# Modalities Combination for Italian Sign Language Extraction and Recognition

Bassem Seddik(✉), Sami Gazzah, and Najoua Essoukri Ben Amara

SAGE Laboratory, National Engineering School of Sousse,
Sousse University, Sousse, Tunisia
bassem.seddik.tn@ieee.org, sami_gazzah@yahoo.fr,
najoua.benamara@eniso.rnu.tn

**Abstract.** We propose in this work an approach for the automatic extraction and recognition of the Italian sign language using the RGB, depth and skeletal-joint modalities offered by Microsoft's Kinect sensor. We investigate the best modality combination that improves the human-action spotting and recognition in a continuous stream scenario. For this purpose, we define per modality a complementary feature representation and fuse the decisions of multiple SVM classifiers with probability outputs. We contribute by proposing a multi-scale analysis approach that combines a global Fisher vector representation with a local frame-wise one. In addition we define a temporal segmentation strategy that allows the generation of multiple specialized classifiers. The final decision is obtained using the combination of their results. Our tests have been carried out on the Chalearn gesture challenge dataset, and promising results have been obtained on primary experiments.

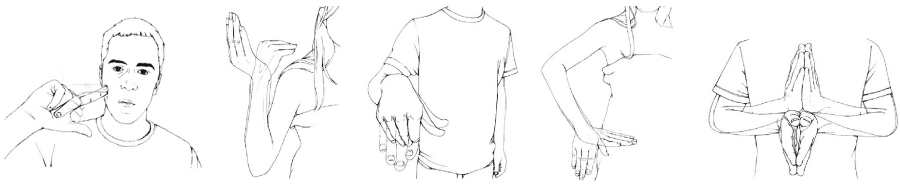
**Keywords:** Motion spotting · Action recognition · Fisher vector · Modalities combination · Classification fusion

## 1 Introduction

With the introduction of the Kinect sensor by Microsoft, a growing interest within the computer vision community has been conducted towards the improvement of human-action recognition solutions. The aimed applications range from education and entertainment to medical rehabilitation and sign language recognition [10]. As the Kinect sensor generates several types of spatial modalities, including the RGB, the depth and the skeletal joint pose streams of Shotton *et al.* [24], the challenge of optimally combining all of them is still open.

This paper is positioned in the context of multi-modal Italian sign language recognition. It aims to combine the different Kinect data streams in order to recognize a predefined set of actions in a continuous streaming real-world-like scenario. The samples of the considered actions are presented in Fig. 1. Human action can manifest in an infinite set of consecutive poses. In a continuous captured stream, we can find the resting poses where the actions are limited, the


V. Murino and E. Puppo (Eds.): ICIAP 2015, Part II, LNCS 9280, pp. 710–721, 2015.
DOI: 10.1007/978-3-319-23234-8_65



(a) Buonissimo (b) Tempo fa (c) Vieni qui (d) Ho fame (e) Cos'hai combinato

**Fig. 1.** Sample illustrations[1] of the Italian sign language vocabulary considered by the Chalearn gesture dataset [5] used in this paper

vocabulary action used for recognition and the rest of the non-significant actions that can be present within the streams.

Recently, efforts have been carried by a number of competitions for the creation of Kinect based dataset taking in consideration both spotting and recognizing actions. The competitive aspect of challenges such as [8] and [5] resulted in a multitude of works that have mainly focused on a dedicated type of modality to rapidly produce their results.

In this paper, we put forward both the combination of the different modalities offered by our sensor and the temporal extraction of motion before recognition. In this context, we propose an approach for the fusion of local feature decisions with global ones. We also contribute by a fusion strategy of multiple specialised SVM classifiers.

The organisation of the following writings will be turning around (1) the feature representation tools, (2) the action-extraction, and (3) the action-recognition solutions. They will be investigated in the next section of the literature review, then there will be dedicated related sections (sections (3) to (5)) detailing our own approach stages for each of them. Our experiments, conclusions and perspectives will be presented afterwards consecutively.

