multiple inter and intra signers variations. Also, different factors such as culture and regions can introduce external variations to sign languages. Such variations imply great challenges to understand and associate semantic language labels to spatio-temporal gestures. Also, for real interactions, current automatic interpretations demand on-line applications to recognize gestures while they are developed. In such sense, the problem is even more difficult because computational strategies must predict incomplete gestures while remaining robust to illumination changes, variations of perspective and even partial occlusion of signers.

The sign language recognition (SLR) has been addressed in literature by multiple approaches that include global shape representations that segment all articulators but with strong limitations due to occlusions and dependences of controlled scenarios. For instance, in [21] a multi-modal analysis was proposed to recover shape information from RGB-D sequences. Local gesture representations include interest points characterization [13,20] and the analysis of appearance and geometrical primitives to represent gestures in videos [16,19]. Zahedi *et al.* [22] proposed a SLR by computing descriptors of appearance that together with gradient of first and second order characterize particular signs. Such approach is dependent of signer appearance and perspective in video sequence. Motion characterization has also been used to recognize gestures being robust to appearance variance and illumination changes [11,13]. For instance, in [11,20] Lukas-Kanade motion fields were computed to characterize gestures in terms of velocity displacements. Nevertheless, this strategy is 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. Also, Konecný *et al.* [11] integrates local shape information with histograms of optical flow to describe gestures. This approach achieved a frame-level representation but lose local and regional information. Wan *et al.* [20] proposed a dictionary of sparse words codified from salient SIFT points and complemented with flow descriptors captured around each point. This representation achieves a proper performance of sign recognition but remains limited to cover much of the variability gestures. In [13] was implemented a local frame motion description for SLR by computing motion trajectories along of the sign but losing spatial representation of the signs.

Additional, machine learning strategies are proposed for gesture recognition from real-time and on-line perspectives [8,14,15]. For instance, Masood *et al.* [14] proposed a deep convolutional model to represent spatial and temporal recurrent features. This approach allows a sign representation of multiple gestures but with several limitations to segment articulators of signers. Also in [15] a 3D convolutional network (3D CNN) was adapted to recognize gestures in sign language. Initially, They normalize the number of video frames. They then apply the CNN model with two layers, one for feature extraction and the other for classification. Although, 3D feature extraction is more suitable for video processing, this method evidently does not take into account motion information. On the other hand, Fan *et al.* [8] recognize frame-level gestures using a simplified two-stream CNNs network. This network is trained with dense optical flow information as

input to the convolutional network. However, this single kinematics is insufficient to describe large human motion, that result fundamental in language recognition. Other alternatives have included multi-modal information [5,12], for instance, Liu *et al.* [12] proposed a computational strategy over RGBD sequences by firstly segmenting and tracking hands. Then a convolutional proposal was adapted to learn hand trajectories but with limitations in the representation of first order kinematics.

The main contribution of this work is a novel strategy to recognize partial gestures by using a cumulative regional mid level representation of kinematic primitives. The proposed approach achieves coding gestures while they are being developed in video sequences. Firstly, a kinematic representation of gestures is carried out by coding features from a dense large displacement optical flow. Then a patch volume based coding is carried out at each frame to code the developed gestures. A set of dictionaries that compute different intervals of the gestures are built from training videos. Finally, a test video is coded as a shape difference VLAD representation to recover main means and variance motion clues. Such representation is carried out at different intervals of the video and mapped to a previously trained support vector machine, allowing a partial gesture recognition. The proposed approach was evaluated in a public sign gesture corpus with 64 different classes and more than 3000 videos. This approach is able to recognize sign gestures using 20% of the sequence with an average accuracy of 53.8% and for 60% of the information, 80% accuracy on average is achieved. For complete sequences, 85% average accuracy is obtained. The rest of the paper is organized as follows: Sect. 2 introduces the proposed method, Sect. 3 presents results and the evaluation of the method, and finally Sect. 4 presents several conclusions and perspectives of the proposed approach.

## 2 Proposed Approach

A cumulative gesture representation is herein proposed to recognize video sequences. The proposed approach starts by a kinematic local level representation to achieve appearance independence characterization. The kinematic primitives are computed from a dense optical flow that take into account large displacements. Multiple temporal and cumulative dictionaries are then built from a patch volume representations of the kinematic primitives. At each defined video interval, the set of recovered patches with relevant motion information are coded w.r.t. the respective cumulative dictionary from a shape difference VLAD [7] representation. Finally the obtained representation of a particular video is mapped to a previously trained support vector machine to obtain a gesture label. The several steps considered in the proposed strategy are explained in detail in the next subsections.

