A multi-scale approach to gesture detection and recognition

Natalia Neverova, Christian Wolf
Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205, F-69621, France
firstname.surname@liris.cnrs.fr

Giulio Paci, Giacomo Sommavilla
ICST, CNR
Padova, Italy
firstname.surname@pd.istc.cnr.it

Graham W. Taylor
University of Guelph
Guelph, Canada
gwtaylor@uoguelph.ca

Florian Nebout
Awabot
Lyon, France
florian.nebout@awabot.com

Abstract

We propose a generalized approach to human gesture recognition based on multiple data modalities such as depth video, articulated pose and speech. In our system, each gesture is decomposed into large-scale body motion and local subtle movements such as hand articulation. The idea of learning at multiple scales is also applied to the temporal dimension, such that a gesture is considered as a set of characteristic motion impulses, or dynamic poses. Each modality is first processed separately in short spatio-temporal blocks, where discriminative data-specific features are either manually extracted or learned. Finally, we employ a Recurrent Neural Network for modeling large-scale temporal dependencies, data fusion and ultimately gesture classification. Our experiments on the 2013 Challenge on Multimodal Gesture Recognition dataset have demonstrated that using multiple modalities at several spatial and temporal scales leads to a significant increase in performance allowing the model to compensate for errors of individual classifiers as well as noise in the separate channels.

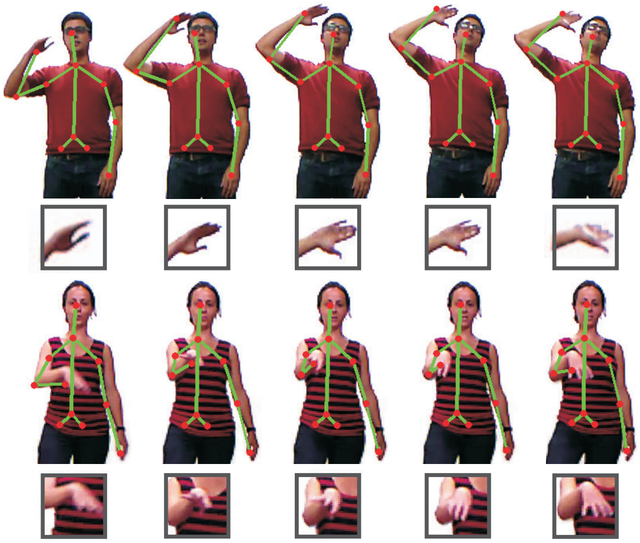

Figure 1. Examples of two-scale gesture decomposition: upper body movement and hand articulation. The gesture in the top row can be fully characterized by large-scale body motion, whereas in the one below, subtle finger movements play the primary role.

1. Introduction

Understanding human motion is one of the most intriguing aspects of computer vision, not only as a captivating scientific puzzle, but also due to its practical importance. Various applications require interpreting near-range human activities, where body parts are clearly visible and articulated, i.e. *gestures*.

Numerous gesture taxonomies have been proposed, but there is little agreement among researchers on what gesture characteristics are the most discriminative and useful. In this work we focus on *intentional* gestures, bearing *communicative function* and targeting enhancement of the communicative value of speech by doubling or conveying additional semantic information. Depending on the application, their functional load may be more significant than simply augmenting the audio channel: in robot-computer interaction, for example, gestures often play the primary role.

Conversational gestures, i.e. hand movements accompanying speech, may convey semantic meaning, as well as information about a speaker's personality and cultural specifics, momentary emotional state, intentions and attitude towards the audience and subject at hand. Interestingly, a number of psychological studies (e.g. [28]) suggest that motor movements do not only aim at illustrating pronounced utterances, amplifying their emotional impact or compensating for the lack of linguistic fluency, but, in turn, influence the process of speech production allowing the speaker to produce more complex and vivid verbal descriptions. Therefore, an ability to recognize and interpret non-verbal cues coupled with verbal information would be a natural step towards understanding and plausible simulation of human behavior and, consequently, making human-machine communication truly effortless and intuitive.

