IET Computer Vision

Research Article

IET Journals
The Institution of Engineering and Technology

Human-action recognition using a multi-layered fusion scheme of Kinect modalities

ISSN 1751-9632
Received on 11th September 2016
Revised 23rd January 2017
Accepted on 11th March 2017
E-First on 18th August 2017
doi: 10.1049/iet-cvi.2016.0326
www.ietdl.org

Bassem Seddik[1,2], Sami Gazzah[1], Najoua Essoukri Ben Amara[1] ✉

[1]*LATIS Laboratory, National Engineering School of Sousse, University of Sousse, Sousse, Tunisia*
[2]*National Engineering School of Sfax, University of Sfax, Sfax, Tunisia*
✉ *E-mail: Najoua.Benamara@eniso.rnu.tn*

Abstract: This study addresses the problem of efficiently combining the joint, RGB and depth modalities of the Kinect sensor in order to recognise human actions. For this purpose, a multi-layered fusion scheme concatenates different specific features, builds specialised local and global SVM models and then iteratively fuses their different scores. The authors essentially contribute in two levels: (i) they combine the performance of local descriptors with the strength of global bags-of-visual-words representations. They are able then to generate improved local decisions that allow noisy frames handling. (ii) They also study the performance of multiple fusion schemes guided by different features concatenations, Fisher vectors representations concatenation and later iterative scores fusion. To prove the efficiency of their approach, they have evaluated their experiments on two challenging public datasets: CAD-60 and CGC-2014. Competitive results are obtained for both benchmarks.

1 Introduction

The recognition of simple human gestures, mid-level actions or even complex behaviours has been in a steady progress during the last years. The works focusing on Red-Green-Blue (RGB) still images [1], depth streams or skeletal poses [2] have made available real-world applications ranging from entertainment to human interfaces, and even outdoor surveillance [2–4]. Sensors such as Microsoft's Kinect allowed synchronised joint, RGB and depth indoor streams and favoured the leaning towards multi-modal three-dimensional (3D) approaches [5]. With the new advances in human pose estimation from indoor and outdoor streams [6, 7], in addition to the progress within the RGB-D-based 3D human-action recognition [8] this field still has a promising future. Each of these data inputs has received a large number of dedicated works, but fewer have benefited from all of them. We focus in this paper on the efficient combination of these modalities for the purpose of human-action recognition. We take advantage here from the inter-correlations existing between the different Kinect streams.

The main challenges in human-action recognition lie in handling: (i) the frame variations unseen during the learning stage for a given action-label and (ii) the possible similarities in sequences having different labels [1, 2, 5]. This is specially the case as most human-action datasets (e.g. KTH [9], JHMDB [10] and CAD-60 [11]) offer pre-segmented action sequences of N frames with unique per-sequence labels. Even with datasets proposing continuous streams (e.g. CGC-2012 [12] and CGC-2014 [13]), the same challenges persist after the sub-actions segmentation stage. Fig. 1*a* illustrates sample frame variations within the CAD-60 [11] and CGC-2014 [13] datasets. It shows the meaningful frames in bigger size compared with the confusing ones.

In recent advancements, learning algorithms leaned towards using large-scale data and generating the higher order statistics [14]. They can be achieved using, for instance, deep neural architectures such as the convolutional neural networks (CNN) [15], or sparse bags of visual words (BoVW) representations such as the Fisher vectors (FV) [16]. When applied to image-based computer vision tasks (e.g. image classification, object detection), remarkable performances have been obtained using the CNN architectures [13, 14]. However when applied to human-action or multi-modal recognition tasks, additional complementarity has been brought using motion features [17], frame-wise descriptors extracted from the joint modality [18] or trajectory-relative FV [19].

This work is related to the sparse FV-based family of methods. We contribute by combining the global BoVW representations with the local frame-wise descriptors and focus on the efficient fusion of the joint, RGB and depth modalities for the purpose of human-action extraction and recognition. We exploit here the joint-relative descriptors efficiency and the sparse representations high linear separability. We extend our previous research [20, 21] with what follows: (i) compared to [20], we improve the joint normalisation, the RGB and depth feature binning and use richer BoVW representations. Furthermore, (ii) we evaluate the local–global approach of [21] within multiple fusion configurations using a variety of features concatenations. In addition, we concatenate the generated FV representations and apply later scores fusion at the decision level. (iii) In addition to the CGC-2014 dataset [13], we

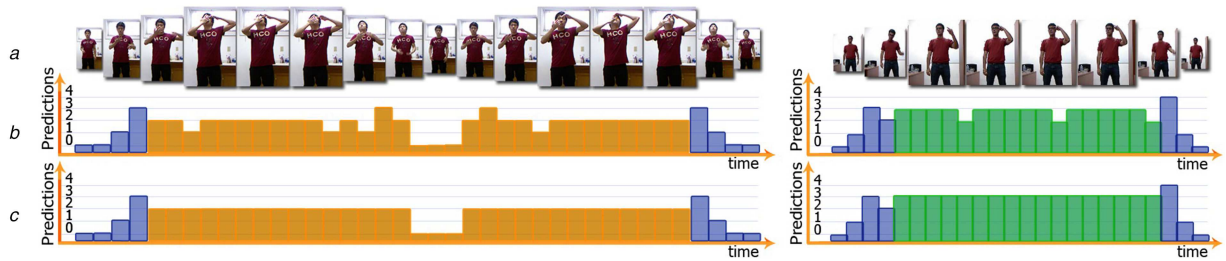

Fig. 1 *Our ImLoDe obtained on the CAD-60 and CGC-2014 datasets* ***(a)*** *Action frame sizes indicate their relativeness to the action labels,* ***(b)*** *Local labels in row 2 are unstable specially at the start and end of actions,* ***(c)*** *Local labels are improved after fusion with the global decisions*

also validate our approach on the CAD-60 dataset [11], bringing new challenges.

In what follows, we review the recent works related to local feature extraction, global BoVW representation and their multi-modal fusion strategies. After that, Section 3 gives the different stages of our approach. Next, we detail our used datasets and our experimental procedure to present later our obtained results in Section 4. Finally, we discuss the obtained performances in Section 5 to conclude and give future perspectives at the end.

2 Related work

According to the study of Jhuang *et al.* [10] for understanding the action-recognition process, the highest uni-modal performances are held by the high-level joint modality in front of RGB and depth low-level modalities. Different bi-modal [22] and tri-modal [18, 23] combinations are also possible from the Kinect streams. In what follows, we consider the descriptors related to limited numbers of action frames as local. In contrast, as the BoVW representations generally encode all the action frames, they are considered as global. We review here the recent advances in feature extraction, BoVW representation and multi-modal fusion, then present our literature analysis.

2.1 Local feature extraction

Hereafter, we review the literature hand-crafted features for the joint, RGB and depth modalities.

Joint based: A lot of successful approaches have relied solely on local joint descriptors for action recognition. They have generally relied on an efficient skeletal normalisation as a first pre-processing stage [24, 25]. Then, they have applied computations relative to the joint positions [24, 26], their accelerations and velocities [23, 27–29] their inter-distances [18, 20, 30, 31] and their rotation angles from different referentials [18, 25, 32]. More advanced approaches have also relied on higher layers of joint-descriptor representations such as the curve lie-groups [25] or the Kendall's manifold [33] in combination with specific decision measures.

RGB based: Still-image feature extractors have also been in constant progress starting with spatio-temporal interest points (STIP) and their extension to 3D temporal volumes [9]. As features, these extractors produce histograms of oriented gradients (HOG) and histograms of optical flows (HOF) [34]. Multiple multi-modal works relied on the HOG and/or HOF descriptors to produce their low-level local features [11–13, 22, 27, 35]. Dense extractors came afterwards and added descriptors relative to the motion boundaries and to the trajectories [36]. The improved dense trajectories (iDT) proposed in [37] relied on a human detector to reduce the trajectory-related features and set the state-of-the-art for hand-crafted image features.

Depth based: Shape-context derived descriptors have been widely used to produce histograms of orientations and distances out of shape landmarks [38] (also with RGB and joint [2]). Other widely used descriptors are the depth temporal-difference energies (displacements). They proved their efficiency in [20, 39] and brought better performances than the RGB-based displacements [12]. Other solutions relied on depth-normals computation [40, 41], STIP-based features [42] or even noise-tolerant depth networks [43].

2.2 Global BoVW feature representation

The core of the BoVW approaches is a meaningful grouping of the local features into more coherent global representations. This grouping is possible with a temporal max, sum or weighted pooling [44], or more commonly, by finding the K most-action-representative local descriptors (codebook centroids) and then attributing the less-important ones to them. To find these important codebook centroids and use them to encode all the features, a number of increasingly efficient methods have been used [45].