## 2 Related Works

We present in what follows the noticeable approaches of feature extraction, temporal segmentation and action recognition within the literature.

**Feature Representation:** For the RGB video streams, a first trend of human-action descriptors relied on region-of-interest gradient features [4,14] or spatio-temporal analysers such as [12]. More recently, global representations relying on sparse supervised [27] and unsupervised [20] encoders have gained growing interest.

For the joint modality, the first families of works focused on the recognition of actions using motion-capture signals from dedicated sensors [2]. Many new research works have used the Kinect joint stream and proposed dedicated measures related to the position, rotation, inter-joint relations and within-time

[1] http://www.theguardian.com/travel/series/learn-italian

behaviour [6,23]. It is noticeable that most of the joint-related representations belong to the hand-crafted feature type.

More recently, a number of research works have focused on the creation of features related to the depth streams. Most of them have relied on the binning of the orientations relative to the depth normals[17], the contextualisation of the point clouds [3] and the quantification of the motion's temporal differences [13].

**Action Extraction:** Also known as action spotting or temporal segmentation, it allows the delimitation of the beginning and the end of an action within a continuous stream of motion. The fastest methods for on-line human-action extraction belong to the heuristic family. In this case, a threshold is applied on a computed measure in order to capture 1D signal changes [1]. Similar thresholds can be found in [18] to evaluate the hand distance from a deduced resting position. Another one is applied in [13] to find whether the left or right hand is in action.

Other works try to analyse the behaviour of computed energy functions using different modalities. The solutions using sliding windows of specified temporal lengths and progressing steps can be found in [6] and [11]. More classic approaches rely on dynamic-programming-derived analysis in order to find temporal cuts within the continuous streams [8]. While more advanced solutions try to combine the advantages of the pre-listed approaches, as in [23], most of them are destined only for the binary classification between actions and resting positions.

**Action Recognition:** The work of Neverova *et al.* [16] can be considered as a frame-wise decision approach. While they used different sampling resolutions for the description of their frame contents with steps of 2,3 and 4, their final decision was on frame-scale. Similar works operating at local frame scale used hand-crafted features in combination with SVM classifiers. A joint-quadruplet descriptor has been proposed in [6] and a motion-trail based one has been presented in [13].

On the other hand, different types of works have focused on the use of global descriptors operating at the scale of the whole sequences concerned by recognition. Sparse representations derived from Bag of Words'(BoW) related representation proved their efficiency within the state of the art winning dense descriptor used in [18]. While the most classic BoW approach is related to vector-quantization global features derived form Fisher vectors have also been widely used [21].

**Our Proposed Approach:** We propose in this work a solution to generate segments with additional content knowledge from the temporal segmentation stage in order to improve recognition afterwards. This allows us to adopt a classification specialisation procedure that uses an adequate recogniser for every type of action. We distinguish in this context between the bi-handed and one-handed action labels.

By analysing the feature representations, we can highlight that while a many works have been interested in global representations of human actions using

sparse features derived from BoW-like approaches [18], recent works focusing more on local scale descriptions have proved their efficiency [16].

We also suggest in this work an approach that combines the strengths of both global and local representations for different modalities. We proceed by a bottom-up analysis using classifiers learned at a frame scale. Then, we apply a fusion with a second up-down analysis derived from the Fisher vector's [19] representation of Gaussian Mixture Model (GMM) probability distributions. The combination of both global and local scales is performed using the weighted sum of the frame-wise recognition probabilities generated by the SVM classifiers.

## 3 Feature Representation

We start in this section by presenting the local features extracted from the different modalities following the stages of our previous work detailed in [22], then we present our introduced global representation.

### 3.1 Frame-wise Feature Representation

We have designed our features for complementarity. We have used the stabilized hand-joint positions to delimit the hand 3D poses from RGB and depth sub-windows. The dynamics of the whole upper body have been deduced afterwards from the 11 upper joints.

