Continuous gesture recognition from articulated poses

Georgios Evangelidis, Gurkirt Singh, Radu Horaud

Continuous gesture recognition from articulated poses *

Georgios D. Evangelidis[1], Gurkirt Singh[2], Radu Horaud[1]

[1]INRIA Grenoble Rhône-Alpes, France
[2]Siemens RTC-ICV, Bangalore, India

Abstract. This paper addresses the problem of continuous gesture recognition from articulated poses. Unlike the common isolated recognition scenario, the gesture boundaries are here unknown, and one has to solve two problems: segmentation and recognition. This is cast into a labeling framework, namely every site (frame) must be assigned a label (gesture ID). The inherent constraint for a piece-wise constant labeling is satisfied by solving a global optimization problem with a smoothness term. For efficiency reasons, we suggest a dynamic programming (DP) solver that seeks the optimal path in a recursive manner. To quantify the consistency between the labels and the observations, we build on a recent method that encodes sequences of articulated poses into Fisher vectors using short skeletal descriptors. A sliding window allows to frame-wise build such Fisher vectors that are then classified by a multi-class SVM, whereby each label is assigned to each frame at some cost. The evaluation in the ChalearnLAP-2014 challenge shows that the method outperforms other participants that rely only on skeleton data. We also show that the proposed method competes with the top-ranking methods when colour and skeleton features are jointly used.

1 Introduction

Gesture and human action recognition from visual information is an active topic with many potential applications in human-computer interaction. The recent release of depth sensors (e.g., Kinect) led to the development of recognition methods based on depth or RGB-D data. Moreover, recent advances on human pose recognition from depth data [23] made the human skeleton extraction possible, so that three information sources are at one's disposal: color, depth, and skeleton. The latter has been proved very effective for human action recognition when used either alone [7, 26] or in conjunction with color/depth features [35, 17].

The majority of gesture (or action) recognition methods consider known boundaries of individual gestures and solve the isolated gesture recognition problem as a single-label assignment (1-of-L) problem, e.g., by invoking a multi-class

* Support from the European Research Council (ERC) through the Advanced Grant VHIA (#340113) is greatly acknowledged.

classfier. However, the continuous case is mostly met in practice, i.e., a video may contain a sequence of gestures in an unknown order and with unknown gesture boundaries. Therefore, one has to solve both the segmentation and the classification problem in order to answer the question: *which* gesture and *when* is performed?

In this paper, we address the above problem within an energy minimization framework, that is, we formulate it as a frame-wise labeling problem under the constraint for a piece-wise constant solution. We build on a recent isolated action recognition method [7] and extend it to the continuous case for gesture recognition. Unlike [7], we use a reduced set of the proposed skeletal descriptors per pose in order to describe the position of the hands with respect to the body. Based on a sliding window, we build frame-wise Fisher vectors that encode the poses of a video segment. Then, a multi-class SVM allows us to assign each label to each frame at some cost, while a binary classifier estimates costs for a "no-gesture" class. All these costs, summarized in a table, are finally exploited by a dynamic programming method that estimates the piece-wise constant labeling which minimizes the total energy. Fig. 1 illustrates the proposed pipeline. We test our method on the ChalearnLAP-2014 dataset [5] and we compare our method with other challenge participants. Note that our primary goal is to first extensively investigate the potential of the skeleton information in a continuous recognition framework, before combining skeletal data with other modalities.

The remainder of the paper is organised as follows. We summarise the related work in Sec. 2 and we present the skeleton-based representation in Sec. 3. Sec. 4 presents the energy minimization framework, while our method is tested on public datasets in Sec. 5. Finally, Sec. 6 concludes this work.

2 Related work

Regardless of the continuous nature of the recognition problem,[1] one initially has to extract features from the available modalities in order to encode the footage. While the potential of color information to provide informative features in a recognition scenario has been extensively studied [30], the release of depth sensors led to the development of depth descriptors, e.g, local occupancy patterns [31, 27] and histogram of spatio-temporal surface normals [18, 37]. More interstingly, the articulated human pose estimation from depth data [23] inspired many researchers to use skeleton as third modality along with RGB-D data, thus building several skeletal descriptors: joint position differences [36, 31], joint angles [3], joint motion characteristics [38], poses of spatio-temporal joint groups [29], relative positions of joint quadruples [7], relative position of joint edges [26], joint angle trajectories [17]. The benefit of combing features from multiple visual sources has been also illustrated [35, 17, 31].