Earlier approaches relied on a hard-assignment strategy using K-means clustering to find the K centroids and then applying vector quantisation (VQ) for encoding [22, 35]. A following improvement was brought by soft-assignments through the probability distributions of the Gaussian mixture models (GMM) [45]. In addition to providing the mean centres, the GMMs allow a better encapsulation of the features distribution variation [35] using the expectation–maximisation algorithm for encoding. Other works rather focused on the encoding of sparse representations from over-complete codebooks (or dictionaries). These methods proved their superiority to the VQ using, for instance, the orthogonal matching pursuit and the sparse coding algorithms [45].

More recently, high-dimensional feature representations such as the FV have been successfully used with large-scale datasets [16]. Coupled with the iDT descriptors [37], they led to highly ranking performances [45, 46]. Four steps are followed to generate the FV: in addition to the codebook generation and the feature encoding, FV first rely on principal components analysis (PCA) with whitening as a preprocessing, then on the global pooling and the double normalisation (power-norm and L^1-norm or L^2-norm) as a post-processing. Different FV pooling strategies use K-means hard-assignments to produce the vector-of-locally-aggregated-descriptor representation suitable for resource-limited systems [45]. Other attempts to improve jointly the representation and classification performances, introduced the max-margin dictionary learning strategies for better intra-class discrimination [47, 48].

2.3 Multi-modal fusion strategies

Multi-modal fusion allows gaining the modalities' complementarity [5], useful especially when designing real-world-working applications [3]. Fusion strategies can be classified into four families: (i) feature-concatenation based, (ii) representation based, (iii) score based and (iv) hybrid ones combining multiple fusion levels [5, 45].

The concatenation of the low-level descriptors, discussed in Section 2.1, can be made using full or conditional combinations. Such strategies were successfully utilised in [37] to fuse multiple features coming from the same RGB modality. They were also used in [22] with both the RGB and depth modalities to reduce feature variability. The use of the joint positions to extract local features from the RGB or depth space has been also commonly used [13, 23]. It can be seen as a conditional fusion of those modalities.

The fusion of multiple BoVW representations is a very efficient source of performance improvement. The obtained feature representations can be then simply concatenated [46] or, in more advanced approaches, pooled from different features [44]. The fusion at the representation level is also found within the CNN architectures with multi-modal inputs [49, 50].

At the score level, the fusion is made out of modality-relative scores. These scores can be generated, for instance, within Bayesian or Markovian probabilistic models, or deduced from the decisions of discriminative models such as Support Vector Machines (SVMs) or random forests. In both cases, the fusion of modal-weighted scores provides better performances than the separate ones [5]. While the multi-modal-decision fusion can be implemented within the neural architectures [18], the weighted fusion of multiple SVM classifiers' outputs proves also efficiency [32].

The fourth family of hybrid-fusion methods make call to combinations of the aforementioned fusions. In [51], the authors proposed a two-layered model that combines SVM-based probabilities with hidden conditional random fields to analyse continuous human actions. Peng *et al.* [45] proposed a hybrid super-vector approach combining multiple BoVW architectures decisions.

2.4 Literature analysis and approach proposition

To handle the pre-stroke and post-stroke artefacts [13, 23] shown in Fig. 1*b*, we consider merging the decision scores of local frame-wise classifiers with global BoVW ones, thus generating improved local decisions (ImLoDe) as in Fig. 1*c*.

The literature states that high-level features (i.e. the joint) bring gains of 19 – 29% [10] in front of the RGB and depth. This is

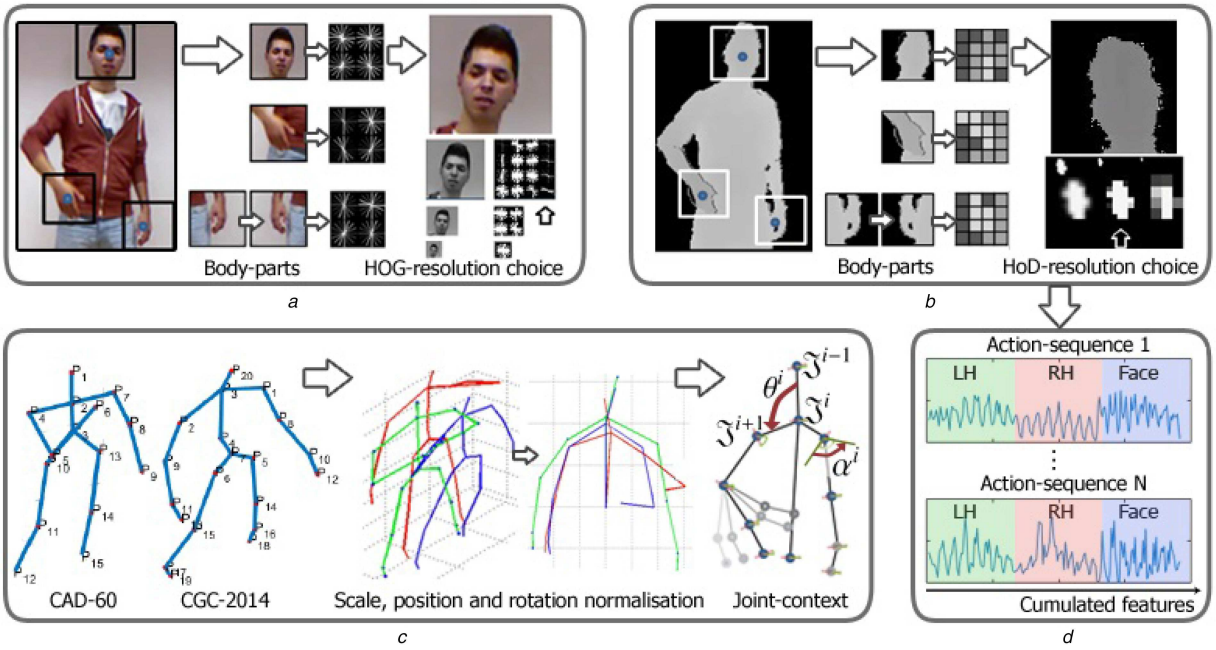

Fig. 2 *Considered features: the local ones are reinforced by global FV representation* ***(a)*** *Local joint descriptors from CAD-60/CGC-2014 for trajectory, speed and shape discrimination,* ***(b)*** *Local HOG descriptor from RGB for appearance discrimination,* ***(c)*** *Local HoD descriptor from depth for displacement discrimination,* ***(d)*** *HoD-BoVW: global temporal pooling of both hands and face HoD*

confirmed by the performances obtained with recent frame-wise joint descriptors [18, 29]. For this reason, we incorporate rich joint-based local features. We also design complimentary pose and motion descriptors from the RGB and depth modalities. To handle the depth inherent noise, we also propose a dedicated pooling-based descriptor. To the difference with [46], we seek to avoid the bottleneck of massive-dense features (three times the dataset size) and their resource consumption problems when concatenating multiple FV representations [45].

In order to choose the optimal global representation, we refer to the BoVW study of Peng *et al.* [45]. Our choice goes to the FV representations as they scale to large datasets [16]. We also decide to concatenate different FV representations generated from multiple local descriptors as they proved particular efficiency [46]. One question to answer at this level is: how to choose our local descriptor-sets in order to obtain efficient later FV-concatenated representations? For this purpose, we evaluate three feature concatenation configurations guided by: (i) full concatenation of all descriptors, (ii) concatenation of features according to their body parts relativeness as in [18] and (iii) concatenation of the features relative to each separate modality (i.e. the joint, RGB and depth). The best performing setup is selected for later FV representations concatenation.

Accordingly, our solution is positioned within the hybrid family of methods. At the feature level, we concatenate different multi-modal descriptors. At the representation level, we also concatenate the different FV representations generated. At the decision level, we merge the scores relative to local–global and multi-modal SVM classifiers with probabilistic outputs. The details of our approach are presented in the next section.

3 Approach details

Based on our previous work [20, 21], we present hereafter our local feature extraction improvements. We also detail the two proposed global representations relative to the HoD-BoVW and the FV. Later, we propose different fusion schemes to evaluate for performance effectiveness.

3.1 Per-modality feature extraction

As shown in Fig. 2, we define a complementarity feature setup designed according to the state-of-art hand-crafted iDT descriptor proposed by Wang and Schmid [37]. For each of the trajectories, HOG, HOF and motion-boundary feature sets, we provide more compact, yet functionally-equivalent, features coming from the joints, RGB and depth modalities. Relying on the joint 2D positions for localised spatial analysis, we are able to sparsely track body parts and generate reduced amounts of localised descriptors around the hands and face. On the temporal axis, we opt for a dense sampling every t frames. The later BoVW global representation allows additional feature-size reduction.

