Multi-scale deep learning for gesture detection and localization

[1,2]Natalia Neverova [1,2]Christian Wolf [3]Graham W. Taylor [4]Florian Nebout

[1]Université de Lyon, CNRS, France firstname.surname@liris.cnrs.fr
[2]INSA-Lyon, LIRIS, UMR5205, F-69621
[3]University of Guelph, Canada gwtaylor@uoguelph.ca
[4]Awabot, Lyon, France florian.nebout@awabot.com

Abstract. We present a method for gesture detection and localization based on multi-scale and multi-modal deep learning. Each visual modality captures spatial information at a particular spatial scale (such as motion of the upper body or a hand), and the whole system operates at two temporal scales. Key to our technique is a training strategy which exploits i) careful initialization of individual modalities; and ii) gradual fusion of modalities from strongest to weakest cross-modality structure. We present experiments on the *ChaLearn 2014 Looking at People Challenge* gesture recognition track, in which we placed first out of 17 teams.

Keywords: gesture recognition, multi-modal systems, deep learning

1 Introduction

Visual gesture recognition is one of the central problems in the rapidly growing fields of human-computer and human-robot interaction. Effective gesture detection and classification is challenging due to several factors: cultural and individual differences in tempos and styles of articulation, variable observation conditions, the small size of fingers in images taken in typical acquisition conditions, noise in camera channels, infinitely many kinds of out-of-vocabulary motion, and real-time performance constraints.

Recently, the field of deep learning has matured and made a tremendous impact in computer vision, demonstrating previously unattainable performance on the tasks of object detection, localization [1, 2], recognition [3] and image segmentation [4, 5]. Convolutional neural networks (ConvNets) [6] have excelled on several scientific competitions such as ILSVRC [3], Emotion Recognition in the Wild (EmotiW 2013) [7], Kaggle Dogs vs. Cats [2] and Galaxy Zoo. Taigman et al. [8] recently claimed to have reached human-level performance using ConvNets for face recognition. On the other hand, extending these models to problems involving the understanding of *video* content is still in its infancy, this idea having been explored in a only a small number of recent works [9–11]. It can be partially explained by lack of sufficiently large datasets and the high cost of data labeling in many practical areas, as well as increased modeling complexity brought about the additional temporal dimension and the interdependencies it implies.

The first gesture-oriented dataset containing a sufficient amount of training samples for deep learning methods was proposed for the the *ChaLearn 2013 Challenge on Multi-modal gesture recognition*. The deep learning method we describe here placed first in the 2014 version of this competition [12].

A core aspect of our approach is employing a multi-modal convolutional neural network for classification of so-called dynamic poses of varying durations (i.e. temporal scales). The best single scale configuration corresponding to a certain formulation of the dynamic pose alone places first (see Section 5 for more details), while introducing parallel multi-scale paths leads to an additional gain in performance. Finally, we find it interesting to provide a comparison of the proposed approach with a baseline model employing a popular ensemble method. The performance of a hybrid solution, leading to another small gain, is reported for a reference.

Data modalities integrated by our algorithm include intensity and depth video, as well as articulated pose information extracted from depth maps. We make use of different data channels to decompose each gesture at multiple scales not only temporally, but also spatially, to provide context for upper-body body motion and more fine-grained hand/finger articulation. We pay special attention to developing an effective learning algorithm since learning large-scale multi-modal networks like the one we train on a limited labeled dataset is a formidable challenge.

Our classification model outputs prediction updates in real-time in frame-wise manner. Nevertheless, since temporal integration is involved, the classification model suffers from a certain degree of inertia. Furthermore, due to high similarity between gesture classes on pre-stroke and post-stroke phases, frame-wise classification at that time is often uncertain and therefore erroneous. To compensate for these negative effects, an additional module is introduced for filtering, denoising and more accurate gesture localization.

The major contributions of the present work are the following: (i) we develop a deep learning-based multi-modal and multi-scale framework for gesture detection, localization and recognition; and (ii) propose a progressive learning procedure enabling our method to scale to a higher number of data modalities.

2 Related work

Gesture recognition — Traditional approaches to action and distant gesture recognition from video typically include sparse or dense extraction of spatial or spatio-temporal engineered descriptors followed by classification [13–18].

Near-range applications may require more accurate reconstruction of hand shapes. In this case, fitting a 3D hand model, as well as appearance-based algorithms provide more appropriate solutions. A group of recent works is dedicated to inferring the hand pose through pixel-wise hand segmentation and estimating the positions of hand joints in a bottom-up fashion [19–22]. In parallel, tracking-based approaches are advancing quickly [23, 24]. Finally, graphical models, exploring spatial relationships between body and hand parts, have recently attracted close attention from the vision community [25, 26].

Multi-modal aspects are of relevance in this domain. In [27], a combination of skeletal features and local occupancy patterns (LOP) are calculated from depth maps to describe hand joints. In [28], the skeletal information is integrated in two ways for extracting HoG features from RGB and depth images: either from global bounding boxes containing a whole body or from regions containing an arm, a torso and a head. Similarly, [29] fuse skeletal information with HoG features extracted from the RGB channel, while [30] propose a combination of a covariance descriptor representing skeletal joint data with spatio-temporal interest points extracted from the RGB modality augmented with audio features.