Much of this work has been conducted in the context of speech recognition, while the vision community has tackled the problem of motion recognition from its own perspective. The fusion of data so different by nature is not straightforward both in terms of modeling the underlying processes and simply due to engineering issues. However, with recent advances in GPU computing together with a rapidly growing market of consumer-grade multimodal sensors, the possibility to collect and process large amounts of multimodal high-dimensional data present new opportunities.

The vision community has recently seen a number of efforts made in applying the idea of multimodality to recognizing human activities. An action recognition system proposed in [25] is built on extracting depth-based spatio-temporal descriptors (in the spirit of HoG) and combining them with joint angle similarities calculated from skeletons. In a similar approach [40] local depth and skeletal features are fused using random forests. There are a number of recent works based on fitting 3D models of hands and dedicated to hand tracking [26] and pose recognition [19] from video RGB-D input. In [24] the authors detect head nodes from video by considering the audio channel as self-context and conditioning sensitivity of the visual detector response on whether or not the person is speaking.

In this work we explore three data modalities: range video data, articulated pose, and speech. In our model, each gesture can be seen as a composition of large-scale motion (for example, of torso, limbs or a head) and subtle movements, such as finger articulation (see Fig. 1). Most modern gesture recognition systems exploit only one of these levels. Existing touchless interfaces based on precise finger tracking have a limited working range of distances (typically 0.5-1.2m) or require that the subject wear special markers or other attire, such as colored gloves or t-shirts [35]. They are therefore often impractical, while full body skeleton-based models typically require unnatural exaggerated articulation, have a very limited vocabulary and therefore lack expressive power. By comparison, our model can be extended or narrowed to any required level or scale based on data.

We also apply the hierarchical approach to the temporal dimension: by decomposing each gesture into a set of (typically ordered) characteristic motion impulses, or "dynamic poses" and integrating consecutive poses extracted from different modalities over a longer time span.

2. Related work

Motion related vision, which encompasses both gesture and action recognition, has gained the attention of many vision scientists over the last two decades. Dozens of papers published each year consider this problem in various contexts: still images, video, multiple views and range data, point clouds, etc. An extensive overview of action recognition from images can be found in numerous surveys (e.g. [36, 14]), while we will focus exclusively on approaches applicable to spatio-temporal data.

Most existing methods proposed for human motion recognition from video are based on extracting spatial and temporal features (representations) followed by classification. Such representations can be either engineered or learned from the data, either separately from each dimension or together from a spatio-temporal 3D volume.

Spatio-temporal descriptors are typically extracted densely (on a regular grid or globally), sparsely around salient interest points or along sparse or dense trajectories [33]. However, it has been shown by [34] that dense sampling generally leads to higher classification rates when it comes to complex realistic data. Among the most widely accepted engineered local descriptors one could name Cuboid [5], HoG/HoF [11], HoG3D [15], ESURF [37] , 3D-SIFT [29] and several others.

In parallel with general approaches, a great amount of ad-hoc methods have been proposed specifically for hand-gesture recognition in narrow contexts. Most of them rely on hand detection, tracking, and gesture recognition based on global hand shape descriptors such as contours, silhouettes, fingertip positions, palm center, number of visible fingers, etc. [30, 20]. Similar descriptors have been proposed for depth and RGBD data [21].

The family of deformable part models forms a separate class of methods broadly explored in the context of actions ([6], [39]). Nevertheless, there are no fundamental obstacles for their adaptation to gesture recognition.