Joint modality: The joint modality offers the trajectory descriptors in addition to the global posture and body motion (e.g. optical flow) information. As we find that our joint descriptors in [20] suffered from inherent-noise and size-normalisation problems, we improve the skeletal normalisation stage by resetting the body rotation while preserving the inter-joint bone sizes. From the raw 3D joint positions $J^i_{x,y,z}$, where $i = 1 \ldots N$ is the joint index, we re-adapt the skeletal models of the Florence3D and UTK skeletal models [25] (having 15 and 20 body joints as in Fig. 2*c*) to generate stabilised 3D joint positions $\mathfrak{J}^i_{x,y,z}$ and their 2D projected positions $\mathfrak{J}^i_{x,y}$ with reference to the hip-centre joint. The different bone sizes are rescaled to match model-predefined distances. From all 15 CAD-60 body joints (respectively 14 upper joints for CGC-2014), we generate the temporal δ and δ^2 gradients [29] and the joint pair-wise distances $\mathfrak{J}^i_{pw}$ to produce the joint-context features as in [20]. We also convert the Euler joint-rotation angles α^i to quaternions and use (1) to compute the inter-bone rotation quaternion angles θ^i shown in Fig. 2*c*

$$\theta^i = \arccos \frac{(\mathfrak{J}^i - \mathfrak{J}^{i-1}) \times (\mathfrak{J}^i - \mathfrak{J}^{i+1})}{||\mathfrak{J}^i - \mathfrak{J}^{i-1}|| \times ||\mathfrak{J}^i - \mathfrak{J}^{i+1}||} \quad (1)$$

The concatenated feature vectors $[\mathfrak{J}_{x,y,z}, \mathfrak{J}_{x,y}, \mathfrak{J}_{pw}, \delta, \delta^2, \alpha, \theta]$ are sized 371 and 368 for CAD-60 and CGC-2014 joint sets, respectively. The redundancy in the obtained descriptor (e.g. $\mathfrak{J}_{x,y,z}$ and $\mathfrak{J}_{x,y}$) proves useful. The later PCA and whitening allows the dimensionality reduction [45].

RGB modality: As the joints save the general posture of the body, the RGB modality is designed to save the localised appearance of the face and hands. Based on previous research, we consider windows of 48×48 pixels for CAD-60 (96×96 for CGC-2014) and extract per-frame HOG descriptors of nine orientations and 4×4 cells using the joint 2D projected positions for the face and both hands. Fig. 2*a* shows the visual representation

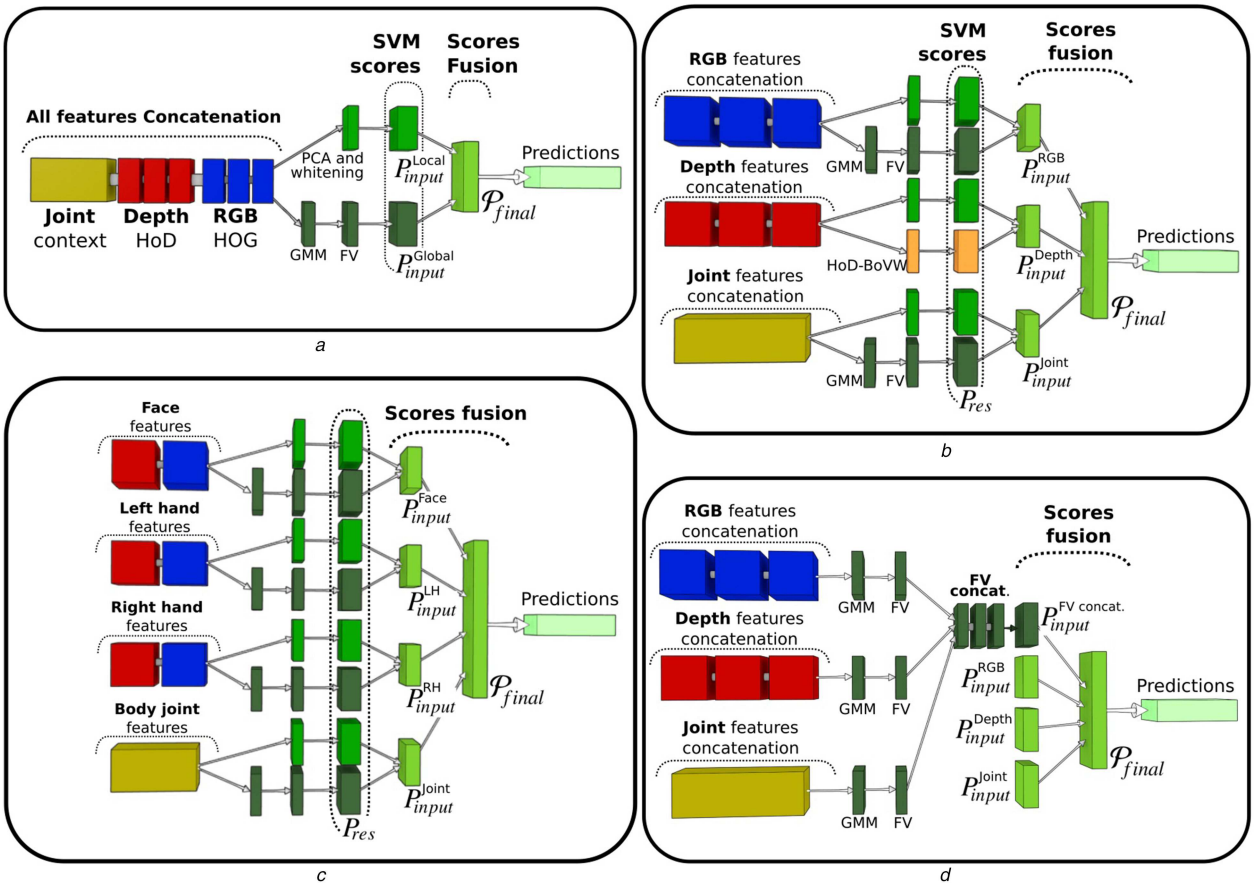

Fig. 3 *Different fusion strategies evaluated* ***(a)*** *Full-concatenation-based fusion,* ***(b)*** *Modality-based concatenation guided by fusion of RGB, depth and joint probability scores,* ***(c)*** *Body-part-based concatenation and fusion,* ***(d)*** *FV representations concatenation and fusion with the modality-based scores*

of the selected configuration containing a total of $9 \times 16 \times 3 = 432$ features.

Depth and mask modalities: The depth stream is mainly useful for measuring the displacements (comparable to the optical flows [34]) of the silhouette. Computing measures related to the depth differences gave interesting results in [12, 39]. We propose here to compute the local histograms of displacements (HoD) descriptor given in (2) from three sub-windows (see Fig. 2*d*) relative to the face, the right and left hands obtained using the 2D joint projections

$$\mathrm{HoD}(t) = S \times [(d(t+1) \times m(t+1)) - (d(t-1) \times m(t-1))] \quad (2)$$

where S is a scale reduction factor, $d(t)$ and $m(t)$ are, respectively, the depth and mask frames at instant t. Fig. 2*b* illustrates our HoD configuration retained on grids of 8×8 pixels for a total of $64 \times 3 = 192$ features.

HoD-BoVW: Each of the aforementioned features receives later global FV for the sake of performance improvement. However, when dealing with the depth modality, we find that weak performances are obtained with the FV representation due to high inherent noise. To obtain a reliable depth-based global representation, we opt for a pooling-based methodology. Using the obtained local HoD features for each frame, we apply a temporal max-pooling for each sample action to produce the complementary global representation that we refer by HoD-BoVW in Fig. 2*d*. We opt for a temporal mean filtering to further reduce the raw signal variations. Compared with the similar method of [12], we avoid the actor-shifting problem by focusing on the body parts.