**RGB Features:** We used the video colour streams in order to deduce the poses of both hands using the HoG [4] descriptor. To achieve this, we exploited the positions indicated by the joints to extract the bounding boxes around the hands and saved 32 descriptive bins (i.e. 8 orientations x 4 cells) per hand.

**Depth Features:** We have utilised the depth information so as to bring the evolution of the 2D HoG features along the Z axis. For this purpose, we have subtracted the background and evaluated the depth motion $DM(t)$ differences in time, as presented in equation (1):

$$\mathrm{DM(t)} = \gamma[(d(t+1).m(t+1) - d(t-1).m(t-1)] \tag{1}$$

where $m(t)$ is the actor mask stream offered by our dataset at each frame $t$, $d(t)$ is the depth and $\gamma$ a downscaling factor. The difference is computed between the next $t+1$ and the past $t-1$ frames. From this stage, we have saved 16 features per hand bounding region obtained from the joint positions.

**Joint Features:** The joint descriptors have been designed to complete the pose captured by the HoG hand features with others relative to the whole upper body's precise position and rotation information. These features have been extracted using the normalised 3D joint positions $J_p^i = [x, y, z]$ in addition to the four quaternion angles $J_q^i = [q_x, q_y, q_z, q_w]$, where $i = 1...11$ is the upper body joint index. Also, similar to the shape-context description [15], we have

analysed the joint pair-wise distances $J_d$ of the 11x11 size, given by equation (2), and subtracted the constantly null ones:

$$J_d = \| J_p^i - J_p^j \|_2 \quad , with \;\; i \neq j \tag{2}$$

Finally, the dynamic evolution of the joint speed $J_s$ and acceleration $J_a$ have been computed using equations (3) and (4) respectively, at the frame instants $t$:

$$J_s(t) = J_p(t+1) - J_p(t-1) \tag{3}$$

$$J_a(t) = J_p(t+2) - 2J_p(t) + J_p(t-2) \tag{4}$$

The obtained feature vector $\mathrm{J} = [J_p, J_q, J_d, J_s, J_a]$ has a size of 251 descriptors.

### 3.2 Global Feature Representation

In order to generate a wider feature representation scale, we have opted for the Fisher vector representation. This feature-space transformation has proven its efficiency, especially in the case of human-action extraction and recognition [18, 21]. It generates a sparse one-dimensional feature vector for each video stream and allows rapid SVM classification afterwards.

Considering a set of training feature vectors $F = [f_1, ..., f_t]$, extracted from $t$ learning frames using a number of $D$ features, we start by learning a GMM using expectation maximization approach [26]. The generated parameters $\Theta = \{\pi_k, \mu_k, \Sigma_k; k = 1, ..., K\}$ are saved such that $\pi_k$, $\mu_k$ and $\Sigma_k$ are respectively the prior probabilities, means and diagonal covariance matrices of every cluster $k$. To initialise the GMM $K$ centroids, we have applied a K-means clustering and considered 3 Gaussian mixtures per action label as presented in [9]. Thus, for a set of 5 bi-handed actions, we have extracted $K = 15$ centroids. They are associated to each $f_i$ sample by the posteriori probability given in equation (5):

$$\Gamma_{ik} = \frac{exp[-\frac{1}{2}(f_i - \mu_k)^T \Sigma_k^{-1}(f_i - \mu_k)]}{\sum_{l=1}^{K} exp[-\frac{1}{2}(f_i - \mu_l)^T \Sigma_k^{-1}(f_i - \mu_l)]} \tag{5}$$

where $i$ and $k$ denote respectively the frame indices and the k-means centroids. The Fisher generated vector of an action sequence $S$ is given by equation (6) where $j \in \{1, ..., D\}$ refers to the feature dimension:

$$\Phi(S) = [u_{j1}, v_{j1}, ..., u_{jK}, v_{jK}] \tag{6}$$

It is constructed using the concatenation of the mean and covariance's partial derivatives given in equations (7) and (8) respectively:

$$u_{jk} = \frac{1}{t\sqrt{\pi_k}} \sum_{i=1}^{t} \Gamma_{ik} \left[ \frac{f_{ji} - \mu_{jk}}{\sigma_{jk}} \right] \tag{7}$$

$$v_{jk} = \frac{1}{t\sqrt{2\pi_k}} \sum_{i=1}^{t} \Gamma_{ik} \left[ \left( \frac{f_{ji} - \mu_{jk}}{\sigma_{jk}} \right)^2 - 1 \right] \tag{8}$$