Instead of hand-crafting, efficient representations can be inferred directly from data by minimizing reconstruction error, e.g. as in autoencoders, or some predefined energy function, e.g. as in Restricted Boltzmann Machines (RBMs). Le et al. [16] used Independent Subspace Analysis (ISA) for computationally efficient learning of hierarchies of invariant spatio-temporal features. In [1] the authors adapted the original work of Ranzato et al. [27] by adding a temporal dimension to 2D sparse convolutional autoencoders. Taylor et al. [7] extended and scaled up the Gated RBM (GRBM) model proposed by Memisevic and Hinton [22] for learning representations of image patch transformations. Chen et al. [3] used convolutional RBMs as basic building blocks to construct the Space-Time Deep Belief Networks (ST-DBN) producing high-level representations of video sequences. Ji et al. [13] applied hardwired filters to obtain low level feature maps and then fed them to a 3D convolutional network (ConvNet) for joint learning of mid-level spatio-temporal representations and classification.

Treating spatial and temporal dependencies in the same way is often problematic, since the particular nature of the temporal dimension is ignored. Alternatively, the temporal dynamics of motion can be modeled by a time-series algorithm. In this context, generative Hidden Markov Models (HMMs), discriminative Conditional Random Fields (CRFs) and their extensions have gained mainstream acceptance and proven to be efficient in many relatively simple recognition tasks. A comparative study [23] has shown that in the context of action recognition CRFs generally perform better than HMMs when using spatial features and worse when optical flow is used. Most of the highly-ranked participants of a recent ChaLearn Gesture recognition challenge claimed to use HMMs, CRFs or similar models [8].

At the same time, sequence models that do incorporate temporal dynamics often oversimplify the data structure and fail to capture high-dimensional, non-linear and long-range dependencies. Drawing an analogy with the previous case of static spatio-temporal features, supporters of connectionist approaches went one step further aiming to learn representations of temporal dynamics from data rather than explicitly modeling them under hardwired assumptions. In this context, Time Delayed Neural Networks (TDNN) were applied to American sign language recognition as early as in 1999 [38]. Recurrent Neural Networks (RNNs), being promising in theory, in practice appear to be difficult to train. To address this issue, several techniques such as Long Short-term Memory RNN (LSTM-RNN) [10] and Echo-State Networks (ESN) [12] have been proposed. In [1] the authors applied LSTM-RNN to action recognition in combination with 3D ConvNets and sparse autoencoders.

The inability to model highly structured multimodal data can be a serious issue in the context of human motion recognition. This has prompted more representationally powerful generative models based on distributed hidden states, such as Temporal RBMs [31]. Their special case, Conditional RBMs [32], where connections between temporal hidden units are removed, are known to be faster and easier to train.

3. Model description

Our multimodal system is based on integrating three sources of information: raw range video data, body skeleton (articulated pose) and an audio stream. It benefits from operating at two spatial scales: upper body motion and hand articulation, and at two temporal scales: gestures and momentary movements, or "dynamic poses".

For each modality, dynamic poses are modeled separately by stacking a small number of consecutive frames in short spatio-temporal blocks, while final data fusion is performed at a larger time scale using a recurrent neural network (RNN) for capturing temporal dependencies and integrating over a time span corresponding to the average duration of a gesture.

We denote by L_1 the duration of a short spatio-temporal block corresponding to a dynamic pose (in practice set to 5 frames, or $1/4$s) and by L_2 time span of a typical gesture (in our model roughly corresponds to 2s). We sample overlapping dynamic poses starting from each second data instance and long sequences starting from each dynamic pose: thus, a video consisting of F frames can be decomposed into $I = (F - L_1)/2 + 1$ dynamic poses and $R = I - L_2 + 1 = (F - L_1)/2 - L_2 + 2$ long sequences.

Let N be the size of the gesture vocabulary to recognize. Here we do not limit ourselves to a pure classification task, since in an unconstrained "real-world" context *detection* is crucial. To address this problem, we introduce an additional "clutter" class number 0 and train our model on a huge amount of instances containing out-of-vocabulary movements using a bootstrapping strategy.