3.2 FV representation

FVs are a special case of the Fisher kernels combining the strengths of both generative and discriminative models [5]. For each of the sample action features $S = [x_1, \ldots, x_t]$ of size D, produced for all t available frames, we first apply PCA with whitening enabled [45] to de-correlate and reduce our data size. After that, we generate a GMM of K centroids associated with their π_k, μ_k and Σ_k parameters, respectively, relative to the prior probabilities, the means and the diagonal covariance matrices. They are associated to each GMM by the posteriori probability in the following equation:

$$\Gamma_{tk} = \frac{\exp[-(1/2)(x_t - \mu_k)^{\mathrm{T}} \Sigma_k^{-1} (x_t - \mu_k)]}{\sum_{l=1}^{K} \exp[-(1/2)(x_t - \mu_l)^{\mathrm{T}} \Sigma_k^{-1} (x_t - \mu_l)]} \quad (3)$$

where t and k are, respectively, the frame and the GMM centroid indexes. Each sample action S is represented using the concatenation of the mean and covariance partial derivatives of all x_t features as in (4) with $j \in [1, \ldots, D]$ being its dimension

$$\Phi(S) = [u_{j1}, v_{j1}, \ldots, u_{jK}, v_{jK}] \quad (4)$$

The generated FV has a size of $2 \times K \times D$ features composed of the mean and covariance's partial derivatives given in (5) and (6), respectively, noting σ_k as the standard deviation

$$u_{jk} = \frac{1}{n\sqrt{\pi_k}} \sum_{t=1}^{n} \Gamma_{tk} \left[\frac{x_{jt} - \mu_{jk}}{\sigma_{jk}} \right] \quad (5)$$

$$v_{jk} = \frac{1}{n\sqrt{2\pi_k}} \sum_{t=1}^{n} \Gamma_{tk} \left[\left(\frac{x_{jt} - \mu_{jk}}{\sigma_{jk}} \right)^2 - 1 \right] \quad (6)$$

One last important step for the FV performance is a double normalisation stage using first the function $f(x) = \mathrm{sign}(x) \times \sqrt{|x|}$ and then the L^2-normalisation. As depicted in Fig. 3, each of the concatenated local frame-wise descriptors receives a GMM modelling then a FV representation.

Fig. 4 *Samples of left-handed and right-handed performances and their relative modalities for* ***(a)*** *CAD-60 dataset with the newly generated user-mask streams,* ***(b)*** *CGC-2014 dataset*

3.3 Fusion and recognition strategy

Fig. 3 gives an insight about the four principal fusion strategies evaluated within our ImLoDe pipeline. They have in common a selective frame-wise feature concatenation as an input. As an output, all of them merge the decision scores relative to SVM classifiers with probabilistic outputs. In addition, the local descriptors in all Fig. 3 fusion schemes follow a double path: (i) a first one applies signal whitening and PCA reduction [45] then generates local SVM scores; (ii) a second one builds BoVW representations and generates global SVM scores.

As shown in Figs. 3*a*–*c*, we build for each set of features their relative specialised SVM models: two for the full-concatenation-based pipeline (i.e. local and global paths), six for the modality-based one and eight for the body-part-based scheme [18]. For each considered input, we first combine the local ($P_{\text{input}}^{\text{Local}}$) and global ($P_{\text{input}}^{\text{Global}}$) resolutions decision scores. Then, we merge the per-input scores P_{input} to produce the final ones $\mathscr{P}_{\text{final}}$. For simplicity sake, we indicate in Figs. 3*b* and *c* only the two most efficient paths. Interestingly, the modality-based pipeline having only six models with a richer descriptors' population gives the best results, at least in a hand-crafted feature case (see experiment results in Section 4.2). To rapidly generate all required models, we pre-compute linear classification kernels from the learning and test sets as implemented in [45]. In Fig. 3*d*, we rely on the best feature-concatenation configuration to generate our concatenated FV. We keep here the best modality-based scores and combine them with FV-concatenation scores. The different classifiers considered bring complementarity in terms of the individual action-label precisions.

To find the best weighting coefficients that maximise the final score $\mathscr{P}_{\text{final}}$, an exhaustive search is iteratively applied for each separate label using the function $f(P)$ given by

$$\mathscr{P}_{\text{final}} = f(P_{\text{input}}) = \max_{C_{\text{input}}} \left(\sum_{j=1}^{m} C_{\text{input}}^{\text{T}} P_{\text{input}}^{(j)} \right) \tag{7}$$

where P_{input} notes the scores obtained after the first fusion step (per input) and $P_{\text{input}} = f(P_{\text{res}})$. Here P_{res} denotes the initial SVM probabilities (i.e. local and global resolutions) generated by the separate models. $C_{\text{input}} = \{c_{\text{input}}^{1}, \ldots, c_{\text{input}}^{m}\} \in [0, 1]$ are the per-label weighting coefficients obtained during the learning stage. Similar coefficients are also generated for each initial resolution. Our weighted fusion scheme advantage lies in its simplicity and extensibility to any other set of modalities (e.g. the sound as in [13, 18]). We highlight that we generate different weighting coefficients per action label during the score maximisation search. This operation reduces the impact of any weakly contributing modality and combines the distinctive classifiers capabilities with different action labels, i.e. when a first classifier finds all 'A' actions and another finds all 'B' actions.

When merging, the global whole-action scores are replicated for every frame and weighted with the local ones. We ensure thus the generation of local frame-wise decisions that improve with every fusion iteration. This choice is justified by the dominance of resting poses at the starting and ending of every sample in most datasets. This is the case of our both benchmarks where pre-strokes and post-strokes are observed. By keeping local decisions, our aim is to subtract these highly fluctuating frames and thus improve the action-recognition process.

4 Experimentation and results

We present hereafter our considered datasets and their evaluation metrics. The experimentation with the different fusion schemes are initialised on the CAD-60 dataset and validated using the CGC-2014 dataset. The concatenated-FV parameter choices are also analysed in detail.

4.1 Considered datasets and metrics

Our experimental choice was oriented towards two commonly used datasets within literature: the CAD-60 [11] and CGC-2014 [13] datasets. For CAD-60, the number of samples is limited (i.e. 68 action samples) but they span over long periods of 1000--4000 s [11]. In contrast, the CGC-2014 is a large-scale dataset with nearly 13,858 labelled instance actions, but performed in short sequences near 10 s [13]. We followed the same file formatting of the CGC-2014 in our experiments and transformed all the CAD-60 still frames into their equivalent CGC-2014 video streams. As the user mask was not offered in CAD-60, we generated it by segmenting the nearest depth region within a central sub-window. As illustrated in Fig. 4, we have created equivalent baselines for both datasets. Thus, easier benchmarking on both datasets is possible. We provide source code for the community at: https://github.com/bassemSeddik/ImLoDe.

Our experiments revealed that a major performance limitation comes from the left-handed and right-handed intrinsic variations of a same action (see Fig. 4). To overcome this challenge, we doubled the size of our learning population with its horizontally mirrored one for our both datasets.

CAD-60: This dataset concerns 12 classes of daily-life actions (e.g. wearing contact glasses, opening pill container, brushing teeth) in addition to two non-action classes relative to the still and random behaviours. It is performed only by four actors and offers images relative to the RGB and depth frames, beside the skeletal streams relative to 15 body joints. Its main challenge is having a one left-handed actor out of those present. As the performances are reported using cross-validation between the different actors, the hardest case is when learning from three right-handed actors and testing on the left-handed one. This is better demonstrated in Table 1, Figs. 5 and 6 focusing on actor-relative performances.

For comparison with the literature, we followed the same new person scheme presented in [11]. In this scheme, we learn from actors different from those used while testing. In our experiments, each time, we learn from three different actors then test over the actions performed by the one lasting. The iteration of this process produces the four-fold actors cross-validation performances. In addition to the accuracy (correctly found samples), we have

Table 1 Full-concatenation fusion accuracy on CAD-60 with four-fold cross-validation

	Actor 1		Actor 2		Actor 3		Actor 4		Average precision
	Loc	FV	Loc	FV	Loc	FV	Loc	FV	
	73.9	64.7	**87.1**	67.9	**54.1**	50	**77.1**	50.0	
Precision($\mathscr{P}_{\text{final}}$)	77.3		96.4		61.0		77.8		78.1

Bold values indicate the best results in each preset.

	Actor 1	**Actor 2**	**Actor 3**	**Actor 4**	***Average***
Precision ($\mathcal{P}_{final}$)	86.3	86.4	68.4	80,2	**81.7**

	Face			**Left hand**			**Right hand**			**Joint**	
	Loc	FV	HoD-BoVW	Loc	FV	HoD-BoVW	Loc	FV	HoD-BoVW	Loc	FV
Precision (P_{res})	45.7	**57.1**	**57.1**	44.1	**50.0**	39.3	41.9	**50.0**	29.8	**78.8**	54.8
Precision (P_{input})	64.6			51.6			58.7			**79.5**	

Fig. 5 *Body-part-based fusion accuracy on CAD-60 with four-fold cross-validation: actor 4 detailed*

	Actor 1	**Actor 2**	**Actor 3**	**Actor 4**	***Average***
Precision ($\mathcal{P}_{final}$)	89.9	96.3	70.4	91.0	**89.0**

	RGB		**Depth**			**Joint**	
	Loc	FV	Loc	FV	HoD-BoVW	Loc	FV
Precision (P_{res})	**53.2**	30.8	30.5	30.8	**38.5**	**56.3**	53.9
Precision (P_{input})	57.7		46.4			**61**	

Fig. 6 *Modality-based fusion accuracy on CAD-60 with four-fold cross-validation: actor 3 detailed*

generated the average confusion matrices and deduced the precision (correctly found labels) and recall average measures, then computed the F_1 score given by

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{8}$$

CGC-2014: The 2014 Chalearn Gesture Challenge dataset (referred as the Montalbano dataset in [13]) has the merit of being one of the first large-scale multi-modal human-action datasets. It contains a vocabulary of 20 Italian sign language (SL) actions (e.g. vattene, vieniqui, perfetto) and offers 940 batch folders divided into three parts relative to the development, validation and test sets. Each folder contains the RGB, depth, user-mask and skeletal streams (relative to 20 body joints) performed by over 40 actors at multiple sessions.