The generated Fisher vector of size $2KD$ is further improved using the function $f(x) = |x|sign(x)$ and applying $l^2$ normalisation [19].

## 4 Action-Segment Spotting

Our temporal segmentation methodology is similar to the one presented in [23], with many additional improvements. The common steps are related to the bi-processing stages going first into the heuristic joint analysis and then applying the SVM classification in order to robustly extract the motion segments. As presented in Fig. 2, the newly adopted steps have been related to the identification of the motion family out of 4 cases: the non-motion (i.e. label 0), the left-handed, the right-handed and the bi-handed actions (i.e. labels 1, 2 and 3 consecutively). We have been able to introduce this pre-classification using the following tests :

$$\begin{cases} \textit{Both hands} : \mathbf{3}\ \textit{if} \quad (J_p^{lh} - J_p^{rest}) > \tau \quad \textit{and} \quad (J_p^{rh} - J_p^{rest}) > \tau \\ \textit{Right hand} : \mathbf{2}\ \textit{if} \quad (J_p^{lh} - J_p^{rest}) > \tau \\ \textit{Left hand} : \mathbf{1}\ \textit{if} \quad (J_p^{rh} - J_p^{rest}) > \tau \\ \textit{No motion} : \mathbf{0}\ \textit{if} \quad (J_p^{lh} - J_p^{rest}) \leqslant \tau \quad \textit{or} \quad (J_p^{rh} - J_p^{rest}) \leqslant \tau \end{cases}$$

where $J_p^{rest}$ is the resting position identified by analysing the joint's most visited cell into a 200x200 grid. We have also saved the binary vector flags indicating whether we have a motion performed by the left hand, the right hand or both. The two first plots (a) and (b) presented in Fig. 2 show the considered thresholds for these vectors. Using the SVM classification, we have been able to extract different types of enriched motion positions. The obtained labels have allowed us to apply a fusion strategy between multiple classifiers with kernels tuned for every motion type.

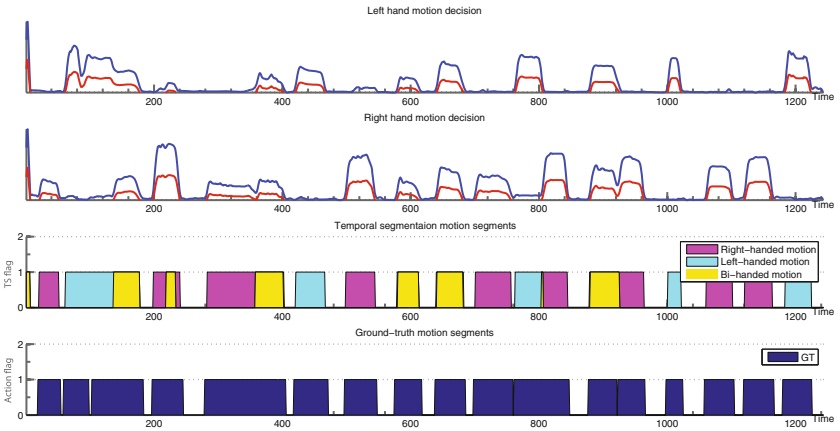



**Fig. 2.** Illustration of the motion spotting results: a- and b- plots showing the motion varition limit considered for the left and right hands respectively, c- the flags deduced from the left and right hand motion analysis, d- the ground-truth manual segmentation