On the first step, we exploit short spatio-temporal blocks to train individual classifiers for all three modalities. For the depth video data we train a convolutional network followed by a multilayer perceptron, for the skeletons we formulate a pose descriptor and, in turn, train another multilayer perceptron on it. We treat the output of the speech recognition system as a "bag-of-words". The exact way of processing of each channel is described in the following subsections.

In all three cases, for each spatio-temporal block and each modality we obtain a distribution over $N + 1$ classes corresponding to neuron activations (for visual data) or class frequencies (for audio channel). These distributions are then concatenated in longer sequences and fed to a recurrent neural network classifier. On the testing stage all dynamic poses are labeled, each dynamic pose $i \in [1 \ldots I]$ participates in the classification process several times as a member of L_2 consecutive overlapping long sequences $r \in [1 \ldots R]$ (or less, at boundaries), and each time it is assigned a distribution of neuron activations $\sigma_{i,r,n}$ over classes $[0 \ldots N]$ (n is a class index). The averaged distribution for each dynamic pose is then calculated as follows:

$$\sigma_{i,n} = \sum_{r=r_{i,1}}^{r_{i,2}} \frac{\sigma_{i,r,n}}{r_{i,2} - r_{i,1}}, \tag{1}$$

$$r_{i,1} = \max(1, i - L_2 + 1), \quad r_{i,2} = \min(i, I - L_2 + 1).$$

The resulting label is then calculated via finding the most probable class and thresholding:

$$l_i = \begin{cases} \arg\max_n (\sigma_{i,n}), & \text{if } \max(\sigma_{i,n}) > \tau \\ 0, & \text{else} \end{cases} \tag{2}$$

The parameter τ is set empirically. All neighboring dynamic poses assigned to the same labels are then aggregated