### 2.1 Computing Kinematic Features

The method starts by characterizing sign gestures with low level kinematic relationships from a local velocity field. In such case, result crucial to quantify large

motion regions developed by independent actuators, such as arm, hands, face or even shoulders. To recover such large displacements a special dense optical flow was herein implemented [1] and then several measures were captured to represent motion. The set of kinematic features herein considered are illustrated in Fig. 1. The set of set of computed features are described as follows:

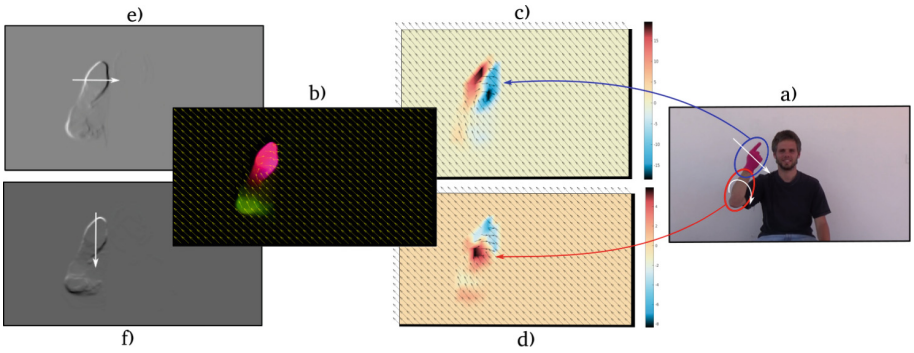


**Fig. 1.** Kinematic features computed along video sequences as low level description of gestures, namely: (b) large displacement optical flow, (c) divergence, (d) curl, (e) and (f) motion boundaries w.r.t. x and y axis

– **Dense flow velocity fields**
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 [1]. This approach consider a variational strategy to minimize classical flow assumptions in which color $E_{color}(w)$ and gradient $E_{gradient}(w)$ changes remain constant among consecutive frames. Likewise, additional assumptions are considered, as:

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

where $\Psi$ represents the atypical values, penalized in a specific neighborhood $\Omega$. Also, a non-local criteria allows the estimation of coherent large displacements. In this case, a sift point matching is carried out among consecutive frames to recover points with large displacements in space. Then the flow regions of such interest matched regions are measured to find flow similar patterns $\mathbf{f}_{t_i}(\mathbf{x})$, described as:

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

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**
  The physical pattern of divergence over the field was also herein considered as kinematic measure of gestures. This kinematic estimation 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 kinematic estimation captures a local field expansion and allows to characterize independent body actuators along a sign description.
- **Rotational fields**
  The rotational flow kinematic estimation was also considered to measure local rotation around of a perpendicular axis. This rotational patterns stand out circular gestures, commonly reported in sign languages [9]. 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 boundaries**
  The relative speed among pixels was also recovered as first spatial derivative in flow components [6]. This kinematic measure allows to code the relative motion among pixels and remove constant motion information. This primitive also highlight main articulator motions.

### 2.2 Coding Motion Gesture Patches

A main drawback of typical gesture strategies 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 volumetric motion patches are computed to represent a sign gesture. In this work only patches with motion information are taken into account, by removing background patches with poor motion information. For doing so, we firstly compute the average background of the video as: $B(\widehat{x}, y) = \frac{1}{t}\sum_{t=1}^{t} f_t(x, y)$. 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 $\tau$ 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 Fig. 2).