The features are typically translated into a single vector, e.g. Bag-of-Words (BoW) histograms [35] or Fisher vectors [7], while a classifier, e.g. SVM, does

[1] we do not distinguish the problems of human action and gesture recognition, since the latter can be roughly seen as the former when the upper-body part is used

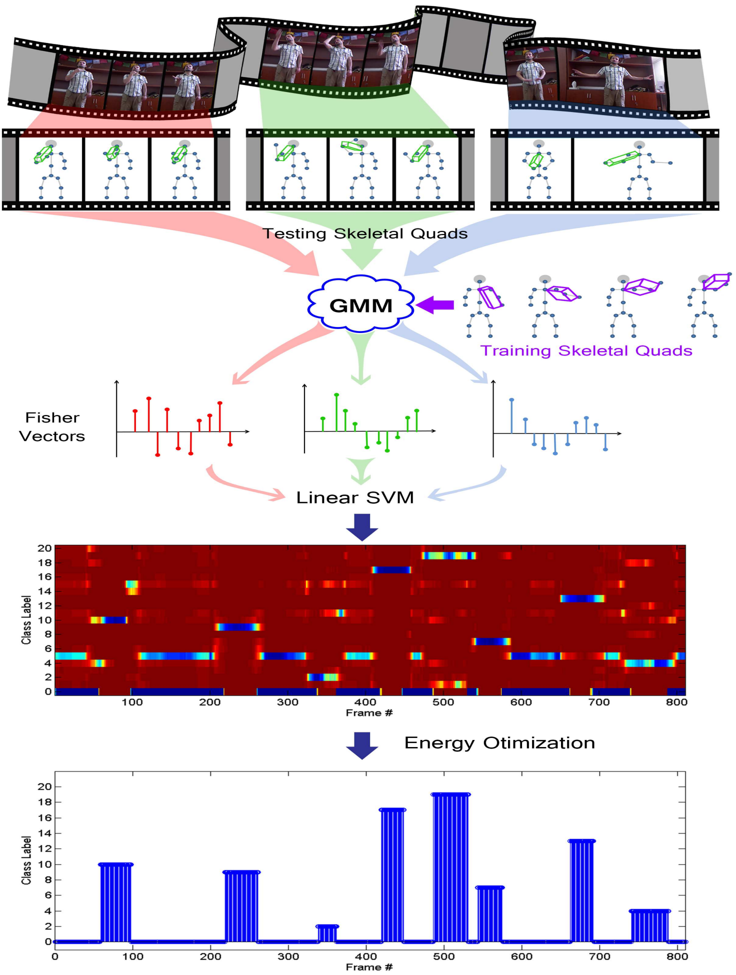


Fig. 1. A Gaussian Mixture Model (GMM), learnt on training data, is supposed to generate skeletal quads. Based on the GMM parameters, the skeletal quads of a gesture segment are encoded into a Fisher vector, and a multi-class SVM assigns a cost per label. A global energy minimizer uses these costs to provide a piece-wise constant labeling.

the 1–of–L label assignment in isolated case. Such a strategy is only locally applicable in a continuous recognition scenario and one has to deal with the temporal nature of the labeling, like in speech recognition. The latter inspired people to develop continuous gesture/action recognition models.

The first continuous sign-language recognizers used hidden Markov models (HMM) [28, 25] for both modeling and recognition. CRF models [24, 32] have been also proposed to avoid HMM's narrow temporal dependencies. Dynamic programming (DP) constitutes a standard framework as well, either in one-pass or two-pass mode, while it can be used in conjunction with either generative or discriminative classifiers [14, 22, 9]. In the DP context, action templates (temporal models that replace HMMs) were recently proposed to be used in a dynamic warping framework [11]. We refer the reader to the latter for a detailed discussion about the pros and cons of all the above models.

ChalearnLAP-2014: Several methods have been proposed in the context of the ChalearnLAP-2014 challenge [5] for the continuous gesture recognition problem. The majority of the methods exploit features from both RGB and depth data [16, 15, 2, 21, 33, 4, 13], while [1] and [19] rely on single modalities, i.e., skeleton and RGB respectively. The silence-based pre-segmentation of sequences is proposed by [4, 13], while the rest of the methods simultaneously solve the segmentation and recognition problems, e.g., with the help of a temporal model [2, 33]. A deep learning framework that employs convolutional neural networks obtained the best results in the challenge [16]. We refer the reader to [5] for a detailed categorization of the above methods in terms of several features.

3 Gesture representation

We rely on [7] in order to encode a set of articulated poses into an informative vector. This approach uses skeletal features, referred to as skeletal quads, to describe a sequence for isolated action recognition. A set of skeletal features is then encoded as a Fisher vector with the help of a trained GMM that explains the generation of any feature set. We briefly discuss these two steps below, that are slightly modified for a gesture recognition scenario.