## 5 Action Recognition Strategy

As presented in Fig. 3, our approach starts by grouping the RGB-D features with an order determined by the dominant hand first. This is obtained by looking at the cumulative motion of both hand joints. During the action recognition, the outputs of the previous segment-labelling stage have been used for classifier specialisation. We have deduced from the label-3 dominance that we had a bi-handed action and from the labels 1 and 2 that we have had a one-handed action. We have then redirected the generated descriptors to two classification pools. The first is related to the bi-handed actions (i.e. *'cheduepalle', 'chevuoi', 'daccordo', 'combinato' and 'basta'*) and the second (presented in section 6) is dedicated to the lasting one-handed ones. This specialization has reduced the inter-variability within the population of descriptors and allowed us to gain recognition improvements.

On a second level of decision, we use the SVM classifiers with the RBF kernels for the local descriptors and linear ones for the global descriptors. We have combined them using a weighted sum of the probability outputs $P_{gl}$ of the SVM classifiers, as in equation 9:

$$P_{gl} = \alpha P_g \oplus (1 - \alpha) P_l \tag{9}$$

where $\alpha$ is an empirically determined weighting value and $\oplus$ denotes the element-wise addition of the global label probability $P_g$ to each of the obtained local frame probabilities $P_l$. A similar fusion process has been repeated for the outputs of the RGB-D and joint decision pools. The obtained frame-wise labels have been grouped following a major voting of the central frame labels, as detailed in [22], to produce a unique global label for each extracted action segment. The evaluation of our approach performance on the ground truth and the temporally segmented partitions is going to be presented in the next section.

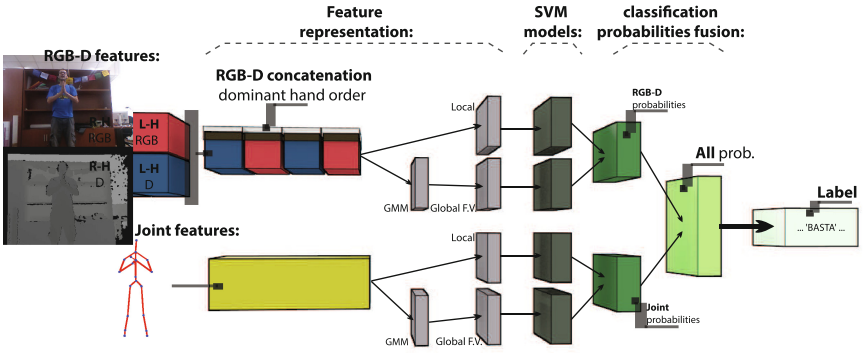



**Fig. 3.** Our approach learns different SVM models: those at a frame-wise local scale and those at a global scale using the Fisher vectors (F.V.). It first fuses global decisions with local ones, then combines the RGB-D probabilities with those of the joints

## 6 Experiments and Results

We present hereafter our used dataset, then evaluate our combined local-global approach using the ground truth and finish with the overall positioning of our solution within the state of the art.

**Used Dataset:** The experiments presented in this work have been carried out on a subset of the Chalearn gesture challenge 2014 dataset [5]. It is organised into 3 subsets relative to learning, validation and test stages. Each of them contains multi-modal data relative to the RGB, depth and user mask videos in addition to the skeletal joints streams as shown in Fig. 4. Each recording is associated to one person performing Italian sign language actions in front of the Kinect sensor. The considered action vocabulary is at a number of 20 different actions. They are labelled as follows: *1.'vattene', 2.'vieniqui', 3.'perfetto', 4.'furbo', 5.'cheduepalle', 6.'chevuoi', 7.'daccordo', 8.'seipazzo', 9.'combinato', 10.'freganiente', 11.'ok', 12.'cosatifarei', 13.'basta', 14.'prendere', 15.'cenepiu', 16.'fame', 17.'tempofa', 18.'buonissimo', 19.'messidaccordo',* and *20.'sonostufo'*.