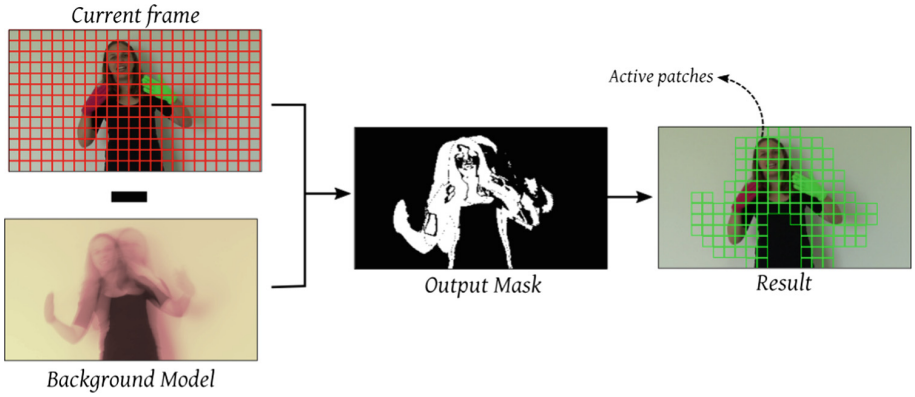


**Fig. 2.** An efficient kinematic patch representation is achieved by only considering patches with relevant motion information. To remove static pixels is herein considered a simple but efficient background model.

### 2.3 Kinematic Patch Description

Each of the recovered volumetric patches are described using the kinematic histograms of local motion information. Then, a histogram is built for every kinematic primitive considered in the proposed approach, 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 determines the particular bin, that-code-codes the local kinematic feature, while the $W(\mathbf{x})$ corresponds to a particular weight for each histogram bin. In case of orientation flow histograms (**HOOF**) the bins $b$ correspond to orientations, while the $W(\mathbf{x})$ is defined by the norm of each vector [4]. Likewise, the motion limits are codified as **MBH** histograms, quantified for each $x, y$ components [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. Then, a particular sign is defined as a set of $n$ spatio-temporal patches $S = \{p_{1\ldots n}^{(c,j)} : j \in [t_1 - t_2]; c \in [x_1, x_2]\}$ bounded in a temporal interval $j$ and spatially distributed in a $c$ region.

### 2.4 Mid Level Partial Accumulated Gesture Representation

A main contribution of this work is the possibility to predict gestures while they are developed in the sequence. For this purpose, a set of cumulative partial dictionaries are obtained at different periods of the sequence. Then, a SD-VLAD descriptor can be updated at different times in the video, achieving a prediction of the signs using cumulative patches information. The whole temporal representation is illustrated in Fig. 3 and will be explained as follows.

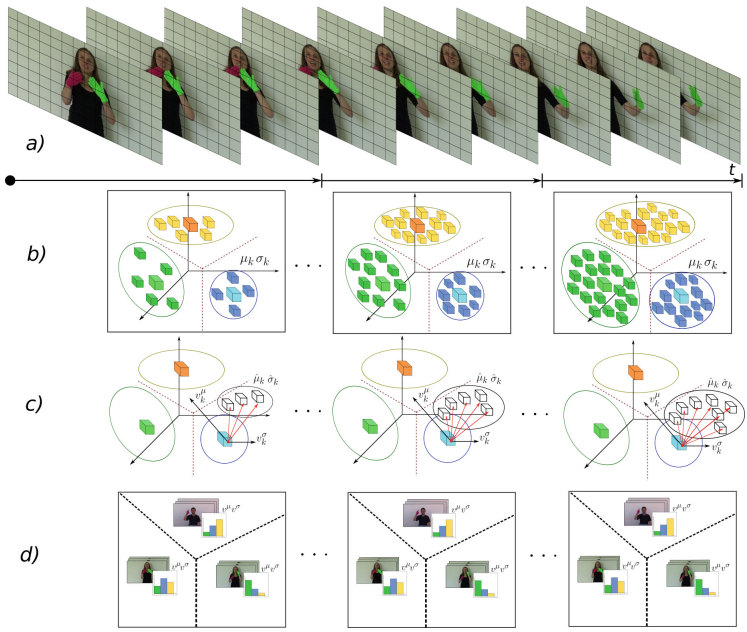


**Fig. 3.** The figure illustrates the mid level partial accumulated gesture representation. With the patches of all the video partial sequences (a), a dictionary adapted to the partial content is created (b) and updated as the information arrives. Finally a coding accumulated representation is obtained using Hard assignment and SD-VLAD (c). The computed descriptors are mapped to support vector machines previously trained with the accumulated partial descriptors (d).