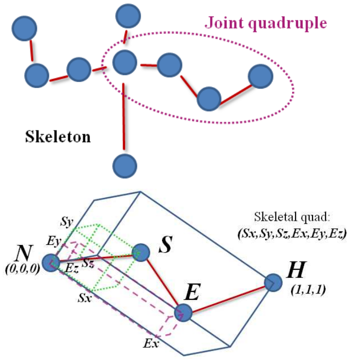
3.1 Skeletal quads

In order to describe a gesture instance, and in particular the hand pose with respect to the body, we use the skeletal quad [7]. This descriptor encodes the geometric relation of joint quadruples thus leading to a low-dimensional descriptor (see Fig 2). This idea originates in [12], while it has been successfully adopted to video synchronisation [8, 6]. In short, if $(\mathbf{x}_1\ \mathbf{x}_2\ \mathbf{x}_3\ \mathbf{x}_4)$ is an ordered set of four joints, i.e., $\mathbf{x}_i \in \mathbb{R}^3$, it is encoded as $\mathbf{q} = [\mathcal{S}(\mathbf{x}_3); \mathcal{S}(\mathbf{x}_4)]$,[2] with $\mathbf{q} \in \mathbb{R}^6$, where

$$\mathcal{S}(\mathbf{x}_i) = s\mathbf{R}[\mathbf{x}_i - \mathbf{x}_1], \quad i = 1 \ldots 4, \tag{1}$$

and s, $\mathbf{R}$ are the scale and rotation respectively, such that $\mathcal{S}(\mathbf{x}_1) = [0, 0, 0]^\top$ and $\mathcal{S}(\mathbf{x}_2) = [1, 1, 1]^\top$. In other words, a similarity normalization is applied to the quadruple, whereby a gesture descriptor is obtained. When the actor looks at the camera (e.g., ChalearnLAP dataset), the above scheme is sufficient. When the camera viewpoint or the body orientation drastically change, any rotation

[2] the notation $[\cdot; \cdot]$ denotes vertical vector concatenation



x_1	x_2	x_3	x_4
HipCenter	Head	HandLeft	HandRight
HipCenter	Head	HandRight	HandLeft
HipCenter	ShoulderCenter	ElbowRight	ElbowLeft
HipCenter	ShoulderCenter	ElbowLeft	ElbowRight
HipCenter	ShoulderCenter	ElbowRight	HandRight
HipCenter	ShoulderCenter	ElbowLeft	HandLeft
Spine	ShoulderLeft	HandLeft	ElbowLeft
Spine	ShoulderRight	HandRight	ElbowRight
ShoulderRight	ShoulderLeft	HandRight	HandLeft
ShoulderRight	ShoulderLeft	HandLeft	HandRight
ShoulderRight	HandRight	WristRight	ElbowRight
ShoulderLeft	HandLeft	WristLeft	ElbowLeft
ShoulderRight	HandRight	ElbowLeft	HandLeft
ShoulderLeft	HandLeft	ElbowRight	HandRight

Fig. 2. (*Left:*) A sketch from [7] that illustrates the coding when ($\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$ $\mathbf{x}_4$) correspond to the joints {Neck, Hand, Shoulder, Elbow}; the quad descriptor is $\mathbf{q} = [S_x, S_y, S_z, E_x, E_y, E_z]^\top$. (*Right:*) The 14 joint quadruples whose quads are locally describe the upper-body pose.

around the local axis $\mathbf{x}_1\mathbf{x}_2$ (axis NH in Fig. 2) must be also normalised, i.e., $\mathbf{R}$ should account for this further rotation as well. This would lead to a fully view-invariant code with less discriminability in principle.

Unlike [7] that enables all possible quads, we choose 14 specific quads based on the upper-body joints, as shown in Fig. 2 (right). Note that this list is not cross-validated based on some data-sets, but it intuitively describes the relation among upper-body joints, by taking into account the body symmetry. For example, the first quad encodes the coordinates of the two hands when the origin is the HipCenter and the Head coincides with the point $[1, 1, 1]^\top$.

3.2 Gesture encoding

Fisher vectors have been proved more discriminative than the the popular BoW representation in a recognition context [20]. We adopt this representation in order to describe a gesture sequence as a set of gesture instances. Note that the low dimension of the skeletal quads compensates for the large inherent dimensionality associated with Fisher vectors.