**Performance Evaluation:** The evaluation of our approach performance has been carried out using a learning set of 80 folders and a test set of 20 other ones (i.e. *Sample0081* to *Sample0100*). Our experiments have led us to choose an empirical value for the weight $\alpha = 0.4$ to allow more influence for the local decision probabilities. Then, the fusion of the decisions relative to both RGB-D and joint probabilities has been applied with equal weights (i.e. $\alpha = 0.5$). The obtained performances using the ground truth action-segments are summarised in Table 1.

The behaviour of the learned global classifiers has allowed the generation of 100% frame-rates if the action is recognised or of 0% if not. This explains the relative superiority of the local-scale classifier (84.21% against 57.06% for the

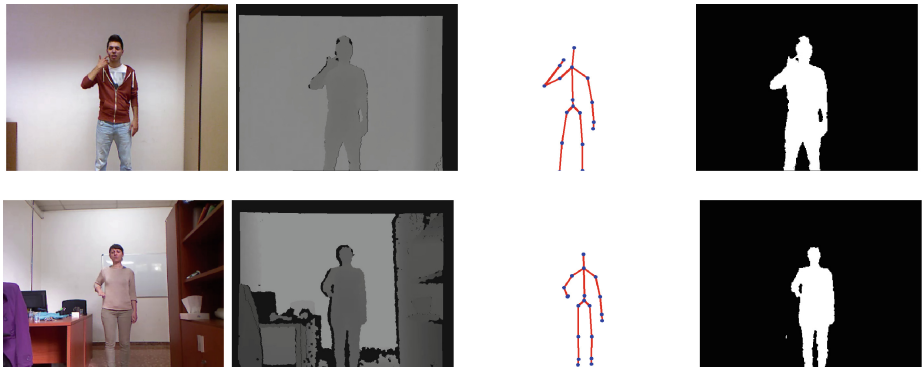


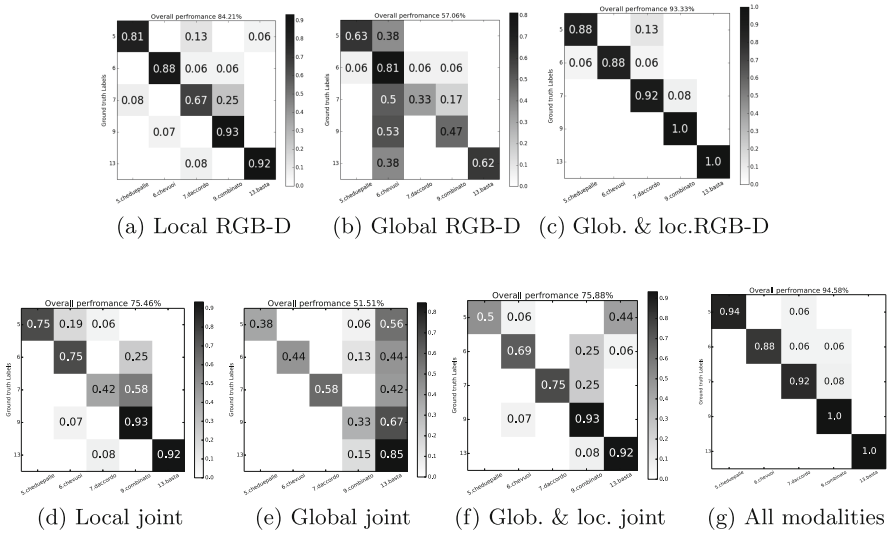

**Fig. 4.** Illustrations of the modalities offered by the Chalearn gesture 2014 dataset: a- RGB, b- depth, c- skeleton and d- user mask for the case of the actions: 'buonissimo' and 'ho fame'

**Table 1.** Evaluation of our approach processing stages using the ground truth in the case of bi-handed actions

| | RGB-D local | RGB-D global | Joint local | Joint global |
|---|---|---|---|---|
| Per modality RR | 84.21% | 57.06% | 75.46% | 51.51% |
| Loc. and Glob. RR | 93.33% | | 75.88% | |
| Multi-modal RR | 94.58% | | | |

RGB-D samples given in Table 1). The combination of both local and global probabilities has noticeably improved the performances, as shown in the confusion matrices of Fig. 5. The fusion of the decisions obtained from both RGB-D and joint classifiers has further improved the results to reach 94.58% in the case of the bi-handed action labels.