**Gesture accumulated dictionaries** to temporally recognize sign gestures, a set of cumulative dictionaries $\boldsymbol{\Lambda} \in \mathbb{R}^{t \times w \times k}$ were built from training sequences with different interval gesture lengths. Then, $\boldsymbol{\Lambda} = [D_1, D_2, \ldots D_t]$ has $t$ temporal dictionaries that are built in a cumulative way each 20% of the sequences, *i.e.*, $D_1$ is a dictionary built only with the first 20% of active patches, $D_2$ summarize a representation of 40% of active sign patches and so on. Each dictionary $D^i = [d_1^i, d_2^i, \ldots d_k^i], \in \mathbf{R}^{\mathbf{w} \times \mathbf{K}}$ has $K$ representative centroids that correspond to $w-$dimensional kinematic features. Every built $D^i$ dictionary is constructed by using a classical *k-means* algorithm from a cumulative set of samples

$X^i = [x_1, x_2, \dots x_N]$ that increase as the gesture is developed. For each dictionary, $K << N$ by considering that a set of $K$ patches in each temporal partition are sufficient to represent particular gestures from incomplete data. Also, It is assumed that each articulator is formed by a set of these mean patches. This dictionary coding progressively achieves a finer representation of gestures because the major density of samples tend to obtain a better statistical representation.

The computed dictionaries are used as reference to code a global representation in each temporal partition of the video. Each local descriptor generated is coded using the respective cumulative dictionary calculated with the information size. 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. In such case the $HA$ corresponds to a voting based strategies that associated each descriptor volume to a specific word in the dictionary, formally defined as:

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

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

**Shape Difference VLAD.** A gesture descriptor is defined by the coding of each volume patch w.r.t. the temporal set of dictionaries $\boldsymbol{\Lambda}$ by using $HA$ association. In literature has been proposed several alternatives to obtain a global representation of patches w.r.t. general dictionaries. For instance, the classical Bag of Words (BoW) codifies patches using simple occurrence but lost information about patch descriptor and also lost particular details of gestures, which can be dramatical in SLR [20]. Currently, the codification Vector of Locally Aggregated Descriptors (VLAD) has shown advantages w.r.t. mid level representations by considering statistics of first order about computed cluster descriptors [10]. Such VLAD representation measures the differences among local descriptor patches of the new video and the closer centroid $c_k$. Formally, such differences can be expressed as: $v_k^{\mu} = \sum_{j=1}^{n_k}(x_j - c_k)$ with dimensionality of $K \times \mathbf{w}$. The association of new patches to specific centroids $c_k$ are achieved by the $HA$ correspondence. This particular strategy can achieve gesture description from dominant kinematic patterns by capturing similarities sign motions w.r.t. centroid dictionaries and adding variance informations. 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, a novel SD-VLAD aggregated the standard deviation of each cluster to complement statistical of VLAD vectors, recovering regional relationships of patches that form a cluster [7].

Following the SD-VLAD representation scheme, at each temporal interval $t$ of the video, defined by a percentage of patches, is taken the difference among local

descriptor patches of new video and the closer cluster $K$. To achieve this variance cluster representation, firstly, the characteristic vectors of each cluster proposed in [10] are weighted by their respective standard deviations and normalized by the number of descriptors as:

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

where the normalization $n_k^t$ is a pooling carried out to VLAD descriptors. To highlight the calculation of the variance descriptor, a new cluster $\hat{c_k}$ is estimated with projected samples of a particular test what are assigned to the pattern $c_k$. Then, the variance of the means is defined as the difference between the new $\hat{c_k^t}$ estimated centroid and the dictionary centroid $c_k^t$, for a particular $t$ cumulative dictionary, defined as:

$$\begin{aligned} v_{k,\mu}^t = \frac{1}{n_k^t} \sum_{j=1}^{n_k^t} (x_j^t - c_k^t) &= \frac{1}{n_k^t} (\sum_{j=1}^{n_k^t} (x_j^t) - n_k^t c_k^t) \\ &= \frac{1}{n_k^t} \sum_{j=1}^{n_k^t} (x_j^t) - c_k^t = \hat{c_k^t} - c_k^t \end{aligned} \tag{8}$$