If statistical independence is assumed, a set of M skeletal quads, $Q = \{\mathbf{q}_i, 1 \leq i \leq M\}$, can be modelled by a K-component Gaussian mixture model (GMM), i.e.,

$$p(Q|\theta) = \prod_{i=1}^{M} \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{q}_i | \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k), \tag{2}$$

where $\theta = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\}_{k=1}^{K}$ is the set of the mixture parameters with mixing coefficients w_k, means $\boldsymbol{\mu}_k \in \mathbb{R}^6$ and *diagonal* covariance matrices, represented here as vectors, i.e., $\boldsymbol{\sigma}_k \in \mathbb{R}^{6\times 1}$. Once the GMM parameters are estimated, e.g., via the standard EM algorithm, any set Q may be described by its Fisher score [10], namely the gradient of the log-probability with respect to the GMM parameters,

$J_\theta^Q = \nabla_\theta \log p(Q|\theta)$. The quadratic form of two such gradient vectors and the inverse information matrix defines the Fisher kernel, which can be written as a linear kernel of the so called Fisher vectors (FV), denoted here as $\mathcal{J}$. The reader is referred to [10] for a detailed analysis.

By considering the gradients with respect to $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k$, the FV consists of the concatenation of two vectors $\mathcal{J}_{\boldsymbol{\mu}_k}^Q$ and $\mathcal{J}_{\boldsymbol{\sigma}_k}^Q$. One can easily show (see [20]) that the $(6k-6+j)$-th element of the above vectors ($1 \leq j \leq 6$, $1 \leq k \leq K$), that is the j-th entry for the k-th block, is given by

$$\mathcal{J}_{\boldsymbol{\mu}_k}^Q(j) = \frac{1}{M\sqrt{\pi_k}} \sum_{i=1}^{M} \gamma_{k,i} \frac{q_i^j - \mu_k^j}{\sigma_k^j},$$
$$\mathcal{J}_{\boldsymbol{\sigma}_k}^Q(j) = \frac{1}{M\sqrt{2\pi_k}} \sum_{i=1}^{M} \gamma_{k,i} \left(\left(\frac{q_i^j - \mu_k^j}{\sigma_k^j} \right)^2 - 1 \right), \quad (3)$$

where $\gamma_{k,i}$ is the posterior probability that $\mathbf{q}_i$ belongs to kth cluster conditioned by Q. The normalization by M is added to avoid dependence on the Q's cardinality. Since quads live in $\mathbb{R}^6$, the Fisher vectors are reasonably long, i.e., of dimension $12K$. Once the FV is computed, we apply a power-normalisation step, i.e., each vector element x is transformed to $sgn(x) * |x|^\alpha$, and the resulting vector is further normalized by its l_2 norm. Note that the power-normalization eliminates the inherent sparseness of the FV; the benefit is discussed in detail in [20].

Unlike [7], we do not use a temporal pyramid since we are going to describe the gestures locally. As a consequence, it is only the following energy minimization scheme that takes into account the temporal nature of gestures towards a continuous labelling.

4 Continuous gesture recognition

We formulate the continuous gesture recognition problem as a labeling problem. In particular, every frame $t \in \{1, ..., T\}$ must be assigned a label $f_t \in \mathcal{L}$ that denotes the gesture ID the frame belongs to. As a consequence, the goal is to find the frame-wise labeling $f = \{f_1, \ldots, f_t, \ldots, f_T\}$ subject to: i) f is consistent with the observations, ii) f is *piece-wise constant*. Note that a frame-wise labeling direclty answers the question "which gesture and when is performed?".

Such a *global* labeling problem can be cast into an energy minimization framework, that is, one estimates the labeling f that minimizes the energy

$$E(f) = E_D(f) + E_S(f), \quad (4)$$

where $E_D(f)$ (the data term) measures the dissimilarity between f and observations and $E_S(f)$ (the smoothness term) penalizes labelings that are not piece-wise constant.

The data-term is typically defined as $E_D(f) = \sum_{t=1}^{T} D_t(f_t)$, where $D_t(f_t)$ measures how appropriate the label f_t is for t-th frame. Here, we use a mutliclass SVM classifier to evaluate this appropriateness. Note that the framework of [34] allows one to compute an empirical probability of assigning a label to an input. We train a multi-class SVM in a one-versus-all manner using FVs of isolated gesture examples. During testing, FVs are built based on a sliding temporal window centered at each frame. Instead of a fully varying window size, we suggest the use of three windows: a narrow, a mid-size and a wide window. By denoting as $p_t(f_t)$ the probability of assigning the label f_t to frame t, the above appropriateness is computed by the following convex combination

$$D_t(f_t) = 1 - \sum_{j=1}^{3} w_j p_t^j(f_t) \tag{5}$$

where j stands for different windows and w_j can balance their contribution.