As demonstrated in Fig. 5, we have been able to obtain recognition gains for both global-local and multi-modal fusion stages. Compared to the results obtained in [22] using the same set of descriptors, the performances have risen from 81.01% on the ground truth to reach 94.58%. In comparison to the works presenting performance evaluation on the Chalearn 2014 dataset ground truth [6,13,18], our solution presents the advantage of reaching 100% recognition rates for multiple action classes. This is, for example, the case for the labels 9 and 13 in Fig. 5-g.




(a) Local RGB-D (b) Global RGB-D (c) Glob. & loc.RGB-D

(d) Local joint (e) Global joint (f) Glob. & loc. joint (g) All modalities

**Fig. 5.** Evaluation of the performance gain for the different labels through the confusion matrices generated using the RGB-D and joint streams for the bi-handed action family

The developed method has proved to be efficient for recognition improvement compared to similar methods [13,22]. Our solution brings two major advantages. The first is related to the exploitation of the action extraction stage for the service of classification specialisation. The second is related to the combination of the local and global classifier decisions. The presented rates are dedicated to the bi-handed actions and similar improvements can be obtained for the one-handed actions.

## 7 Conclusion and Perspectives

We have presented in this paper an approach for the extraction and the recognition of human actions from continuous streams of multi-modal data. We have contributed by proposing a fusion of the decisions offered by the Fisher vector's global representation with those obtained from the frame-wise local descriptions of actions. We have also put forward a combination strategy for features extracted from the RGB, depth and joint data streams offered by the Kinect sensor beside a specialisation approach of the classifiers for the one-handed and bi-handed actions. The experiments on the Chalearn gesture challenge dataset have proven the effectiveness of our approach for the recognition improvement.

Future perspectives for our work include the evaluation over the whole test dataset offered by the Chalearn gesture challenge using the Jaccard index [5] and the investigation of more advanced fusion strategies derived from the probabilistic theory [28]. In addition, we are in the process of considering data regulation strategies [7] as the extra classes of non-vocabulary and non-motion segments come with over-balanced learning population rates within our dataset.

## References

1. Alippi, C., Boracchi, G., Roveri, M.: Just-In-Time Classifiers for Recurrent Concepts. IEEE Transactions on Neural Networks and Learning Systems **24**, 620–634 (2013)
2. Bao, L., Intille, S.S.: Activity recognition from user-annotated acceleration data. In: Ferscha, A., Mattern, F. (eds.) PERVASIVE 2004. LNCS, vol. 3001, pp. 1–17. Springer, Heidelberg (2004)
3. Belongie, S., Malik, J., Puzicha, J.: Shape Matching and Object Recognition Using Shape Contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**, 509–522 (2002)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, pp. 886–893. IEEE Press, San Diego (2005)
5. Escalera, S., Baró, X., Gonzàlez, J., Bautista, M.A., Madadi, M., Reyes, M., Ponce-López, V., Escalante, H.J., Shotton, J., Guyon, I.: ChaLearn looking at people challenge 2014: dataset and results. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014 Workshops. LNCS, vol. 8925, pp. 459–473. Springer, Heidelberg (2015)
6. Evangelidis, G.D., Singh, G., Horaud, R.: Continuous gesture recognition from articulated poses. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014 Workshops. LNCS, vol. 8925, pp. 595–607. Springer, Heidelberg (2015)