For initial evaluation, the CGC-2014 competition organisers allowed testing on the validation set. The final competition results are reported on the test-set using the Jaccard-Index (J-Ind) given by [13]

$$\text{Jaccard}(A, B)_{s,n} = \frac{A_{s,n} \cap B_{s,n}}{A_{s,n} \cup B_{s,n}} \tag{9}$$

where $A_{s,n}$ is the obtained segmentation with label $n \in [1, 20]$ and $B_{s,n}$ is the ground truth, given a sequence s. To the difference of the accuracy and precision measures, the J-Ind measures two aspects: (i) the correspondence between the segmented actions and their ground-truth (start, end) positions within the continuous streams and (ii) the percentage correctly labelled frames.

In what follows, the obtained results are first detailed for the CAD-60 dataset for the first three fusion schemes of Fig. 3. The results detailed afterwards for the CGC-2014 dataset are built on top of first ones. Reader may refer to [20, 21] for detailed experiments and results on CGC-2014.

4.2 Fusion-scheme evaluation on CAD-60 dataset

The analysis of the concatenation-based fusion (i.e. Fig. 3*a*) is presented in Table 1. It shows that when having the local joint features concatenated with those of the RGB-D, the local models perform better than the global ones for all actors. The combined local–global decisions improve the precision to reach 96.4% for the easiest actor and 61% for the hardest left-handed. The concatenation average precision is 78.1%.

Additional complementarity is brought by separating the feature contents. By analysing Fig. 5 relative to the finest body-part-based fusion scheme (i.e. Fig. 3*c*), we obtain a better average precision of 81.7%. We could at this level extract the HoD-BoVW global representation. Within this pipeline, the best results are detained by the global FV models for RGB-D and the local ones for the joint. Interestingly, the face-relative model outperforms those of the hands. This conclusion is also confirmed by the performances shown in Fig. 7*a*.

The best feature concatenation choice is provided by the modality-based fusion scheme (i.e. Fig. 3*b*). At this level, the HoD-BoVW representation has an advantage in front of the local and FV depth-based models. This improvement is explained by the richer set of descriptors obtained from each modality when compared with the body-part-based models. For the joint and RGB model, the local resolution is the best. We detail in Fig. 6 the hardest case of actor-3 (left-handed), where we reach a precision of 70.4%.

The last fusion scheme (i.e. Fig. 3*d*) concatenates the different FV representations generated into a unique sparse vector and combines its decision scores with the best modality-based ones. The FV-concatenated global representation depends on two parameters: The PCA-reduced feature dimensionality (D) and the number (K) of GMM are used for dictionary construction. Fig. 7*a* evaluates the per-modality accuracies using different GMM numbers for CAD-60 actor-1 samples. It shows that the reduced number of $K = 32$ (respectively 64) GMMs gives the best performances with the joint (respectively the face) features. In the presence of correlated features K is augmented to 256 (e.g. the all-feature-concatenated vector).

Fig. 7*b* overlays the cumulative impacts of the joint, RGB and HoD PCA energies on the FV-concatenation performance. Using factors around 0.86 and 0.88 of the joint and HOG respective feature energies, their concatenated FV representations give a stable performance of 95%. By adding a factor of 0.79 from the HoD feature energy, the accuracy oscillates between 95 and 100% on actor-1 samples.

As expected, the fusion of the FV-concatenation scores with those relative to the modality-based (i.e. $P_{\text{input}}^{\text{RGB}}$, $P_{\text{input}}^{\text{Depth}}$ and $P_{\text{input}}^{\text{Joint}}$) results in raising the average precision from 89% in Fig. 6 to

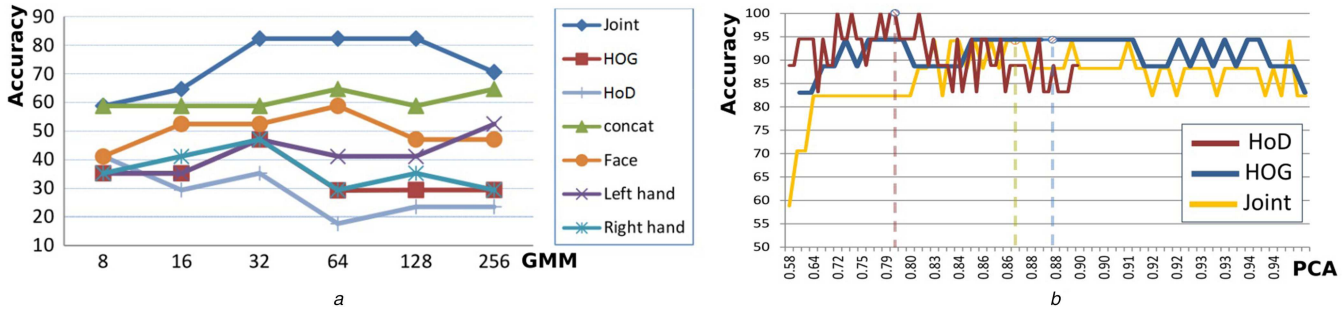

Fig. 7 *FV experimental setup for CAD-60* ***(a)*** *GMM number impact on each separate modality or input considered,* ***(b)*** *PCA energy impact on FV concatenation: joint, HOG and HoD cumulative factors indicated*

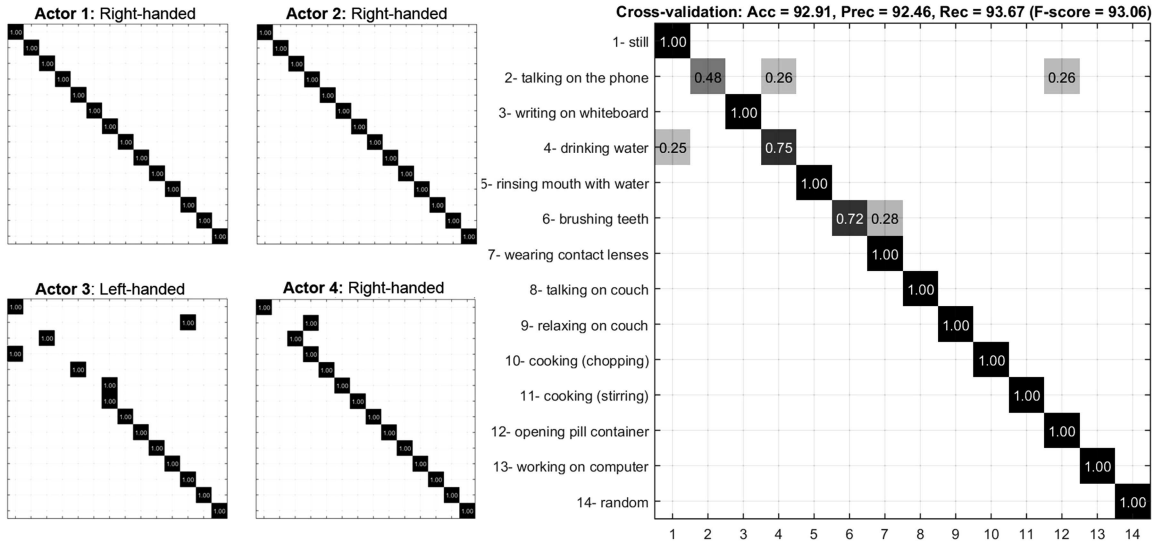

Fig. 8 *Cross-validation confusion matrices for CAD-60 12 actions, still and random motions*

92.5%. After the fusion, an additional grouping is applied by searching for the most occurring label within the central frames within $[(N/6), (5N/6)]$ (N is the sample length). As illustrated in Fig. 8, this operation allows the 100% precisions reported for actors 1 and 2.

4.3 Experimental validation on CGC-2014

The evaluation of our full-feature concatenation scheme with the CGC-2014 dataset resulted in 81% accuracy in [20]. Additional performances are also obtained in [21] by combining our local and global representations. In order to apply our best fusion scheme merging the scores of FV-concatenation and the per-modality classifiers, we start by optimising the FV parameters.