The smoothness term penalizes the assignment of different labels to successive frames. If we consider that each frame interacts with its immediate neighbors (first-order Markov assumption), the smoothness term can be written as a summation of pairwise potentials

$$E_S(f) = \sum_{t=1}^{T-1} V(f_t, f_{t+1}). \tag{6}$$

Since we seek a piece-wise constant labeling, we define the potential function

$$V(f_t, f_{t+1}) = \begin{cases} 0 & \text{if } f_t = f_{t+1} \\ \beta & \text{otherwise.} \end{cases} \tag{7}$$

To efficiently solve the global optimization problem

$$\min_f E(f), \tag{8}$$

we use a dynamic programming approach, while we leave more sophisticated solutions for a future paper. If $C_t[f_t]$ denotes the minimum labeling cost for the first t frames provided that the t-th frame has label f_t, the following recursive equation (t is increasing) fills in a table C of size $|\mathcal{L}| \times T$:

$$\begin{aligned} C_1[f_1] &= D_1(f_1) \\ C_t[f_t] &= D_t(f_t) + \min_{f_{t-1}}(C_{t-1}[f_{t-1}] + V(f_{t-1}, f_t)) \ . \end{aligned} \tag{9}$$

Once the table is filled, the optimal solution can be obtained by first finding the label of the last frame, i.e., $f_n^* = \arg\min_{f_n} C_n[f_n]$, and then by tracing back in order of decreasing t:

$$f_t^* = \arg\min_{f_i} \big(C_t[f_t]) + V(f_t, f_{t+1}^*)\big). \tag{10}$$

In order to take into account the silent part of a gesture sequence, we train a binary classifier based on silent and non-silent parts of a sufficient number of training examples (see details in Sec. 5). This leads to an extra row in table C that corresponds to a "no-gesture" class. Note that this does not imply a presegmentation, but rather, we let the energy minimizer decide which frames are silent or no.

Note that if one wants to solve the isolated recognition problem, dynamic programming is not required (C reduces to a vector). In such a case, the single label is directly obtained by $f^* = \arg\max_f p(f)$.

5 Experiments

5.1 MSR-Action3D Dataset

Although we focus on the continuous recognition problem, it is important to show that our gesture representation is quite discriminative in the isolated recognition scenario. Our method reduces to [7] for isolated case, being the only difference the few quads that are considered here. We refer the reader to [7] for a performance analysis of the method in isolated case. Here, we just present an updated comparison table based on the widely used MSR-Action3D dataset, by including very recent methods that are missed in [7].

We skip the details of the dataset (see [7]). We just notice that 20 actions are performed by 10 actors and that a cross-subject splitting is considered, i.e. five actors for training and five actors for testing. FVs are built based on 128-component GMM, while they are power-normalized with $\alpha = 0.3$. Instead of all possible joint quadruples (4845) in [7], only 44 meaningful quads are used; they relate hand and leg joints with the body. [3] Note that this leads to a more efficient recognition since the running time of building FVs is reduced by a factor of 100.

Table 1 shows the performance of various methods for this dataset. We do not observe significant loss in performance when using fewer quads. Our method achieves similar performance with the state-of-the-art methods that count on skeleton joints, while it competes with the best-performing methods that use multi-modal features.

5.2 Multi-modal Gesture Dataset 2014

The Multi-modal Gesture dataset was released for the ChalearnLAP-2014 challenge (Track3) [5]. More than $14,000$ gestures drawn from a vocabulary of 20 Italian sign gesture categories were performed by several actors. Each actor, being captured by a Kinect camera, performed a non-fixed number of gestures. Multi-modal data are available, i.e., skeleton (20 joints), RGB and depth image sequences.

The dataset is divided into three parts, development data ($7,754$ manually labeled gestures), validation data ($3,362$ labelled gestures) and evaluation data (2742 gestures). Training and parameter cross-validation was done on development and validation data respectively; testing was done on the evaluation data.

[3] we used the code provided by https://team.inria.fr/perception/research/icpr2014/

Table 1. Recognition accuracy on MSRAction3D datasets.