7. Gazzah, S., Essoukri Ben Amara, N.: Writer identification using modular MLP classifier and genetic algorithm for optimal features selection. In: Wang, J., Yi, Z., Żurada, J.M., Lu, B.-L., Yin, H. (eds.) ISNN 2006. LNCS, vol. 3972, pp. 271–276. Springer, Heidelberg (2006)
8. Guyon, I., Athitsos, V., Jangyodsuk, P., Escalante, H.J.: The ChaLearn Gesture Dataset (CGD 2011), MVA (2013)
9. Hernandez-Vela, A., Bautista, M.A., Perez-Sala, X., Ponce-Lpez, V., Escalera, S., Bar, X., Pujol, P., Angulo, C.: Probability-based Dynamic Time Warping and Bag-of-Visual-and-Depth-Words for Human Gesture Recognition in RGB-D. Pattern Recognition Letters **50**, 112–121 (2014)
10. Ibanez, R., Soria, A., Teyseyre, A., Campo, M.: Easy gesture recognition for kinect. AES **76**, 171–180 (2014)
11. Ortiz Laguna, J., Olaya, A.G., Borrajo, D.: A dynamic sliding window approach for activity recognition. In: Konstan, J.A., Conejo, R., Marzo, J.L., Oliver, N. (eds.) UMAP 2011. LNCS, vol. 6787, pp. 219–230. Springer, Heidelberg (2011)
12. Laptev, I.: On space-time interest points. IJCV **64**(2–3), 107–123 (2005)
13. Liang, B., Zheng, L.: Multi-modal gesture recognition using skeletal joints and motion trail model. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014 Workshops. LNCS, vol. 8925, pp. 623–638. Springer, Heidelberg (2015)
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**(2), 91–110 (2004)
15. Mori, G., Malik, J.: Recovering 3d Human Body Configurations Using Shape Contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence **7**, 1052–1062 (2006)
16. Neverova, N., Wolf, C., Taylor, G.W., Nebout, F.: Multi-scale deep learning for gesture detection and localization. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014 Workshops. LNCS, vol. 8925, pp. 474–490. Springer, Heidelberg (2015)
17. Oreifej, O., Zicheng, L.: HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In: CVPR, pp. 716–723. IEEE Press, Los Alamitos (2013)
18. Peng, X., Wang, L., Cai, Z., Qiao, Y.: Action and gesture temporal spotting with super vector representation. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014 Workshops. LNCS, vol. 8925, pp. 518–527. Springer, Heidelberg (2015)
19. 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, Part IV. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010)
20. Pigou, L., Dieleman, S., Kindermans, P.-J., Schrauwen, B.: Sign language recognition using convolutional neural networks. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014 Workshops. LNCS, vol. 8925, pp. 572–578. Springer, Heidelberg (2015)
21. Rostamzadeh, N., Zen, G., Mironică, I., Uijlings, J., Sebe, N.: Daily living activities recognition via efficient high and low level cues combination and fisher kernel representation. In: Petrosino, A. (ed.) ICIAP 2013, Part I. LNCS, vol. 8156, pp. 431–441. Springer, Heidelberg (2013)
22. Seddik, B., Gazzah, S., Essoukri Ben Amara, N.: Hands, face and joints for multi-modal human-actions spotting and recognition. In: EUSIPCO (2015)
23. Seddik, B., Gazzah, S., Chateau, T., Essoukri Ben Amara, N.: Augmented skeletal joints for temporal segmentation of sign language actions. In: IPAS, pp. 1–6. Hammamet (2014)

24. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from a single depth image. In: CVPR (2011)
25. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from RGBD images. In: ICRA, pp. 842–849 (2012)
26. Vedaldi, A., Fulkerson, B.: VLFeat: An Open and Portable Library of Computer Vision Algorithms (2008)
27. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV, pp. 3551–3558 (2013)
28. Yazid, H., Kalti, K., Essoukri Ben Amara, N.: A performance comparison of the Bayesian graphical model and the possibilistic graphical model applied in a brain MRI cases retrieval contribution. In: SSD, pp. 16. IEEE Press, Hammamet (2013)