Such variance of the means $v_k^t$ is coded at each time $t$, w.r.t. the particular cumulative dictionary $D_t$. From same analysis, a new representation is added to descriptor by computing differences among standard deviation, such as:

$$v_{k,\sigma}^t = \hat{\sigma_k^t} - \sigma_k^t = \left( \frac{1}{n_k^t} \sum_{j=1}^{n_k^t} (x_j^t - c_k^t)^2 \right)^{\frac{1}{2}} - \sigma_k^t \tag{9}$$

where $\hat{\sigma_k^t}$ is the standard deviation of assigned local descriptors in VLAD and $\sigma_k^t$ is the standard deviation of assigned descriptors $c_k$. Such difference recover shape information of descriptor. The SD-VLAD descriptor is form by the concatenation of vectors $v_{\mu}^t$ and $v_{\sigma}^t$ at each time $t$. Finally is applied a normalization at each dimension of the descriptor as suggested in [17] as: $f(\mathbf{p}) = sign(\mathbf{p})|\mathbf{p}|^{\frac{1}{2}}$. In such way is possible to obtain a partial representation at each interval $t$ of the video.

### 2.5 SVM Sign Recognition

The recognition of each potential sign is carried out by a Support Vector Machine (SVM) [3] 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. Taking into account that our representation constitutes several SD-VLAD partial representations, for each defined interval of representation we built a particular SVM model. A $(\gamma, C)$-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 [18] 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 $1920 \times 1080$ at 60 frames per second. The selected signs involve articular motions with 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 $346 \times 194$, since proposed approach is mainly based in kinematic features. Five different intervals in time were defined to recover partial gestures, *i.e*, each 20% of the video were built a dictionary and SD-VLAD descriptor. 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. The experimental framework was stated to recognize new gestures developed for different signers, that are independent of signers employed during training. In LSA64 corpus, each signer has recorded several repetition of the same sign, and therefore the random partition of samples can lead to a mistake in the validation of recognition. Then, a leave-one-out scheme based on the number of signers was herein implemented. In this scheme, one different signer was tested while the other 9 signers were used for the training model, running 10 different experiments to obtain a robust statistical meaning.

The first experiment was performed for general sign gesture recognition using the whole video sequences. From such experiment was possible to obtain the best performance of our approach by using all available patch information in video. The performance of the proposed approach was fix for individual signers 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 build a dictionary with $K = 64$ and descriptors, all calculated patches for each sign were used. As reported (see in Fig. 4), the SD-VLAD representation achieved a better description of the gestures and its able to properly code the salient sign patterns in both local and regional characterization. For most people, the accuracy of our approach was about 85%, thanks to shape representation of kinematic clusters and the quality of the dictionary due to the background

removed. Particularly for signer 8 there exist some limitations because strong noise variation in temporal recording such as long hair movement and the signers $1, 4$ where the highest accuracy was obtained, with 90%, because they perform the sign in an ideal way and do not exaggerate facial expressions or capture additional movement as in person 8 due to the short hair (Fig. 4).

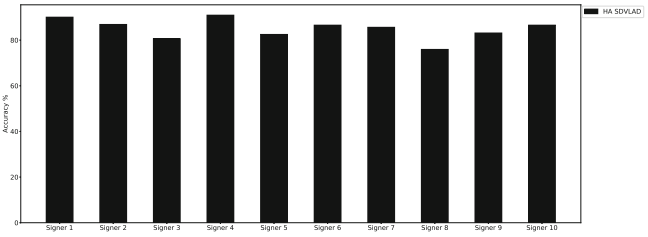


**Fig. 4.** A individual signer analysis is carried out for LSA64 by using the SD-VLAD descriptor. An average accuracy of 85% was achieved in the task of recognition. Some errors are reported for signer 8 because some variabilities of gestures as well as external motions are captured without correspondence with gesture information

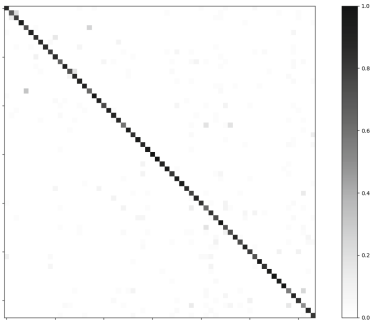

**Fig. 5.** The confusion matrix obtained in first experiment with the LS64 dataset. The proposed approach achieves an average score close to 85.45% for the multi-class recognition

A second experiment was designed to evaluate the performance of the proposed approach for temporal signs recognition task in several intervals of the video. In such case, each 20% of the video, in terms of number of patches, a new SD-VLAD descriptor is obtained and mapped to the SVM to obtain a prediction. Because each video records only one gesture we can determine the percentage of gesture information by simple counting the number of patches while the video is running. Nevertheless, in real application, the proposed approach is able to compute a prediction at each frame of the sequence. A first evaluation was carried out by only computing a general dictionary of gestures. In fact it was used

the same trained dictionary representation used for the task of classification. Figure 6-left illustrates the performance achieved by the proposed approach. As expected, the proposed approach has a poor performance with initial intervals of video, because the obtained sparse SD-VLAD representations are not taken into account during training. Nevertheless, for the 60% of the video, the proposed approach achieve a 70% in average of accuracy, showing an appropriate performance by only taking half of the information of the gesture. This result is fundamental because using only one dictionary is possible to recover sign gestures with half of the video information.