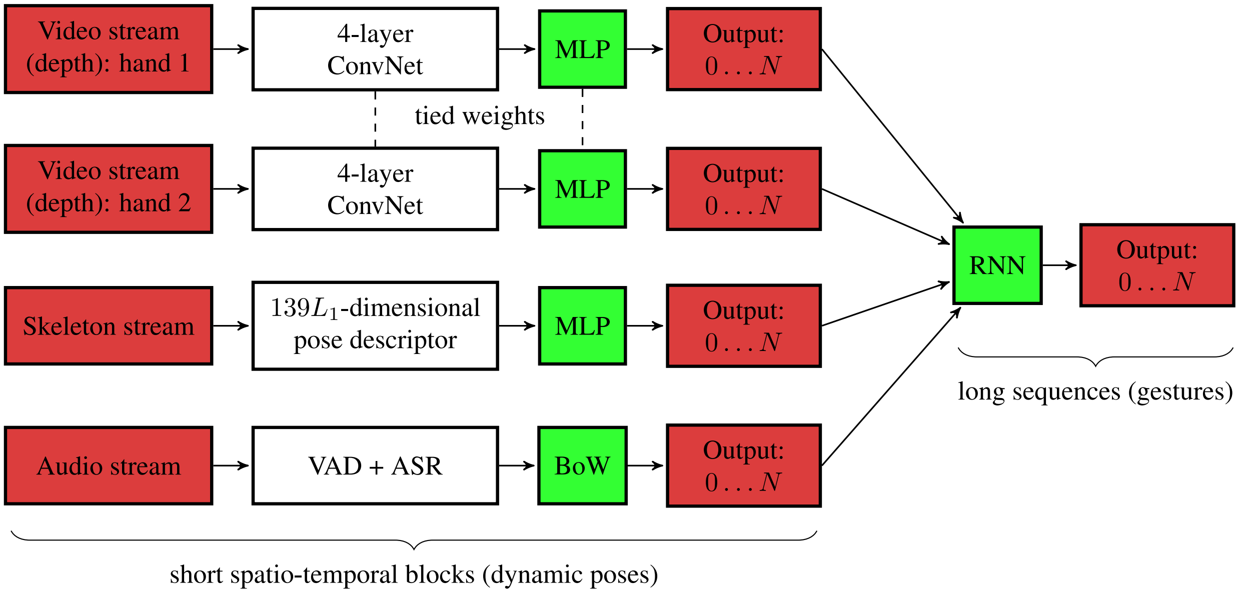

Figure 2. The proposed model operates at two temporal scales: first, we train individual classifiers for each data modality (depth video, articulated pose, audio stream) at the level of short spatio-temporal blocks (which we call "dynamic poses"); second, the outputs of all classifiers are fused using a recurrent neural network (RNN) at the time span of an average gesture.

into a single gesture. The full diagram of the model workflow is shown in Fig. 2.

3.1. Video stream

Starting with the depth video stream, we do not rely on hand-crafted descriptors and instead consider pixel values of the depth map as low-level features. Since in this case we are interested in capturing fine movements of palms and fingers, from each frame we extract two bounding boxes around the right and left hand centered at the hand positions provided by the skeleton (joints HandRight and HandLeft). To eliminate the influence of the person's position with respect to the camera and keep the hand size approximately constant, the size of each bounding box is normalized by the distance between the hand and the sensor:

$$H_{\{x,y\}} = h_{\{x,y\}} \frac{\{X,Y\}}{z \cdot \tan\left(\alpha_{FoV,\{x,y\}}\right)}, \qquad (3)$$

where $H_{\{x,y\}}$ is the hand size (px) along the x and y axes of the camera sensor, $h_{\{x,y\}}$ – the physical size of an average hand, in mm, $X \times Y$ – the frame size in pixels, z – the distance between the hand and the camera in mm, $\alpha_{\mathrm{FoV},\{x,y\}}$ – the camera field of view along the x and y axes respectively.

Due to built-in smoothing parameters, skeleton tracking is often inertial. Thus positions of quickly-moving joints are typically detected less accurately. Reduced smoothing results in introducing additional noise and jitter. To compensate for these effects, we correct positions of the hand joints by minimizing inter-frame square root differences between corresponding blocks within each dynamic pose.

As a preprocessing step, we subtract background from each frame by simple thresholding along the depth axis and apply local contrast normalization to zero mean and unit variance over local neighborhoods.

Finally, we use extracted blocks for supervised training of a convolutional network [17] consisting of 2 convolutional layers with $\tanh$ activations and 2 sub-sampling layers (ConvNet in Fig. 2). The first pair of layers performs 3D convolutions of short spatio-temporal blocks (dynamic poses), followed by max pooling over spatial and temporal dimensions. The second pair of layers performs 2D convolutions and spatial max pooling. The output of the 4-th layer is fully connected to a multilayer perceptron (MLP).

All layers of the network are trained jointly and in a supervised way by backpropagation. During training, we do not differentiate between "right hand" and "left hand" images. Instead, all left-hand blocks are mirrored horizontally to eliminate differences in hand orientation.

In real-life contexts both single-hand and two-hand gestures may take place. To avoid manual annotation of which hand is active and which is passive, we calculate variances in positions of both hands within each dynamic pose. Then, if the gesture is known to involve both hands or if the given hand is more active, it gets assigned with the ground truth label corresponding to the gesture, otherwise it is labeled with the class 0.