Modality	Methods	Average Accuracy
Skeleton	EigenJoints [36]	82.33%
	Joint Angles [17]	83.53%
	FV of skeletal quads (**less quads**)	89.07%
	Skeleton Lie group [26]	89.45%
	FV of skeletal quads [7]	89.86%
	Pose Set [29]*	90.22%
	Moving pose [38]	**91.70**%
Skeleton, RGB-D	Joints+Actionlets [31]	88.20%
	HON4D [18]*	88.89%
	Joints + Depth Cuboids [35]	89.30%
	Super Normal vectors [37]	93.09%
	Joints+STIP [39]*	94.30%
	Joint Angles+MaxMin+HOG^2 [17]*	**94.84**%

*different cross-subject splitting

In particular, the GMM and the linear multi-class SVM [4] were learned on the development data while the parameter β of the smoothness term was cross-validated on the validation set; best performance was obtained with $\beta = 3$. As mentioned, 14 quads were invoked (see Fig. 2). We used a GMM with 128 components that led to 1536D Fisher vectors, which were power-normalized with $\alpha = 0.5$. While we observed that less components lead to a lower performance in validation set, we did not test more components to avoid very long FVs. The size of the three sliding windows were 15, 37 and 51, while equal weights ($w_j = 1/3$) were used to compute the cost of assigning each label per frame. There was no significant difference in the performance when changing the window size and the weights. As with the multi-class classifier, the cost of assigning a "no-gesture" label is the average cost obtained from three classifications of short windows, i.e., with 5, 7 and 9 frames. We used 50 videos of the development data for this training, by considering each FV of a window as an example. The performance of the recognition was quantified by the average Jaccard index.

Table 2 summarizes the Jaccard indices of the participants in the challenge. We first sort out the methods that count on skeleton data, and then the methods that combine different modalities. The proposed method and Camgoz *et al.* [1] achieve the best performance when only skeleton joints are used (almost same index). Note that the Jaccard index 0.745 corresponds to the performance of the software version submitted to the challenge. A fixed bug led to a higher index, i.e., 0.768. As previously, the best performing methods use multi-modal features. Neverova *et al.* [16] exploit all the modalities, thus obtaining the highest Jaccard index (0.850). Note that Peng *et al.* [19] performs quite well by using only RGB data.

[4] The scikit-learn Python package was used

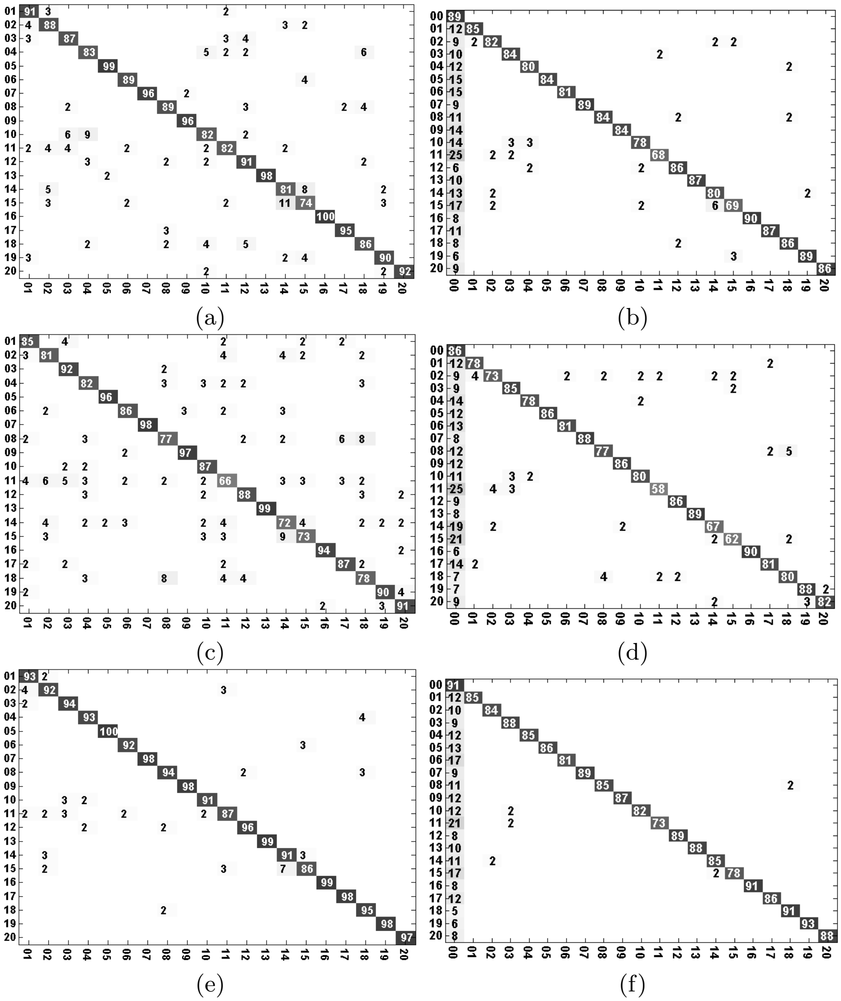
Table 2. Results of the ChalearnLAP-2014 challenge (Track 3) [5]