Representation learning — Various fundamental architectures have been proposed in the context of motion analysis for *learning* (as opposed to handcrafting) representations directly from data, either in a supervised or unsupervised way. Independent subspace analysis (ISA) [31] as well as autoencoders [32, 9] are examples of efficient unsupervised methods for learning hierarchies of invariant spatio-temporal features. Space-time deep belief networks [33] produce high-level representations of video sequences using convolutional RBMs.

Vanilla supervised convolutional networks have also been explored in this context. A method proposed in [34] is based on low level preprocessing of the video input and employs a 3D convolutional network for learning of mid-level spatio-temporal representations and classification. Recently, Karpathy et al. [10] have proposed a convolutional architecture for general purpose large-scale video classification operating at two spatial resolutions (a fovea stream and a context stream).

A number of deep architectures have recently been proposed specifically for multi-modal data. Ngiam et al. [35] employ sparse RBMs and bimodal deep antoencoders for learning cross-modality correlations in the context of audio-visual speech classification of isolated letters and digits. Srivastava et al. [36] use a multi-modal deep Boltzmann machine in a generative fashion to tackle the problem of integrating image data and text annotations. Kahou et al. [7] won the 2013 Emotion Recognition in the Wild Challenge by building two convolutional architectures on several modalities, such as facial expressions from video frames, audio signal, scene context and features extracted around mouth regions. Finally, in [37] the authors propose a multi-modal convolutional network for gesture detection and classification from a combination of depth, skeletons and audio.

3 Gesture classification

On a dataset such as *ChaLearn 2014*, we face several key challenges: learning representations at multiple spatial and temporal scales, integrating the various modalities, and training a complex model when the number of labeled examples is not at *web-scale* like static image datasets (e.g. [3]). We start by describing how the first two challenges are overcome at an architectural level. Our training strategy to overcome the last challenge is described in Sec. 3.4.

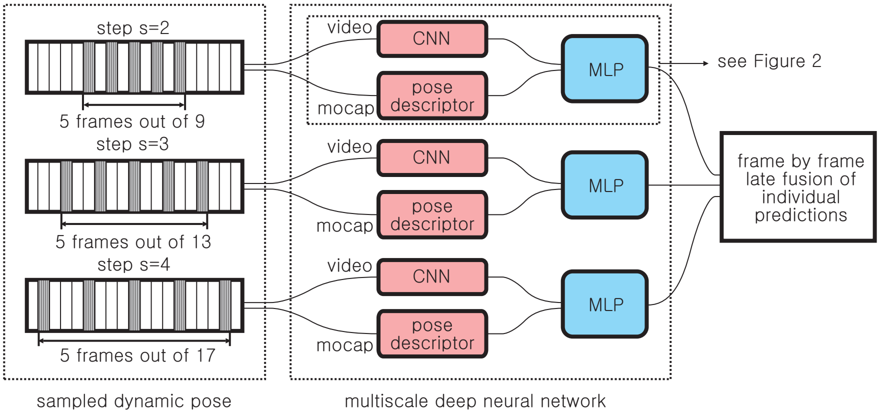

Fig. 1. The deep convolutional multi-modal architecture operating at 3 temporal scales corresponding to dynamic poses of 3 different durations.

Our proposed multi-scale deep neural model consists of a combination of single-scale paths connected in a parallel way (see Fig. 1). Each path independently learns a representation and performs gesture classification at its own temporal scale given input from RGB-D video and articulated pose descriptors. Predictions from all paths are then aggregated through additive late fusion. This strategy allows us to first extract the most salient (in a discriminative sense) motions at a fine temporal resolution and, at the same time, consider them in the context of global gesture structure, smoothing and compensating for per-block errors typical for a given gesture class.

To differentiate among temporal scales, a notion of dynamic pose is introduced. By *dynamic pose* we mean a sequence of video frames, synchronized across modalities, sampled at a given temporal step s and concatenated to form a spatio-temporal 3d volume. Varying the value of s allows the model to leverage multiple temporal scales for prediction, thereby accommodating differences in tempos and styles of articulation of different users. Our model is therefore different from the one proposed in [4], where by "multi-scale" the authors imply a multi-resolution spatial pyramid rather than a fusion of temporal sampling strategies. Regardless of the step s, we use the same number of frames (5) at each scale. Fig. 1 shows the three such paths used in this work (with $s = 2 \ldots 4$). At each scale and for each dynamic pose, the classifier outputs a per-class score.

All available modalities, such as depth, gray scale video, and articulated pose, contribute to the network's prediction. Global appearance of each gesture instance is captured by the skeleton descriptor, while video streams convey additional information about hand shapes and their dynamics which are crucial for discriminating between gesture classes performed in similar body poses.

Due to the high dimensionality of the data and the non-linear nature of cross-modality structure, an immediate concatenation of raw skeleton and video signals is sub-optimal. However, initial discriminative learning of individual data representations from each isolated channel followed by fusion has proven to be efficient in similar tasks [35]. Therefore, in our approach, discriminative data representations are first learned within each separate channel, followed by joint fine tuning and fusion by a meta-classifier (independently at each scale, for more details see Sec. 3.4). A shared set of hidden layers is employed at different levels for, first, fusing of "similar by nature" gray scale and depth video streams and, second, combining the obtained joint video representation with the transformed articulated pose descriptor.

3.1 Articulated pose

We formulate a pose descriptor, consisting of 7 logical subsets, and allow the classifier to perform online feature selection. The descriptor is calculated based on 11 upper body joints, relevant to the task, whose raw, i.e. pre-normalization, positions in a 3D coordinate system associated with the depth sensor are denoted as $\mathbf{p}^{(i)}_{\text{raw}} = \{x^{(i)}, y^{(i)}, z^{(i)}\}$, $i = 0...10$ ($i = 0$ corresponds to the *HipCenter* joint).