The output of the MLP is a $N + 1$-way softmax, indicat-

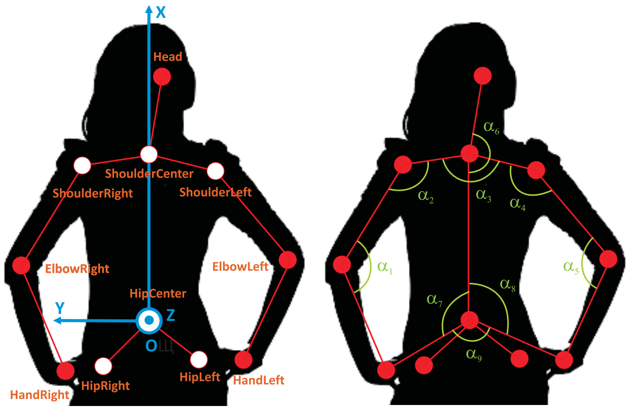

Figure 3. The pose descriptor is calculated from normalized coordinates of 9 upper body joints (shown on the left) as a set of angles formed by triples of joints (shown on the right) and pairwise distances. The body coordinate system (shown in blue) is calculated based on 6 torso joints (shown in white).

ing the probability of the hand pose instance to be assigned to each of the $N + 1$ gesture classes.

3.2. Articulated pose

The full body skeleton provided by current consumer depth cameras consists of 20 joints identified by their coordinates in a 3D coordinate system aligned with the depth sensor. For our purposes we exploit only 11 of them corresponding to the upper body (see Fig. 3), we also do not use wrist joints as their detected positions are often unstable.

To calculate pose descriptors we follow the logic proposed in [2]. First, we translate the origin of the coordinate system to the HipCenter position and normalize all coordinates by the distance between the HipCenter and ShoulderCenter joints in order to compensate for differences in human heights. Second, we calculate 3 vectors $\vec{x}$, $\vec{y}$ and $\vec{z}$ describing body orientation: to do that we apply PCA on 6 torso joint coordinates (shown in white on Fig. 3). As a result, vectors $\vec{x}$ and $\vec{y}$ form the body plane, while $\vec{z}$ is approximately perpendicular and points towards the camera.

On the next step, we calculate 9 inclination angles $\alpha_{1...9}$ between real and virtual "bones", i.e. vectors connecting pairs of joints (see Fig. 3). Also, to characterize the orientation of all joints with respect to the body coordinate system, for each pair of "bones" used in the previous case, we calculate another 9 angles $\beta_{1...9}$ between projections of the first bone vector and the vector $\vec{y}$ on the plane perpendicular to the orientation of the second bone. The third set of angles $\gamma_{1...11}$ is calculated between each of vectors connecting joints with the camera sensor position and vector $\vec{z}$.

Finally, we calculate pairwise distances $d_{i,j}$ between all 11 joints ($i \neq j$). Overall, it gives us a 139-dimensional pose descriptor for each video frame:

$$PD = [\alpha_{0...9}, \beta_{0...9}, \gamma_{0...11}, d_{0...110}]. \quad (4)$$

We stack pose descriptors extracted from L_1 consecutive frames into a single feature vector which thus describes the corresponding dynamic pose. Such vectors are then used for training a fully-connected MLP with sigmoid units (see Fig. 2). To account for the fact that each gesture can be performed by either right or left hand (or both), we augment the training dataset with pose descriptors mirrored horizontally with respect to the camera coordinate system.

As in the case of the depth video stream, the output at this step is a $N + 1$-way softmax, with the only difference being that here we consider dynamic poses not with respect to each hand but to the whole body.

3.3. Audio stream

Our audio processing module uses a simple word-spotting strategy which assumes that each gesture has a limited verbal vocabulary associated with it. However, numerous practical issues such as illiterate speech, variations in dialects, differences in speech levels (i.e. idiomatic, casual, even ungrammatical colloquial speach vs grand style) make high demands on the level of system generalization.

In order to avoid using complex language models, we associate each class with a single virtual gesture "word": a set of verbal words (in the linguistic sense), word-combinations or short phrases having the same semantic meaning and typically accompanying each given gesture. To do that, we construct a dictionary including all possible utterances associated with each gesture word in the form of sequences of phonemes. Here we do not differentiate between slight variations in phrase constructions and different pronunciations of the same phrase. For example, an Italian gesture "sei_pazzo?" ("are you crazy?") can be associated with phonetic transcriptions "s E i p a tts o", "m a s E i p a tts o" and "s E i m a tt o" (corresponding to the "sei pazzo", "ma sei pazzo" and "sei matto" orthographic forms). Depending on the training data and the task at hand, the dictionary can be populated by hand, aiming on including the greatest possible number of ways to express every single gesture.