For the CGC-2014 dataset, the optimal parameter choices are detailed in Table 2 using a random subset of 100 sequences. The results are reported after duplicating the learning set (two-third of the considered population) with its symmetrical, resulting in 2784 action samples. By analysing the individual per-modality FV accuracies, the best performances (underlined values in Table 2) are obtained by maximising the K and D parameters. However, when considering the concatenated FV relative to Fig. 3*d*, higher performances are obtained using reduced parameter setups. The highlighted cells in Table 2 show that using only 25, 20 and 48% of the joint, HoD and HOG dimensionalities, respectively, give the best results using PCA reduction with whitening. In addition, only choices among 32 and 64 GMM are sufficient. When using the full CGC-2014 dataset, the FV-concatenation performance is improved from 79% (value shown in Table 2) to reach 87.4% with the same selected parameters.

The obtained results for each separate input within our best fusion scheme are detailed in Table 3. We show that the joint and the FV-concatenation relative classifiers offer the highest individual accuracies and that their combination in a first iteration leads to 91.5% accuracy. The fusion of the different scores available leads to our best ImLoDe performance of 93.4% after selecting the middle most significant frames within $[(N/3), (2N/3)]$ where N is the sample length. Fig. 9 details our separate label accuracies. It shows that, from the 20 present, 11 action labels are recognised with accuracies above 95%.

5 Discussion

Table 4 proves the competitiveness of our approach on the CAD-60 dataset. Using the F-score measure, our solution ranks in the second position and demonstrates a robust precision/recall ratio. It reaches 92.5% as action-label precision and a higher value of 92.9% as accuracy in terms of correctly labelled samples. Note that it admits the highest recall of 93.7% due to its labels grouping stage as shown on the confusion matrix in Fig. 8. The solutions [28, 31] are not reported as they use different joints and labels than existing ones [43].

For the CGC-2014 dataset, Table 5 shows that our solution provides an accuracy of 93.4% over the test-set ground-truth segmented actions and a robust J-Ind of 0.819. The CGC-2014 ranking concerns works having applied a temporal-segmentation stage out of the continuous streams. Contributions with upper J-Ind bounds such as [54] are not listed. In our case, we have used our joint descriptor with a binary SVM classifier to distinguish between the actions of interest and the rest of the motion [20]. Our temporal-segmentation stage finds the (start, end) action position-couples with an accuracy of 92.3% on the test-set.

By analysing Table 5, we find that the action-extraction stage impacts the performance difference perceived between final results (J-Ind) and the ground-truth accuracies. The two top-ranking methods [49, 18] exploit CNN features for recognition and detain very accurate neural-network-based temporal-segmentation stages. Our proposed SVM-based solution belongs to the next group of performances and shows a good balance between accuracy and action spotting. Interestingly, our solution outperforms the

Table 2a Continued

	Joint dimension reduction → resulting energy						
	1%→~**0.51**	2%→~**0.66**	4%→~**0.79**	6%→~**0.88**	8%→~**0.92**	16%→~**0.98**	25%→~**1.00**
8 GMM	13.6	23.1	28.0	33.2	37.4	49.7	56.3
FV-concat.	65.0	63.6	65.7	68.9	68.5	75.2	77.3
16 GMM	19.2	33.6	33.9	35.3	44.1	56.9	60.1
FV-concat.	68.5	67.1	68.9	71.3	69.2	**76.9**	75.5
32 GMM	23.1	32.2	33.6	37.8	44.4	58.4	60.1
FV-concat.	68.5	**72.4**	71.7	68.2	68.9	74.5	78.0
64 GMM	25.5	30.4	39.2	40.6	46.9	59.1	63.3
FV-concat.	67.8	70.3	**72.4**	69.6	68.9	73.8	**79.0**
128 GMM	24.5	35.7	40.6	46.5	50.7	63.6	66.4
FV-concat.	70.3	71.0	69.6	72.0	67.8	74.8	76.2
256 GMM	*26.9*	*37.8*	*42.0*	*48.3*	*50.4*	*64.7*	*66.4*
FV-concat.	**71.0**	71.7	70.9	**72.7**	**70.6**	76.6	77.3

Bold values indicate the best results in each preset.

Italic values are relative to the highest value of GMMs.

Table 2b Continued

	HoD dimension reduction→ resulting energy					
	5.5%→~**0.51**	10%→~**0.61**	20%→~**0.72**	40%→~**0.85**	60%→~**0.92**	70%→~**0.95**
8 GMM						
FV-concat.						
16 GMM	39.5	49.7	56.3	58.7	58.0	53.5
FV-concat.	66.8	70.6	72.7	75.9	76.6	**75.5**
32 GMM	40.9	48.6	61.2	62.9	61.9	55.9
FV-concat.	**70.6**	74.5	**78.0**	**77.3**	**76.9**	75.5
64 GMM	43.7	58.0	60.5	62.6	62.6	61.2
FV-concat.	69.9	**75.5**	78.0	76.2	75.2	75.5
128 GMM	*47.9*	*58.0*	*64.7*	*64.0*	*66.1*	*64.0*
FV-concat.	69.2	73.8	76.2	76.9	75.2	75.2
256 GMM						
FV-concat.						

Bold values indicate the best results in each preset.

Italic values are relative to the highest value of GMMs.

Table 2c Impact of the dimensionality *D* and GMM number *K* on FV accuracy with a random CGC-2014 subset

	HOG dimension reduction → resulting energy					
	6%→~**0.50**	12%→~**0.66**	18%→~**0.75**	24%→~**0.81**	48%→~**0.93**	64%→~**0.97**
8 GMM						
FV-concat.						
16 GMM	24.1	35.0	39.5	32.9	38.8	42.3
FV-concat.	68.2	72.4	71.7	71.3	72.4	73.4
32 GMM	32.5	38.1	41.6	42.0	45.5	47.2
FV-concat.	72.0	71.3	71.0	73.4	75.5	75.2
64 GMM	35.7	40.9	45.1	47.2	47.6	46.2
FV-concat.	**73.8**	70.6	74.1	**76.2**	**76.6**	74.8
128 GMM	*40.9*	*46.2*	*45.5*	*47.9*	*49.3*	*48.6*
FV-concat.	73.1	**76.2**	**75.2**	73.4	75.5	***75.2***
256 GMM						
FV-concat.						

Bold values indicate the best results in each preset.

Italic values are relative to the highest value of GMMs.

3DCNN-based approach of Wu *et al.* [50] and the comparable super-vector approach of Peng *et al.* [46] making use of the iDT descriptors and FV concatenation. In our case, a major performance optimisation is due to the joint high-level features.

Data augmentation impact: For both considered datasets, doubling the size of the learning set with its horizontally mirrored one resulted in noticeable gains of 2.6 and 2.8% in CAD-60 and CGC-2014 respective accuracies. In addition, for CGC-2014, learning from both the development and validation sets resulted in an additional gain of 1.3% to reach 93.4% as in Table 5. These results confirm that different data augmentation methods are prone to better performances. For CAD-60, it is interesting to note that the mirroring was not sufficient to leverage the left-handed actor variability (see actor-3 confusion matrix in Fig. 8) within our machine-learning configuration. This can be explained by other factors such as the clothe colours and the actor's position variations. Beyond the scope of this paper, augmenting the learning

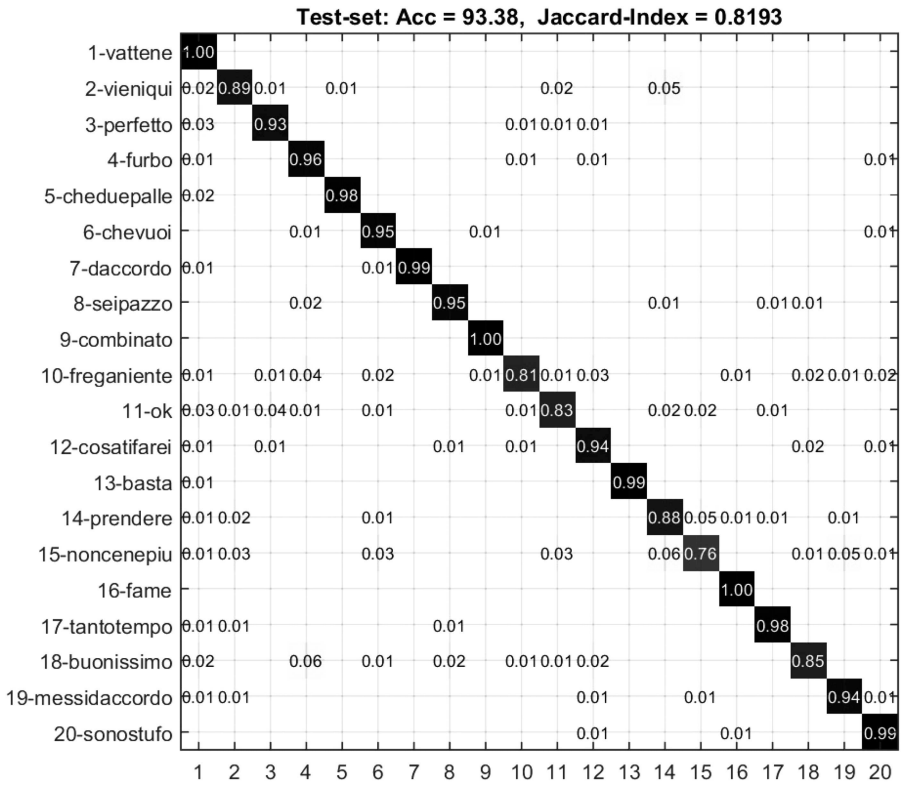

Fig. 9 *Confusion matrix for the CGC-2014 20 actions*

Table 3 Best fusion scheme performances on CGC-2014: modality-based and FV-concatenation scores fusion