Following the procedure proposed in [38], we first calculate normalized joint positions, as well as their velocities and accelerations, and then augment the descriptor with a set of characteristic angles and pairwise distances.

Joint positions. The skeleton is represented as a tree structure with the *HipCenter* joint playing the role of a root node. Its coordinates are subtracted from the rest of the vectors $\mathbf{p}_{\text{raw}}$ to eliminate the influence of position of the body in space. To compensate for differences in body sizes, proportions and shapes, we start from the top of the tree and iteratively normalize each skeleton segment to a corresponding average "bone" length estimated from all available training data. It is done in the way that absolute joint positions are corrected while corresponding orientations remain unchanged:

$$\mathbf{p}^{(i)}(t) = \mathbf{p}^{(i-1)}_{\text{raw}}(t) + \frac{\mathbf{p}^{(i)}_{\text{raw}}(t) - \mathbf{p}^{(i-1)}_{\text{raw}}(t)}{||\mathbf{p}^{(i)}_{\text{raw}}(t) - \mathbf{p}^{(i-1)}_{\text{raw}}(t)||} b^{(i-1,i)} - \mathbf{p}^{(0)}_{\text{raw}}(t), \tag{1}$$

where $\mathbf{p}^{(i)}_{\text{raw}}$ is a current joint, $\mathbf{p}^{(i-1)}_{\text{raw}}$ is its direct predecessor in the tree, $b^{(i-1,i)}$, $i = 1 \ldots 10$ is a set of estimated average lengths of "bones" and $\mathbf{p}$ are corresponding normalized joints. Once the normalized joint positions are obtained, we perform Gaussian smoothing along the temporal dimension ($\sigma = 1$, filter size 5×1) to decrease the influence of skeleton jitter.

Joint velocities are calculated as first derivatives of normalized joint positions: $\delta\mathbf{p}^{(i)}(t) \approx \mathbf{p}^{(i)}(t+1) - \mathbf{p}^{(i)}(t-1)$.

Joint accelerations correspond to the second derivatives of the same positions: $\delta^2\mathbf{p}^{(i)}(t) \approx \mathbf{p}^{(i)}(t+2) + \mathbf{p}^{(i)}(t-2) - 2\mathbf{p}^{(i)}(t)$.

Inclination angles are formed by all triples of anatomically connected joints (i, j, k), plus two "virtual" angles (Right,Left)*Elbow*-(Right,Left)*Hand*-*HipCenter*:

$$\alpha^{(i,j,k)} = \arccos \frac{(\mathbf{p}^{(k)} - \mathbf{p}^{(j)})(\mathbf{p}^{(i)} - \mathbf{p}^{(j)})}{||\mathbf{p}^{(k)} - \mathbf{p}^{(j)}|| \cdot ||\mathbf{p}^{(i)} - \mathbf{p}^{(j)}||} \tag{2}$$

Azimuth angles β provide additional information about the pose in the coordinate space associated with the body. We apply PCA on the positions of 6 torso joints (*HipCenter*, *HipLeft*, *HipRight*, *ShoulderCenter*, *ShoulderLeft*, *ShoulderRight*) to obtain 3 vectors forming the basis: $\{\mathbf{u}_x, \mathbf{u}_y, \mathbf{u}_z\}$, where $\mathbf{u}_x$ is approximately parallel to the shoulder line, $\mathbf{u}_y$ is aligned with the spine and $\mathbf{u}_z$ is perpendicular to the torso.

Then for each pair of connected bones, β are angles between projections of the second bone ($\mathbf{v}_2$) and the vector $\mathbf{u}_x$ ($\mathbf{v}_1$) on the plane perpendicular to the orientation of the first bone. As in the previous case of inclination angles, we also include two virtual “bones” (Right,Left)*Hand-HipCenter*.

$$\begin{aligned} \mathbf{v}_1 &= \mathbf{u}_x - (\mathbf{p}^{(j)} - \mathbf{p}^{(i)}) \frac{\mathbf{u}_x \cdot (\mathbf{p}^{(j)} - \mathbf{p}^{(i)})}{||\mathbf{p}^{(j)} - \mathbf{p}^{(i)}||^2} \\ \mathbf{v}_2 &= (\mathbf{p}^{(k)} - \mathbf{p}^{(j)}) - (\mathbf{p}^{(j)} - \mathbf{p}^{(i)}) \frac{(\mathbf{p}^{(k)} - \mathbf{p}^{(j)}) \cdot (\mathbf{p}^{(j)} - \mathbf{p}^{(i)})}{||\mathbf{p}^{(j)} - \mathbf{p}^{(i)}||^2} \\ \beta^{(i,j,k)} &= \arccos \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{||\mathbf{v}_1|| ||\mathbf{v}_1||} \end{aligned} \tag{3}$$

Bending angles γ are a set of angles between a basis vector $\mathbf{u}_z$, perpendicular to the torso, and normalized joint positions:

$$\gamma^{(i)} = \arccos \frac{\mathbf{u}_z \cdot \mathbf{p}^{(i)}}{||\mathbf{p}^{(i)}||} \tag{4}$$

Pairwise distances. Finally, we calculate pairwise distances between all normalized joint positions: $\rho^{(i,j)} = ||\mathbf{p}_n^{(i)} - \mathbf{p}_n^{(j)}||$.