Team	Modality	Jaccard index
Camgoz *et al.* [1]	Skeleton	0.746
FV of quads (**this work**)	Skeleton	0.745 (0.768*)
Team-13 (Terrier)	Skeleton	0.539
Team-17 (YNL)	Skeleton	0.271
Neverova *et al.* [16]	Skeleton, Depth, RGB	**0.850**
Monnier *et al.* [15]	Depth, RGB	0.834
Ju Yong Chang [2]	Skeleton, RGB	0.827
FV of quads+HOF (**this work**)	Skeleton, RGB,	0.816
Peng *et al.* [19]	RGB	0.792
Pigou *et al.* [21]	Depth, RGB	0.789
Di Wu and Ling Shao [33]	Skeleton, Depth	0.787
Team-9 (TelePoints)	Skeleton, Depth, RGB	0.689
Chen *et al.* [4]	Skeleton, Depth, RGB	0.649
Bin Liang and Lihong Zheng [13]	Skeleton, Depth	0.597
Team-12 (Iva.mm)	Skeleton, Depth, RGB	0.556
Team-14 (Netherlands)	Skeleton, Depth, RGB	0.431
Team-16 (vecsrel)	Skeleton, Depth, RGB	0.408

* A fixed software bug led to a higher index

We also report the performance of our method when skeletal features are combined with color features, i.e., histograms of flows (HOF) [30]. We reduce the dimensionality of HOF feature from 108 to 48 using PCA, and then, we learn another GMM with 64 components which led to 6144D Fisher vector. An early fusion is performed, i.e., the two FVs are concatenated into a single input to be classified. This leads to a higher index (0.816) and makes the proposed method comparable with the top-ranking methods. Note that we did not investigated what is the best color feature to be combined with the quads. It is worth noticing that the performance when using only HOF features is 0.693. Apparently, the combination of skeletal quads with more sophisticated features [30] would lead to a higher index.

Fig. 3 depicts the confusion matrices for the continuous case when each frame is considered as an example (the reader should not confuse these numbers with the Jaccard indices of Table 2). We also show the confusion matrices that correspond to the isolated recognition scenario, i.e. when the gesture boundaries are known. The accuracy of our method in the isolated case is 90%, 86% and 94% when using skeleton, HOF and both features, respectively. The proposed representation is quite discriminative in either case. However, quite similar gestures like "Le vuoi prendere" (id 14) and "Non ce ne piu" (id 15) may be confused without finger joints. The confusion matrix in the continuous case is more sparse since the confusions are concentrated in the column of the no-gesture class (label 00), owing to the frame-wise labeling. It is important to note that the recognition accuracy of this class, i.e., 89%, 86% and 91%, is based on the final global



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 3. Confusion matrices of our recognition method for Multi-modal gesture dataset when using skeletal quads (a,b), HOF (c,d), skeletal quads + HOF (e,f) in isolated (left) and continuous recognition scenario (right).

labeling, as with any other class. Apparently, the use of color features improves the discrimination and the matrices tend to be more diagonal. Note that the percentages do not sum up to 100 since we keep the integer part of the numbers.

6 Conclusions

We dealt with the continuous gesture recognition problem from series of articulated poses. The problem was cast into a labeling framework and it was solved as a global energy minimization problem. We proposed a dynamic programming solver that exploits the outputs of multiple SVMs and provides a piece-wise constant labeling. We mainly tested our method on the Moltabagno Gesture dataset, released for the ChalearnLAP-2014 challenge purposes. Despite the use of the skeleton information only, the proposed method achieved high recognition scores, while its performance was boosted when extra modalities, e.g. colour data, were used. Future work consists of extending Fisher vectors to deal with the temporal nature of a pose sequence, and of investigating the optimal fusion with other modalities.