The proposed framework consists of two modules (implemented with the Julius LVCSR engine [18]). First, Voice Activity Detection (VAD) is applied to isolate single speech gesture events (with start and end timestamps), then an automatic speech recognition (ASR) system takes over considering each isolated event as a word instance.

Typically, ASR systems provide a lattice (also called a "wordgraph") that for each recognised word gives timing, scores and possible connections with other recognised words. For this task, we simplified the algorithm to produce an n-best list for every gesture event. Each list contains an infinite-dimension sparse vector of "hypotheses" (gesture classes with associated confidence scores).

As the last step, we select a fixed number W of hypotheses with the highest scores and treat the ASR system output in the bag-of-words fashion calculating frequencies of appearances of each of N gesture classes and normalizing the counts by W. As a result, we obtain a distribution of class probabilities that has the same structure as the outputs produced from the video and skeleton modalities.

3.4. Data fusion

All outputs of the ensemble of single-modality classifiers are fused with a recurrent neural network (RNN) which serves as a meta-classifier. By *output* here we, as before, mean a distribution over $N + 1$ classes produced either by a softmax output of an MLP (in the case of the video and skeleton streams) or the bag-of-words model (in case of the audio stream). Here we prefer to fuse outputs of sole classifiers rather than merging extracted features since, apart from the fact that this strategy is extremely computationally demanding and depending on training data may result in the curse of dimensionality, it allows the model to benefit from pre-training of individual modalities in discriminative way.

Since we assume that all gestures are independent from each other and their order is randomized, incorporating really long-term dependencies in the model is not beneficial and even harmful. Therefore, for RNN training and testing we do not use a continuous data stream, but rather split the input into sequences with the length L_2 roughly corresponding to the duration of a typical gesture. As a result, the RNN input is a sequence of L_2 $4L_1$-dimensional feature vectors.

During meta-classification all modalities are synchronized and the combined stream is represented as a sequence of overlapping integrated multimodal dynamic poses. All such poses are processed successively, while for each given pose its components are fed to their own preliminary classifiers in parallel (see Fig. 2).

Straightforward training of RNNs by backpropagation through time is known to be problematic due to exponentially vanishing gradients [9]. To address this issue, we employ a 2-stage training procedure. All recurrent weights are first set to 0 and the RNN is trained as a vanilla MLP. Once the process has converged, we fine tune the weights keeping the feedforward connections unchanged.

4. Experimental results

We participated in the 2013 Multi-modal Gesture Recognition Challenge (see Fig. 4) ranking 6th overall (54 teams participated with 20 submissions on the final test set, 17 method descriptions provided, see Table 1). The competition dataset consists of RGB-D video and audio recordings (shot with the Kinect sensor) of 13,858 manually annotated Italian conversational gestures, where 20 classes of "useful" gestures (i.e. recognizable) are augmented with arbitrary out-of-vocabulary gestures, movements and sounds. Skeleton data is also provided.

The dataset is initially split into training, validation and test subsets. In our experiments, we combine the first two subsets, use 90% of the data for training and the rest for validation. The performance is reported on the test subset.

For these experiments, we set the temporal length of a dynamic pose equal to $L_1 = 5$ frames and the duration of a gesture equal to $L_2 = 16$ dynamic poses (estimated from the training data).

As video input, we use depth blocks of 72×72 pixels and perform local contrast normalization, where standard deviation is estimated over a 11×11 neighborhood. The hand size is set to 180 mm with a safety factor of 1.5. The first layer of the convolutional network consists of 25 filters $11 \times 11 \times 3$ (the last dimension is temporal) and is followed by spatio-temporal max pooling $2 \times 2 \times 3$ transforming 3D feature maps into 2D. The third layer again consists of 25 filters 5×5 followed by spatial pooling 2×2. The last, fully connected layer of the video path consists of 700 neurons.