Combined together, this produces a 183-dimensional pose descriptor for each video frame: $\mathbf{D} = [\mathbf{p}, \delta\mathbf{p}, \delta^2\mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\rho}]^T$. Finally, each feature is normalized to zero mean and unit variance.

A set of consequent 5 frame descriptors sampled at a given step s are concatenated to form a 915-dimensional dynamic pose descriptor which is further used for gesture classification. The two subsets of features involving derivatives contain dynamic information and for dense sampling may be partially redundant as several occurrences of same frames are stacked when a dynamic pose descriptor is formulated. Although theoretically unnecessary, this is beneficial in the context of a limited amount of training data.

3.2 Depth and intensity video: convolutional learning

In our approach, two video streams serve as a source of information about hand pose and finger articulation. Bounding boxes containing images of hands are cropped around positions of the *RightHand* and *LeftHand* joints. Within each set of frames forming a dynamic pose, hand position is stabilized by minimizing inter-frame square-root distances calculated as a sum over all pixels, and corresponding frames are concatenated to form a single spatio-temporal volume. The color stream is converted to gray scale, and both depth and intensity frames are

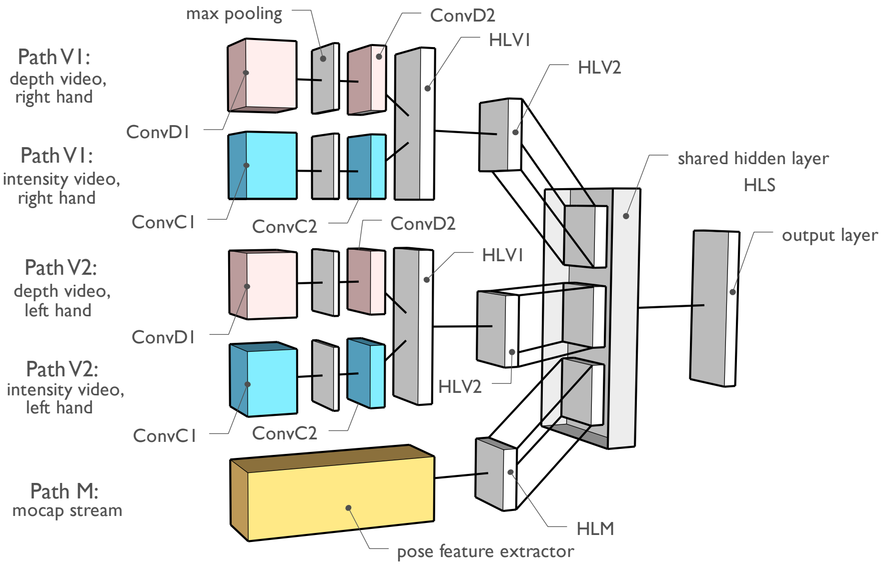

Fig. 2. Single-scale deep architecture. Individual classifiers are pre-trained for each data modality (paths V1, V2, M) and then fused using a 2-layer shared fully connected network initialized in a specific way (see Sec. 3.4). The first layers perform 3D convolutions followed by 3D max pooling shrinking the temporal dimension. The second layers are exclusively spatial. Weights are shared across V1 and V2 paths.

normalized to zero mean and unit variance. Left hand videos are flipped about the vertical axis and combined with right hand instances in a single training set.

During pre-training, video pathways are adapted to produce predictions for each hand, rather than for the whole gesture. Therefore, we introduce an additional step to eliminate possible noise associated with switching from one active hand to another. For one-handed gesture classes, we estimate the motion trajectory length of each hand using the respective joints provided by the skeleton stream (summing lengths of hand trajectories projected to the x and y axes):

$$\varDelta = \sum_{t=2}^{5} (|x(t) - x(t-1)| + |y(t) - y(t-1)|), \quad (5)$$

where $x(t)$ is the x-coordinate of a hand joint (either left or right) and $y(t)$ is its y-coordinate. Finally, the hand with a greater value of $\varDelta$ gets assigned to the label class, while the second one is assigned the zero-class label.

For each channel and each hand, we perform 2-stage convolutional learning of data representations independently (first in 3D, then in 2D space, see Fig. 2) and then fuse the two streams with a set of fully connected hidden layers. Parameters of the convolutional and fully-connected layers at this step are shared between the right hand and left hand pathways. Our experiments have demonstrated that relatively early fusion of depth and intensity features leads to a significant increase in performance, even though the quality of predictions obtained from each channel alone is unsatisfactory.

3.3 Fusion

Once individual single-scale predictions are obtained, we employ a simple voting strategy for fusion with a single weight per model. We note here that introducing additional per-class per-model weights and training meta-classifiers (such as an MLP) on this step quickly leads to overfitting.

At each given frame t per-class network outputs o_k are obtained via per-frame aggregation and temporal filtering of predictions at each scale with corresponding weights μ_s defined empirically through cross-validation on a validation set:

$$o_k(t) = \sum_{s=2}^{4} \mu_s \sum_{j=-4s}^{0} o_{s,k}(t+j), \tag{6}$$

where $o_{s,k}(t+j)$ is the score of class k obtained for a spatio-temporal block sampled starting from the frame $t+j$ at step s. Finally, the frame is assigned the class label $l(t)$ having the maximum score: $l(t) = \arg\min_k o_k(t)$.