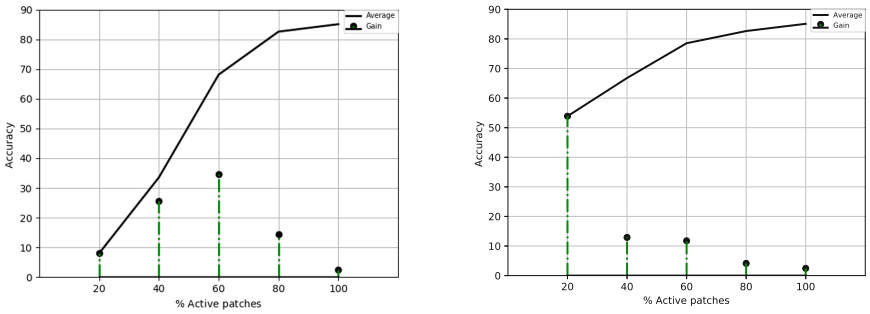


**Fig. 6.** A temporal recognition performance of the proposed approach. In left is illustrated the performance of the proposed approach by only using one dictionary. In right a complete implementation of the method is presented by computing $t$ different dictionaries. In last case, $t$ different SVM models are built for prediction.

In a third experiment, a complete version of the proposed approach was considered. In such case, several $t$ dictionaries were firstly trained to coded partial information. Also, a set of $t$ different SVM models were trained for each interval of the video. The experiment considered intervals of 20%, i.e. $t = 5$, as in the previous experiment. As illustrated in Fig. 6-right the proposed approach achieves competitive results even in few intervals of the video. For instance, using only the 20% and 40% of the video the strategy achieves on average 53.8% and 66.7%, respectively. It is worth noting, that such intervals corresponds to approximately 10 frames of the recorded gesture which is negligible in terms of the gesture description. For 60% of the video, the proposed approach achieves almost 80% of accuracy, and stable result with few information of the sequence. Such result using the half of the video sequences is comparable to the classification task carried out in the first experiment (Fig. 7).

The performance obtained with the proposed approach is competitive and result promising for develop more advanced tools that allows to carried out on-line recognition, as well as, the independent analysis of articulators during time. The corpus herein evaluated has been used in other recognition works. For example, in [14] was reported 95.2% accuracy but selecting 18 fewer classes. The framework of validation of this work is randomly and result difficult to evaluate

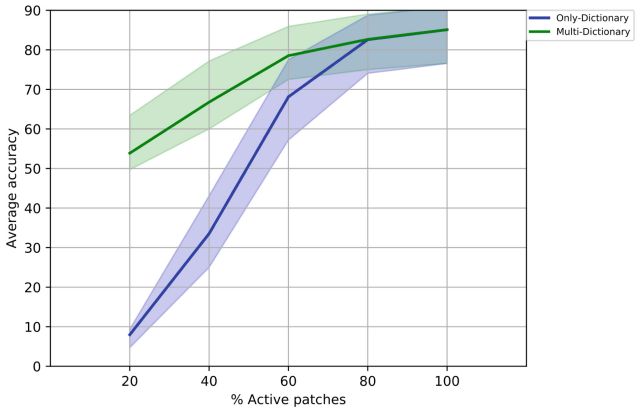


**Fig. 7.** Performance of the proposed approach using only one dictionary and a set of temporal dictionaries. In this illustration is shown the mean accuracy with the maximum an minimum score obtained by signer.

the real performance of the proposed approach with the complete corpus. Also, in [15] was reported an accuracy of 93.9% using only one random partition of corpus with 80% for training and 20% for testing. In such case, some repetitions of the same signer can be spread in both partition, losing the real meaning of gesture recognition accuracy.