	$P_{\text{input}}^{\text{RGB}}$	$P_{\text{input}}^{\text{Depth}}$	$P_{\text{input}}^{\text{Joint}}$	$P_{\text{input}}^{\text{FV concat}}$
Accuracy(P_{input})	77.7	73.7	86.2	**87.4**
Accuracy(first fusion iteration)		83.6		**91.5**
Accuracy($\mathscr{P}_{\text{final}}$)			**93.4**	
Accuracy(modality)		92.0		

Bold values indicate the best results in each preset.

Table 4 Comparison with the state-of-the-art on the CAD-60 dataset

Rank	Approach	Precision, %	Recall, %	*F*-score, %
1	Cippitelli *et al.* [32]	93.9	93.5	*93.7*
2	ImLoDe (this approach)	**92.5**	**93.7**	***93.1***
3	Parisi *et al.* [43]	91.9	90.2	*91,0*
4	Zhu *et al.* [42]	93.2	84.6	*88.7*
5	Gaglio *et al.* [26]	77.3	76.7	*77.0*
6	Koppula *et al.* [23]	80.8	71.4	*75.8*
7	Ni *et al.* [52]	75.9	69.5	*72.6*
8	Yang and Tian [53]	71.9	66.6	*69.2*
9	Sung *et al.* [11]	67.9	55.5	*61.1*

Bold values indicate our own obtained results.

Italic values are relative to the used metric for literature comparison.

set with such variations (e.g. translation, colour contrasting etc.) can improve the performances as in [15, 54].

Memory consumption and speed: We compare here our method to the solution of Peng *et al.* [46] based on the iDT descriptor. As our ImLoDe framework relies on a different set of features, we present in Table 6 the comparative evaluation of the two approaches. We refer to the detailed study presented by the same authors in [45] for the K and D parameters conditioning the size of FV and therefore the main memory consumption factor. In terms of speed, the encoding time is reported under the same number of frames (relative to 50 action samples) as used in [45]. Using our parameter setup indicated previously in Table 2, we are able to reach a J − Ind = 0.819 higher than 0.792 with the comparative method. Most interestingly we use lower FV sizes (∼41 k instead of ∼205 k), require less encoding time (1.62 instead of 45 s) and thus allow a better runtime.

For implementation, we have used a Xeon E3-1220 CPU of 3.1 GHz frequency running under a 64-bit Windows7 and 16 GB of RAM, making it a lower configuration than the one required in [45]. We have used python for feature extraction and Matlab for encoding and classification. Within our best fusion scheme, the runtime is conditioned by the speed of the local classifiers. Using all of them, our solution runs at 89 fps. Interestingly, by using only the fastest joint-based local classifier, our ImLoDe approach is only 1.9% less accurate while running at 402 fps.

6 Conclusion and perspectives

In this paper, we have presented a multi-layered approach for human-action recognition using the Kinect joint, RGB and depth modalities. We have focused on combining the performances of local frame-wise joint-based classifiers with global BoVW and FV-based ones. Four types of fusion schemes have been studied using a concatenation of: (i) all generated descriptors, (ii) body-part relative descriptors, (iii) modality specific ones and (iv) FV-based representations. The retained hybrid-fusion scheme – ImLoDe – is a framework combining multiple inputs at the description, representation and score levels. Two challenging benchmarks have been considered for experimentation: CAD-60 and CGC-2014. Using them, an optimisation of the FV scalability has been carried using a balance of the GMM number and dimensionality parameters. For both benchmarks, competitive ranking and runtime are obtained compared with the state-of-the-art.

A straight forward perspective is using better-performing methods to improve the per-modality decision scores. In this work, we have used our own set of per-modality complementary descriptors. Other available possibilities include using the CNN with the RGB and depth modalities [18]. For instance, the 3DCNN used in [50] offers the global analysis comparable to our local–global consideration. More advanced joint-based representations [25, 33] can also be adopted. Another interesting path is the investigation of different data augmentation operations such as image translation, colour contrasting and temporal variation on the FV performances. The FV scalability can, theoretically,

Table 5 Comparison with the state-of-the-art on the CGC-2014 dataset

Rank	Approach	Acc., %	J-Ind	Rank	Approach	Acc.	J-Ind
1	Pigou *et al.* [49]	97.2	*0.906*	8	Camgoz *et al.* [24]	—	*0.747*
2	Neverova *et al.* [18]	96.8	*0.870*	9	Evangelidis *et al.* [55]	94.0%	*0.745*
3	Monnier *et al.* [27]	97.9	*0.834*	10	Team Telepoints [13]	—	*0.689*
4	Chang [30]	92.6	*0.827*	11	Team Fortiss [13]	—	*0.649*
5	**ImLoDe (this approach)**	**93.4**	*0.819*	12	**Seddik *et al.*** [20] **(ours)**	81.0%	*0.618*
6	Wu *et al.* [50]	92.3	*0.809*	13	Liang and Zheng [39]	92.8%	*0.597*
7	Peng *et al.* [46]	97.0[a]	*0.792*	14	Team iva.mm [13]	—	*0.556*

[a]Result using the validation set.

Bold values indicate our own obtained results.

Italic values are relative to the used metric for literature comparison.

Table 6 Performance comparison in terms of the FV size in memory and the encoding time

	D	K	FV size ($2KD$)	Enc. time, s
FV Joint	92	64	11,776	0.22
FV HOG	207	64	26,496	1.19
FV HoD	38	32	2432	0.21
our features FV concatenation			**40,704**	**1.62**
iDT FV concatenation [46]	200	512	204,800	45

Bold values indicate our own obtained results.

encapsulate such variations at the cost of bigger memory requirements and slower encoding time. As a last extension, we are interested in the recognition of SLs with richer action vocabularies. As facial expressions [56] play an important role in SL, they deserve more focus as an additional modality.

7 References

[1] Choudhary, A., Chaudhury, S.: 'Video analytics revisited', *IET Comput. Vis.*, 2016, **10**, (4), pp. 237–247

[2] Aggarwal, J.K., Xia, L.: 'Human activity recognition from 3d data: a review', *Pattern Recognit. Lett.*, 2014, **48**, pp. 70–80

[3] Dondi, P., Lombardi, L., Porta, M.: 'Development of gesture-based human–computer interaction applications by fusion of depth and colour video streams', *IET Comput. Vis.*, 2014, **8**, (6), pp. 568–578

[4] Guo, H., Wang, J., Lu, H.: 'Multiple deep features learning for object retrieval in surveillance videos', *IET Comput. Vis.*, 2016, **10**, (4), pp. 268–271

[5] Vrigkas, M., Nikou, C., Kakadiaris, I.A.: 'A review of human activity recognition methods', *Front. Robot. AI*, 2015, **2**, p. 28

[6] Haque, A., Peng, B., Luo, Z., *et al.*: 'Towards viewpoint invariant 3d human pose estimation'. Proc. ECCV, 2016, pp. 160–177

[7] Wang, L., Qiao, Y., Tang, X.: 'Video action detection with relational dynamic-poselets'. Proc. ECCV, 2014, pp. 565–580

[8] Hadfield, S., Lebeda, K., Bowden, R.: 'Hollywood 3d: what are the best 3d features for action recognition?', *Int. J. Comput. Vis.*, 2017, **121**, pp. 95–110

[9] Laptev, I., Marszalek, M., Schmid, C., *et al.*: 'Learning realistic human actions from movies'. Proc. CVPR, 2008, pp. 1–8

[10] Jhuang, H., Gall, J., Zuffi, S., *et al.*: 'Towards understanding action recognition'. Proc. ICCV, 2013, pp. 3192–3199

[11] Sung, J., Ponce, C., Selman, B., *et al.*: 'Unstructured human activity detection from rgbd images'. Proc. ICRA, 2012, pp. 842–849