3.4 Training

With an increasing number of data modalities, efficient training of large-scale deep architectures becomes one of the most practically important issues in domains such as gesture understanding. Due to an exploding number of parameters, direct modeling of joint data distributions from all available data sources is not always possible. The problem becomes even more crucial if we aim on simultaneous data fusion and modeling temporal sequences. In this work we used several strategies, such as pre-training of individual classifiers on separate channels and iterative fusion process of all modalities.

Recall Fig. 2 illustrating a one scale deep multi-modal convolutional network. Initially it has 5 separate pathways: depth and intensity video channels for right (V1) and left hands (V2), and a mocap stream (M).

We start with transforming of each data input to the form which is discriminative for the given classification task by passing the data through a modality-specific 3-step convolutional neural network (ConvD1-ConvD2 in the case of depth data and ConvC1-ConvC2 in the case of intensity video) or by manual feature extraction (in the case of mocap data, as it was described in Sec. 3.1).

From our observations, inter-modality fusion is effective at early stages if both channels have the same nature and convey overlapping information. On the other hand, mixing modalities which are weekly correlated, is rarely beneficial until the final stage. Accordingly, in our architecture two video channels corresponding to the same hand are fused immediately after feature extraction (hidden layers HLV1 and HLV2), while exploring cross-modality correlations of complementary skeleton motion and hand articulation is postponed by two layers (the fusion is performed only at a shared layer HLS).

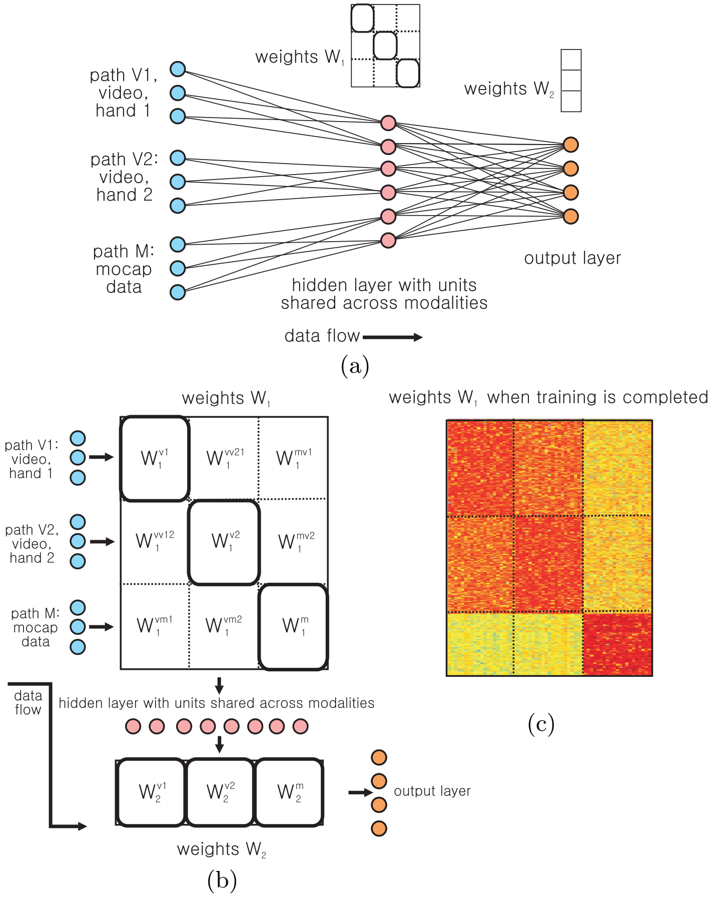

Fig. 3. (a) Architecture of shared hidden and output layers. Output hidden units on each path are first connected to a subset of neurons of the shared layer; (b) Structure of parameters of shared hidden and output layers (corresponds to the architecture above); (c) Energy structure of weights W1 after training. Diagonal blocks are dominated by individual modalities, off-diagonal elements reflect cross-modality correlations.

Furthermore, proper initialization of the shared layer HLS before data fusion is important. Direct fully-connected wiring of pre-trained paths to the shared layer with randomly initialized weights leads to quick degradation of pre-trained connections and as a result, our experience suggests that the joint representation performs worse than a linear combination of the predictions of individual classifiers. This may be related to the fact that the amount of data at our disposal is still not sufficient for straightforward training of such large-scale architectures.

To address this issue, one possible strategy would be to train a classifier on data that is arranged in a specific meaningful order, starting from clean samples that are easy to classify and proceeding to the most complex ones, allowing the network to learn more and more sophisticated concepts (as it is done, for example, in curriculum learning [39]). This approach has shown to yield better generalization in less time. We employ a similar but alternative strategy, changing the network itself in an iterative way, evolving from a weak prediction model to increasingly more complex prediction models. The network is divided into meaningful parts that are pre-trained separately and then combined. We begin training by presenting modality-specific parts of the network with samples where only one modality is present. In this way, we pre-train initial sets of modality-specific layers that extract features from each data channel and create more meaningful and compact data representations.

Once pre-training is completed, we proceed with integrating all channels, one by one, in an iterative manner (see Fig. 3). We choose the order of modalities in a specific way to first combine the data where the strongest cross-modality structure is expected. This permits the model to gradually and effectively learn a joint distribution, focusing representational power on where it is most effective, while keeping the input compact and the number of parameters relatively small. In the task of multi-modal gesture recognition, the video stream and articulated pose alone convey sufficient information about the gesture, i.e. recognition can be performed reasonably well from each channel independently. However, data in the two depth channels, representing the articulation of each of the two hands, is complementary and can improve accuracy.