## 4 Conclusions

In this work was proposed a novel computational strategy to temporally predict sign gestures by using a robust cumulative SD-VLAD representation. The proposed approach built a set of cumulative kinematic patch-based dictionaries to represent gestures in different intervals of the video. Each of these dictionaries has cumulative information of patches captured along the sequence. For any new sign gesture video, a SD-VLAD descriptor is obtained at different instances of the video by coding difference vectors w.r.t. the respective dictionary. The computed descriptor is mapped to a SVM model to obtain a recognition of the sign. The proposed approach achieves 80% of accuracy using only the 60% of the video sequences. Also for very few frames the approach achieve competitive results, demonstrating the capabilities to recognize signs from partial temporal information. The proposed approach is able to be used in on-line applications requiring few frames to capture an appropriate set of patches and built the SD-VLAD descriptor. Future works include frame-level evaluation to built a grammatically more complex models. Also, additional experiments will be doing with additional datasets.

**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. Brox, T., Bregler, C., Malik, J.: Large displacement optical flow. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 41–48. IEEE (2009)
2. Vaughan, G.: Deafness and hearing loss (2018). http://www.who.int/en/news-room/fact-sheets/detail/deafness-and-hearing-loss
3. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Tech. (TIST) **2**(3), 27 (2011)
4. 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: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1932–1939. IEEE (2009)
5. Chen, X., Koskela, M.: Online RGB-D gesture recognition with extreme learning machines. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 467–474. ACM (2013)
6. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006). https://doi.org/10.1007/11744047_33
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. Multimed. Tools Appl. **76**, 1–28 (2017)
8. Fan, Z., Lin, T., Zhao, X., Jiang, W., Xu, T., Yang, M.: An online approach for gesture recognition toward real-world applications. In: Zhao, Y., Kong, X., Taubman, D. (eds.) ICIG 2017. LNCS, vol. 10666, pp. 262–272. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-71607-7_23
9. 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)
10. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3304–3311. IEEE (2010)
11. Konecnỳ, J., Hagara, M.: One-shot-learning gesture recognition using HOG-HOF features. J. Mach. Learn. Res. **15**, 2513–2532 (2014)
12. Liu, Z., et al.: Real-time sign language recognition with guided deep convolutional neural networks. In: Proceedings of the 2016 Symposium on Spatial User Interaction, pp. 187–187. ACM (2016)
13. Martínez, F., Manzanera, A., Gouiffès, M., Braffort, A.: A Gaussian mixture representation of gesture kinematics for on-line sign language video annotation. In: Bebis, G. (ed.) ISVC 2015. LNCS, vol. 9475, pp. 293–303. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-27863-6_27
14. Masood, S., Srivastava, A., Thuwal, H.C., Ahmad, M.: Real-time sign language gesture (word) recognition from video sequences using CNN and RNN. In: Bhateja, V., Coello Coello, C.A., Satapathy, S.C., Pattnaik, P.K. (eds.) Intelligent Engineering Informatics. AISC, vol. 695, pp. 623–632. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-7566-7_63

15. Neto, G.M.R., Junior, G.B., de Almeida, J.D.S., de Paiva, A.C.: Sign language recognition based on 3D convolutional neural networks. In: Campilho, A., Karray, F., ter Haar Romeny, B. (eds.) ICIAR 2018. LNCS, vol. 10882, pp. 399–407. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93000-8_45
16. 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: International Conference on Electronic Design, ICED 2008, pp. 1–6. IEEE (2008)
17. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_11
18. 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)
19. Tofighi, G., Monadjemi, S.A., Ghasem-Aghaee, N.: Rapid hand posture recognition using adaptive histogram template of skin and hand edge contour. In: 2010 6th Iranian Machine Vision and Image Processing (MVIP), pp. 1–5. IEEE (2010)
20. Wan, J., Ruan, Q., Li, W., Deng, S.: One-shot learning gesture recognition from RGB-D data using bag of features. J. Mach. Learn. Res. **14**(1), 2549–2582 (2013)
21. 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)
22. Zahedi, M., Keysers, D., Ney, H.: Appearance-based recognition of words in American sign language. In: Marques, J.S., Pérez de la Blanca, N., Pina, P. (eds.) IbPRIA 2005. LNCS, vol. 3522, pp. 511–519. Springer, Heidelberg (2005). https://doi.org/10.1007/11492429_62