[12] Guyon, I., Athitsos, V., Jangyodsuk, P., *et al.*: 'The chalearn gesture dataset (cgd 2011)', *Mach. Vis. Appl.*, 2014, **25**, (8), pp. 1929–1951

[13] Escalera, S., Baró, X., Gonzàlez, J., *et al.*: 'Chalearn looking at people challenge 2014: dataset and results'. Proc. ECCV Workshops, 2014, pp. 459–473

[14] Guo, Y., Liu, Y., Oerlemans, A., *et al.*: 'Deep learning for visual understanding: a review', *Neurocomputing*, 2016, **187**, pp. 27–48

[15] Krizhevsky, A., Sutskever, I., Hinton, G.E.: 'ImageNet classification with deep convolutional neural networks'. Proc. NIPS, 2012, pp. 1097–1105

[16] Perronnin, F., Sánchez, J., Mensink, T.: 'Improving the Fisher kernel for large-scale image classification'. Proc. ECCV, 2010, pp. 143–156

[17] Pfister, T., Charles, J., Zisserman, A.: 'Flowing convNets for human pose estimation in videos'. Proc. ICCV, 2015, pp. 1913–1921

[18] Neverova, N., Wolf, C., Taylor, G., *et al.*: 'Moddrop: adaptive multi-modal gesture recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, **38**, (8), pp. 1692–1706

[19] Wang, L., Qiao, Y., Tang, X.: 'Action recognition with trajectory-pooled deep-convolutional descriptors'. Proc. CVPR, 2015, pp. 4305–4314

[20] Seddik, B., Gazzah, S., Essoukri Ben Amara, N.: 'Hands, face and joints for multi-modal human-action temporal segmentation and recognition'. Proc. EUSIPCO, 2015, pp. 1143–1147

[21] Seddik, B., Gazzah, S., Essoukri Ben Amara, N.: 'Modalities combination for Italian sign language extraction and recognition'. Proc. ICIAP, 2015, pp. 710–721

[22] Wan, J., Ruan, Q., Li, W., *et al.*: 'One-shot learning gesture recognition from rgb-d data using bag of features', *J. Mach. Learn. Res.*, 2013, **14**, pp. 2549–2582

[23] Koppula, H.S., Gupta, R., Saxena, A.: 'Learning human activities and object affordances from RGB-D videos', *Int. J. Robot. Res.*, 2013, **32**, (8), pp. 951–970

[24] Camgöz, N.C., Kindiroglu, A.A., Akarun, L.: 'Gesture recognition using template based random forest classifiers'. Proc. ECCV Workshops, 2014, pp. 579–594

[25] Vemulapalli, R., Arrate, F., Chellappa, R.: 'R3DG features: relative 3d geometry-based skeletal representations for human action recognition', *Comput. Vis. Image Underst.*, 2016, **152**, pp. 155–166

[26] Gaglio, S., Re, G.L., Morana, M.: 'Human activity recognition process using 3-d posture data', *IEEE Trans. Hum.-Mach. Syst.*, 2015, **45**, (5), pp. 586–597

[27] Monnier, C., German, S., Ost, A.: 'A multi-scale boosted detector for efficient and robust gesture recognition'. Proc. ECCV Workshops, 2014, pp. 491–502

[28] Shan, J., Akella, S.: '3d human action segmentation and recognition using pose kinetic energy'. Proc. ARSO, 2014, pp. 69–75

[29] Zanfir, M., Leordeanu, M., Sminchisescu, C.: 'The moving pose: an efficient 3d kinematics descriptor for low-latency action recognition and detection'. Proc. ICCV, 2013, pp. 2752–2759

[30] Chang, J.Y.: 'Nonparametric gesture labeling from multi-modal data'. Proc. ECCV Workshops, 2014, pp. 503–517

[31] Faria, D.R., Premebida, C., Nunes, U.: 'A probabilistic approach for human everyday activities recognition using body motion from rgb-d images'. Proc. RO-MAN, 2014, pp. 732–737

[32] Cippitelli, E., Gasparrini, S., Gambi, E., *et al.*: 'A human activity recognition system using skeleton data from rgbd sensors', *Comput. Intell. Neurosci.*, 2016, **2016**, pp. 1–14

[33] Ben Amor, B., Su, J., Srivastava, A.: 'Action recognition using rate-invariant analysis of skeletal shape trajectories', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, **38**, (1), pp. 1–13

[34] Chun, S., Lee, C.: 'Human action recognition using histogram of motion intensity and direction from multiple views', *IET Comput. Vis.*, 2016, **10**, (4), pp. 250–256

[35] Hernández-Vela, A., Bautista, M.Á., Perez-Sala, X., *et al.*: 'Probability-based dynamic time warping and bag-of-visual-and-depth-words for human gesture recognition in RGB-D', *Pattern Recognit. Lett.*, 2013, **50**, pp. 112–121

[36] Wang, H., Kläser, A., Schmid, C., *et al.*: 'Action recognition by dense trajectories'. Proc. CVPR, 2011, pp. 3169–3176

[37] Wang, H., Schmid, C.: 'Action recognition with improved trajectories'. Proc. ICCV, 2013, pp. 3551–3558

[38] Dominio, F., Donadeo, M., Zanuttigh, P.: 'Combining multiple depth-based descriptors for hand gesture recognition', *Pattern Recognit. Lett.*, 2014, **50**, pp. 101–111

[39] Liang, B., Zheng, L.: 'Multi-modal gesture recognition using skeletal joints and motion trail model'. Proc. ECCV Workshops, 2014, pp. 623–638

[40] Oreifej, O., Liu, Z.: 'Hon4d: histogram of oriented 4d normals for activity recognition from depth sequences'. Proc. CVPR, 2013, pp. 716–723

[41] Zhang, C., Tian, Y.: 'Histogram of 3d facets: a depth descriptor for human action and hand gesture recognition', *Comput. Vis. Image Underst.*, 2015, **139**, pp. 29–39

[42] Zhu, Y., Chen, W., Guo, G.: 'Evaluating spatiotemporal interest point features for depth based action recognition', *Image Vis. Comput.*, 2014, **32**, (8), pp. 453–464

[43] Parisi, G.I., Weber, C., Wermter, S.: 'Self-organizing neural integration of pose-motion features for human action recognition', *Front. Neurorobot.*, 2015, **9**, p. 3

[44] Zhou, W., Wang, C., Xiao, B., *et al.*: 'Human action recognition using weighted pooling', *IET Comput. Vis.*, 2014, **8**, (6), pp. 579–587

[45] Peng, X., Wang, L., Wang, X., *et al.*: 'Bag of visual words and fusion methods for action recognition: comprehensive study and good practice', *Comput. Vis. Image Underst.*, 2016, **150**, pp. 109–125

[46] Peng, X., Wang, L., Cai, Z., *et al.*: 'Action and gesture temporal spotting with super vector representation'. Proc. ECCV Workshops, 2014, pp. 518–527

[47] Onofri, L., Soda, P., Iannello, G.: 'Multiple subsequence combination in human action recognition', *IET Comput. Vis.*, 2014, **8**, (1), pp. 26–34

[48] Iosidis, A., Tefas, A., Pitas, I.: 'Discriminant bag of words based representation for human action recognition', *Pattern Recognit. Lett.*, 2014, **49**, pp. 185–192

[49] Pigou, L., Van Den Oord, A., Dieleman, S., *et al.*: 'Beyond temporal pooling: recurrence and temporal convolutions for gesture recognition in video', *Int. J. Comput. Vis.*, 2016, **124**, pp. 1–10
[50] Wu, D., Pigou, L., Kindermanz, P.J., *et al.*: 'Deep dynamic neural networks for multimodal gesture segmentation and recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, **38**, (8), pp. 1583–1597
[51] Selmi, M., El-Yacoubi, M.A., Dorizzi, B.: 'Two-layer discriminative model for human activity recognition', *IET Comput. Vis.*, 2016, **10**, (4), pp. 273–278
[52] Ni, B., Moulin, P., Yan, S.: 'Order-Preserving sparse coding for sequence classification'. Proc. ECCV, 2012, pp. 173–187
[53] Yang, X., Tian, Y.: 'Effective 3d action recognition using eigenjoints', *J. Vis. Commun. Image Represent.*, 2014, **25**, (1), pp. 2–11
[54] Molchanov, P., Yang, X., Gupta, S., *et al.*: 'Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks'. Proc. CVPR, 2016, pp. 4207–4215
[55] Evangelidis, G.D., Singh, G., Horaud, R.: 'Continuous gesture recognition from articulated poses'. Proc. ECCV Workshops, 2014, pp. 595–607
[56] Seddik, B., Maâmatou, H., Gazzah, S., *et al.*: 'Unsupervised facial expressions recognition and avatar reconstruction from kinect'. Proc. SSD, 2013, pp. 1–6