To ensure that the joint model is meaningful, both the shared representation layer and output layer are first configured to produce an optimal weighted sum of individual modalities. The network parameters are further optimized starting from this initialization. We start the fusion procedure by integrating two highly dependent video channels (V1 and V2) with shared parameters, then add the third visual modality (articulated pose, path M) (see Fig. 3).

4 Gesture localization

With increasing duration of a dynamic pose, recognition rates of the classifier increase at a cost of loss in precision in gesture localization. Using wider sliding windows leads to noisy predictions at pre-stroke and post-stroke phases, in some cases overlapping several gesture instances at once. On the other hand, too short dynamic poses are not discriminative either as most gesture classes at their initial and final stages have a similar appearance (e.g. raising or lowering hands).

To address this issue, we introduce an additional binary classifier to distinguish resting moments from periods of activity. Trained on dynamic poses at the finest temporal resolution $s = 1$, this classifier is able to precisely localize starting and ending points of each gesture.

The module is implemented based on the same articulated pose descriptor input to the MLP. All training frames labeled with some gesture class are used as positive examples, while a set of frames right before and after such gesture are considered as negatives. This strategy allows us to assign each frame with a label "motion" or "no motion" with accuracy of 98%.

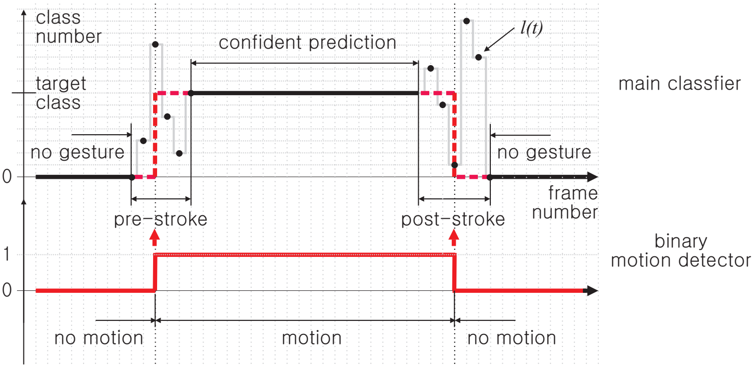

Fig. 4. Gesture localization. Top: output predictions of the main classifier; Bottom: output of the binary motion detector. Noise at pre-stroke and post-stroke phases in the first case is due to high similarity between gesture classes at these time periods and temporal inertia of the classifier.

To combine the classification and localization modules, frame-wise gesture class predictions are first obtained as described in Section 3.3. Output predictions at the beginning and at the end of each gesture are typically noisy (see the top curve at Fig. 4). For each spotted gesture, its boundaries are extended or shrunk towards the closest switching point produced by the binary classifier.

5 Experiments

The *Chalearn 2014 Looking at People Challenge (track 3)* dataset consists of 13,858 instances of Italian conversational gestures performed by different people and recorded with a consumer RGB-D sensor. It includes color, depth video and articulated pose streams. The gestures are drawn from a large vocabulary, from which 20 categories are identified to detect and recognize [12]. Training data is accompanied by a ground truth label for each gesture, as well as information about its starting and ending points. The corpus is split into development, validation and test set, where the test data has been released after code submission.

5.1 Experimental setup

The hyperparameters used for the convolutional nets are provided in Table 1. They were constant across temporal scales. Gesture localization was performed with an MLP with 300 hidden units. All hidden units of both modules (framewise classification and following localization) had rectified linear (ReLU) activations. Hyperparameters were optimized on the validation data. Early stopping based on a validation set was employed to prevent the models from overfitting. Optimal fusion weights for the different temporal scales were found to be: $\mu_{s=2} = 0.26$, $\mu_{s=3} = 1.02$, $\mu_{s=4} = 2.20$ and the weight of the baseline model (see Section 5.2) was set to $\mu_{\text{ERT}} = 1$.

Table 1. Hyperparameters chosen for the deep learning models.

Layer	Parameters	Layer	Parameters
Convolutional layer ConvD1	25×5×5×3	Hidden layer HLV1	900
Convolutional layer ConvD2	25×5×5	Hidden layer HLV2	450
Convolutional layer ConvC1	25×5×5×3	Hidden layer HLM	300
Convolutional layer ConvC2	25×5×5	Hidden layer HLS	63
Max pooling, steps	2×2×3	Output layer	21

Table 2. *ChaLearn 2014 "Looking at people Challenge (track 3)"* results (top 10).

Rank	Team	Score	Rank	Team	Score
1	**Ours**	**0.8500**	6	stevenwudi	0.7873
2	CraSPN	0.8339	7	ismar	0.7466
3	JY	0.8268	8	Quads	0.7454
4	CUHK-SWJTU	0.7919	9	Telepoints	0.6888
5	lpigou	0.7888	10	TUM-fortiss	0.6490

We followed the evaluation procedure proposed by the challenge organizers and adopted the Jaccard Index to quantify model performance:

$$J_{s,n} = \frac{A_{s,n} \cap B_{s,n}}{A_{s,n} \cup B_{s,n}}, \tag{7}$$

where $A_{s,n}$ is the ground truth label of gesture n in sequence s, and $B_{s,n}$ is the obtained prediction for the given gesture class in the same sequence. Here $A_{s,n}$ and $B_{s,n}$ are binary vectors where the frames in which the given gesture is being performed are marked with 1 and the rest with 0. Overall performance was calculated as the mean Jaccard index among all gesture categories and all sequences, with equal weights for all gesture classes.