References

1. Camgoz, N.C., Kindiroglu, A.A., Akarun, L.: Gesture recognition using template based random forest classifiers. In: ECCV Workshops (2014)
2. Chang, J.Y.: Nonparametric gesture labeling from multi-modal data. In: ECCV Workshops (2014)
3. Chaudhry, R., Ofli, F., Kurillo, G., Bajcsy, R., Vidal, R.: Bio-inspired dynamic 3d discriminative skeletal features for human action recognition. In: CVPR Workshops (CVPRW) (2013)
4. Chen, G., Clarke, D., Weikersdorfer, D., Giuliani, M., Gaschler, A., Knoll, A.: Multi-modality gesture detection and recognition with un-supervision, randomization and discrimination. In: ECCV Workshops (2014)
5. Escalera, S., Bar, X., Gonzlez, J., Bautista, M.A., Madadi, M., Reyes, M., Ponce, V., Escalante, H.J., Shotton, J., Guyon, I.: Chalearn looking at people challenge 2014: Dataset and results. In: ECCV Workshops (2014)
6. Evangelidis, G., Bauckhage, C.: Efficient subframe video alignment using short descriptors. IEEE T PAMI 35, 2371–2386 (2013)
7. Evangelidis, G., Singh, G., Horaud, R., et al.: Skeletal quads: Human action recognition using joint quadruples. In: ICPR (2014)
8. Evangelidis, G.D., Bauckhage, C.: Efficient and robust alignment of unsynchronized video sequences. In: DAGM (2011)
9. Hoai, M., Lan, Z.Z., De la Torre, F.: Joint segmentation and classification of human actions in video. In: CVPR (2011)
10. Jaakola, T., Haussler, D.: Exploiting generative models in discriminative classifiers. In: NIPS (1999)
11. Kulkarni, K., Evangelidis, G., Cech, J., Horaud, R.: Continuous action recognition based on sequence alignment. IJCV (2014), preprint
12. Lang, D., Hogg, D.W., Mierle, K., Blanton, M., Roweis, S.: Astrometry.net: Blind astrometric calibration of arbitrary astronomical images. The astronomical journal 137, 1782–2800 (2010)
13. Liang, B., Zheng, L.: Multi-modal gesture recognition using skeletal joints and motion trail model. In: ECCV Workshops (2014)
14. Lv, F., Nevatia, R.: Recognition and segmentation of 3-d human action using HMM and multi-class AdaBoost. In: ECCV (2006)
15. Monnier, C., German, S., Ost, A.: A multi-scale boosted detector for efficient and robust gesture recognition. In: ECCV Workshops (2014)
16. Neverova, N., Wolf, C., Taylor, G.W., Nebout, F.: Multi-scale deep learning for gesture detection and localization. In: ECCV Workshops (2014)

17. Ohn-Bar, E., Trivedi, M.M.: Joint angles similiarities and hog^2 for action recognition. In: Computer Vision and Pattern Recognition Workshops (CVPRW) (2013)
18. Oreifej, O., Liu, Z.: Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In: CVPR (2013)
19. Peng, X., Wang, L., Cai, Z.: Action and gesture temporal spotting with super vector representation. In: ECCV Workshops (2014)
20. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: ECCV (2010)
21. Pigou, L., Dieleman, S., Kindermans, P.J., Schrauwen, B.: Sign language recognition using convolutional neural networks. In: ECCV Workshops (2014)
22. Shi, Q., Cheng, L., Wang, L., Smola, A.: Human action segmentation and recognition using discriminative semi-markov models. IJCV 93(1), 22–32 (2011)
23. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR (2011)
24. Sminchisescu, C., Kanaujia, A., Metaxas, D.: Conditional models for contextual human motion recognition. CVIU 104(2), 210–220 (2006)
25. Starner, T., Weaver, J., Pentland, A.: Real-time american sign language recognition using desk and wearable computer based video. IEEE T PAMI 20(12), 1371–1375 (1998)
26. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: CVPR (2014)
27. Vieira, A.W., Nascimento, E.R., Oliveira, G.L., Liu, Z., Campos, M.F.: On the improvement of human action recognition from depth map sequences using spacetime occupancy patterns. Pattern Recognition Letters 36, 221–227 (2014)
28. Vogler, C., Metaxas, D.: ASL recognition based on a coupling between HMMs and 3D motion analysis. In: ICCV (1998)
29. Wang, C., Wang, Y., Yuille, A.L.: An approach to pose-based action recognition. In: CVPR (2013)
30. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV (2013)
31. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: CVPR (2012)
32. Wang, S.B., Quattoni, A., Morency, L., Demirdjian, D., Darrell, T.: Hidden conditional random fields for gesture recognition. In: CVPR (2006)
33. Wu, D., Shao, L.: Deep dynamic neural networks for gesture segmentation and recognition. In: ECCV Workshops (2014)
34. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. The Journal of Machine Learning Research 5, 975–1005 (2004)
35. Xia, L., Aggarwal, J.: Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In: CVPR (2013)
36. Yang, X., Tian, Y.: Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: CVPR Workshops (CVPRW) (2012)
37. Yang, X., Tian, Y.: Super normal vector for activity recognition using depth sequences. In: CVPR (2014)
38. Zanfir, M., Leordeanu, M., Sminchisescu, C.: The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In: ICCV. pp. 2752–2759 (2013)
39. Zhu, Y., Chen, W., Guo, G.: Fusing spatiotemporal features and joints for 3d action recognition. In: CVPR Workshops (CVPRW). pp. 486–491 (2013)