5.2 Baseline model

In addition to the main pipeline, we have created a baseline model based on an ensemble classifier trained in a similar iterative fashion but on purely handcrafted descriptors. It was done to explore relative advantages (and disadvantages) of using learned representations and also the nuances of fusion. In addition, due to differences in feature formulation as well as in the nature of classifiers, we found it beneficial to combine the proposed deep network with the baseline method in a hybrid model as separately two models make different errors (see Table 4).

We use depth and intensity hand images and extract three sets of features. HoG features describe the hand pose in the image plane, and histograms of depths describe pose along the third spatial dimension. The third set reflects temporal dynamics of the hand shape.

HoG features from intensity images. First, we make use zero-mean and unit variance-normalized intensity images to extract HoG features $\mathbf{h}_{\text{int}}$ [40] at 9 orientations from a 2-level spatial pyramid [41], i.e. from the whole image and a magnified version of it containing 3 × 3 cells.

Histograms of depths. 9-bin depth histograms $\mathbf{h}_{\text{dep}}$ are extracted on two scales from depth maps of both hands: from a whole map and from each quarter of its upsampled version (by a factor of 2).

Derivatives of histograms. First derivatives of HoGs and depth histograms are calculated: $\delta\mathbf{h}(t) \approx \mathbf{h}(t+1) - \mathbf{h}(t-1)$, where $\mathbf{h}$ can stand for both $\mathbf{h}_{\text{int}}$ and $\mathbf{h}_{\text{dep}}$. Combined together, these three sets of features form a 270-dimensional descriptor $[\mathbf{h}_{\text{int}}, \mathbf{h}_{\text{dep}}, \delta\mathbf{h}_{\text{int}}, \delta\mathbf{h}_{\text{dep}}]$ for each frame and, consequently, a descriptor of dimension of 1350 for the dynamic pose of each hand.

Extremely randomized trees (ERT) [42] are adopted for data fusion and gesture classification. Ensemble methods of this sort have generally proven to be especially effective in conjunction with handcrafted features. During training, we followed the same iterative strategy as in the case of the neural architecture. First, three ERT classifiers are trained independently on (i) skeleton descriptors (the same as described in Section 3.1)), (ii) video features for the right hand and (iii) video features for the left hand. Once training is completed, features from all modalities with importance above the mean value are selected and once again fused for training a new, general ERT classifier. Feature importance is calculated as mean decrease in impurity (i.e. total decrease in node impurity weighted by proportion of samples reaching that node and averaged over all trees [43]).

At each step, ERT classifiers are trained with 300 estimators, an information gain criterion, no restrictions in depth and $\sqrt{N_f}$ features considered at each step (where N_f is the total number of features).

5.3 Results and discussion

The top 10 scores of the challenge are reported in Table 2. Our winning entry corresponding to a hybrid model (i.e. a combination of the proposed deep neural architecture and a baseline model (see Section 5.2)) surpasses the second best score by a margin of 1.61 percentage points. We also note that the multi-scale neural architecture still demonstrates the best performance, as well as the top one-scale neural model alone (see Tables 3 and 4).

Detailed information on the performance of neural architectures at each scale is provided in Table 3, including the multi-modal setting and per-modality tests. Interestingly, the discriminative power of articulated pose strongly depends on the sampling step and achieves a maximum value in the case of large sliding windows. On the other hand, video streams, containing information about hand shape and articulation, seem to be less sensitive to this parameter and demonstrate very good performance even for short spatio-temporal blocks. This signifies that in the context of this dataset, a body pose is interesting exclusively in terms of its dynamics, while hand postures are fairly discriminative alone, even in nearly static mode. The overall highest performance is nevertheless obtained in the case of a dynamic pose with duration roughly corresponding to the length of an average gesture (s=4, i.e. 17 frames).

Table 3. Performance at different temporal scales (deep learning + binary motion detector). All numbers reported in the table are the Jaccard Index.

Step	Articulated pose	Video	All
2	0.6938	0.7862	0.8188
3	0.7734	0.7926	0.8255
4	0.7891	0.7990	0.8449
all	**0.8080**	**0.8096**	**0.8488**

Table 4. Performance of different architectures (Jaccard Index).

Model	W. motion detector	W/o motion detector	(Rank)
Deep learning (proposed)	0.8118	**0.8488**	(1)
ERT (baseline)	0.7278	0.7811	(6)
Deep learning + ERT (hybrid)	0.8143	**0.8500**	(1)

The comparative performances of the baseline and hybrid models are reported in Table 4. In spite of low scores of the isolated ERT baseline model, fusing its outputs with the ones provided by the neural architecture is still slightly beneficial, mostly due to differences in feature formulation in the video channel (adding ERT to mocap alone did not result in a significant gain).

For each combination, we also provide results obtained with a classification module alone (without additional gesture localization) and coupled with the binary motion detector. The experiments have shown that the localization module contributes significantly to overall performance.

The deep learning architecture is implemented with the Theano library. A single scale predictor operates at frame rates close to real time (24 fps on GPU).

6 Conclusion

We have presented a general method for gesture and near-range action detection from a combination of depth and intensity video and articulated pose data. The model can be extended by adding alternative sensory pathways without significant changes in the architecture. It can elegantly cope with more spatial or temporal scales. Beyond scaling, an interesting direction for future work is a deeper exploration into the dynamics of cross-modality dependencies. Considering full signal reconstruction (similar to [35]), or explicit feedback connections as in the case of Deep Boltzmann Machines [36] would be helpful in the case when the input from one or more modalities is missing or noisy.

Acknowledgments This work was partially funded by a French grant INTERABOT, call "Investissements d'Avenir / Briques Génériques du Logiciel Embarqué", and by a French region "Rhônes Alpes" through the ARC6 research council.

References

1. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. (2014)
2. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. In: ICLR. (2014)
3. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet Classification with Deep Convolutional Neural Networks. In: NIPS. (2012)
4. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning Hierarchical Features for Scene Labeling. In: PAMI. Volume 35(8). (2013) 1915–1929
5. Couprie, C., Clment, F., Najman, L., LeCun, Y.: Indoor Semantic Segmentation using depth information. In: ICLR. (2014)
6. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11) (1998) 2278–2324
7. Kahou, S.E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, c., Memisevic, R., Vincent, P., Courville, A., Bengio, Y.: Combining modality specific deep neural networks for emotion recognition in video. In: ICMI. (2013)
8. Taigman, Y., Yang, M., Ranzato, M.A., Wolf, L.: DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In: CVPR. (2014)
9. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A.: Spatio-Temporal Convolutional Sparse Auto-Encoder for Sequence Classification. In: BMVC. (2012)
10. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Li, F.F.: Large-scale Video Classification with Convolutional Neural Networks. In: CVPR. (2014)
11. Simonyan, K., Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos. In: arXiv preprint arXiv:1406.2199v1. (2014)
12. Escalera, S., Baró, X., Gonzàlez, 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 workshop. (2014)
13. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. IJCV **103**(1) (2013) 60–79
14. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. BMVC (2009) 124.1–124.11
15. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior Recognition via Sparse Spatio-Temporal Features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. (2005)
16. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008)
17. Kläser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: BMVC. (2008)
18. Willems, G., Tuytelaars, T., Gool, L.V.: An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. In: ECCV. (2008) 650–663
19. Keskin, C., Kiraç, F., Kara, Y., Akarun, L.: Real time hand pose estimation using depth sensors. In: ICCV Workshop on Consumer Depth Cameras, IEEE (2011)
20. Półrola, M.a., Wojciechowski, A.: Real-time hand pose estimation using classifiers. In: Computer Vision and Graphics. Volume 7594., Springer (2012) 573–580
21. Tang, D., Yu, T.H., Kim, T.K.: Real-time Articulated Hand Pose Estimation using Semi-supervised Transductive Regression Forests. In: ICCV. (2013)
22. Tompson, J., Stein, M., LeCun, Y., Perlin, K.: Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks. In: ACM Transaction on Graphics. (2014)

23. Oikonomidis, I., Kyriazis, N., Argyros, A.: Efficient model-based 3D tracking of hand articulations using Kinect. In: BMVC. (2011) 101.1–101.11
24. Qian, C., Sun, X., Wei, Y., Tang, X., Sun, J.: Realtime and Robust Hand Tracking from Depth. In: CVPR. (2014)
25. Wang, F., Li, Y.: Beyond Physical Connections: Tree Models in Human Pose Estimation. In: CVPR. (2013)
26. Tang, D., Chang, H.J., Tejani, A., Kim, T.K.: Latent Regression Forest: Structured Estimation of 3D Articulated Hand Posture. In: CVPR. (2014)
27. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: CVPR. (2012)
28. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured Human Activity Detection from RGBD Images. In: ICRA. (2012)
29. Chen, X., Koskela, M.: Online RGB-D gesture recognition with extreme learning machines. In: ICMI. (2013)
30. Nandakumar, K., Wah, W.K., Alice, C.S.M., Terence, N.W.Z., Gang, W.J., Yun, Y.W.: A Multi-modal Gesture Recognition System Using Audio, Video, and Skeletal Joint Data Categories and Subject Descriptors. In: 2013 Multi-modal Challenge workshop in conjunction with ICMI. (2013)
31. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR. (2011) 3361–3368
32. Ranzato, M., Huang, F.J., Boureau, Y.L., LeCun, Y.: Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition. In: CVPR. (2007)
33. Chen, B., Ting, J.A., Marlin, B., de Freitas, N.: Deep learning of invariant Spatio-Temporal Features from Video. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning. (2010)
34. Ji, S., Xu, W., Yang, M., Yu, K.: 3D Convolutional Neural Networks for Human Action Recognition. PAMI **35**(1) (January 2013) 221–31
35. Ngiam, J., Khosla, A., Kin, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML. (2011)
36. Srivastava, N., Salakhutdinov, R.: Multimodal learning with Deep Boltzmann Machines. In: NIPS. (2013)
37. Neverova, N., Wolf, C., Paci, G., Sommavilla, G., Taylor, G.W., Nebout, F.: A multi-scale approach to gesture detection and recognition. In: ICCV Workshop on Understanding Human Activities: Context and Interactions (HACI). (2013)
38. Zanfir, M., Leordeanu, M., Sminchisescu, C.: The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Recognition and Detection. In: ICCV. (2013)
39. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML. (2009)
40. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: CVPR. (2005)
41. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: CVPR. (2006)
42. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. In: Machine learning, 63(1), 3-42. (2006)
43. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and regression trees. (1984)