The pose descriptor in our implementation is a $139 \times 5 = 695$ dimensional vector. The MLP operating on pose descriptors consists of 350 hidden units.

An acoustic model of the ASR system has been built using the EVALITA speech data set, which is a subset of the CLIPS corpus published for the EVALITA 2011 Forced Alignment task [4]. This is a relatively small Italian adult speech corpus featuring colloquial speech and including numerous Italian dialects. The audio data from the gesture recognition dataset has not been used, neither for training nor for adapting the ASR module. The number of hypothesis used in a bag-of-words model is set to $W = 9$.

Finally, a RNN having 200 hidden units is used for data fusion. For this challenge, we put a prior on the sequence length and the confidence threshold τ is set accordingly (for the proposed method $\tau = 0.925$).

Following the procedure originally proposed by the challenge organizers, we evaluate the performance as the edit (Levenshtein) distance (ED) between an ordered sequence of gestures recognized by the system and the ground truth (one sequence corresponds to one video from the dataset). This metric is calculated as a number of uniformly penalized edit operations (substitution, insertion, deletion) necessary to transform one sequence into another. The overall score is a sum of the edit distances over the whole test set divided by the real number of gesture instances. We also provide values of precision and recall.

As is shown in Table 2, combining multiple modalities leads to a significant gain in performance when the quality of predictions from individual channels is not satisfactory. The model has proven to be invariant to person position and height, and insensitive to environment and illumination.

Team	ED	Rank	Team	ED	Rank
Team 1	0.1276	1	**Our team**	0.1773	6
Team 2	0.1539	2	Team 7	0.2445	7
Team 3	0.1711	3		. . .	
Team 4	0.1722	4	Team 16	0.8746	16
Team 5	0.1733	5	Team 17	0.9207	17

Table 1. Official results of the challenge on gesture recognition.

Modalities used	Recall	Precision	ED
Independent dynamic poses (MLP)			
Depth video only	0.5433	0.5494	0.6613
Articulated pose	0.7298	0.7420	0.4250
Audio stream	0.6754	0.6590	0.4966
Depth + Pose	0.7618	0.7739	0.3809
Depth + Audio	0.7669	0.7796	0.3742
Pose + Audio	0.8765	0.8885	0.2125
Depth + Pose + Audio	**0.8784**	**0.8920**	**0.2091**
Modeling time dependencies (RNN)			
Depth + Pose	0.7810	0.7972	0.3440
Depth + Pose + Audio	**0.8939**	**0.9072**	**0.1786**
Random predictions			1.4747

Table 2. Experimental results on the competition dataset. Increase in recall and precision and decreasing edit distance indicate improvement in performance.

The major weakness of the audio module (partially compensated by combining with visual predictions) is mapping all sounds into predefined 20-gesture categories, which gives a large amount of false positives in gesture detection. In addition, the speech recognition module often fails when speech is too fast and continuous. In the case where several gesture-associated words are uttered in the same breath, it is likely that only one will be recognized. Finally, the ASR may produce ambiguous results when several gesture classes have the same meaningful words associated with them (e.g. “messi daccordo” can be associated with both “MESSIDACCORDO” and “DACCORDO” classes).

The visual module alone is sensitive to quality of annotations and data input: its performance can be significantly decreased if gestures are weakly articulated, too fast or too slow in comparison with training instances and have no pronounced boundaries between them. Too short gestures are the most difficult to recognize, since fast movements often cause errors in detection of skeleton joint positions (due to smoothing) and motion blur, which create additional difficulties for recognition from the video channel. In addition, our system, trained as a dynamic pose-based RNN, is not automatically time-scale invariant and assumes a certain range of possible speeds learned from the training data.

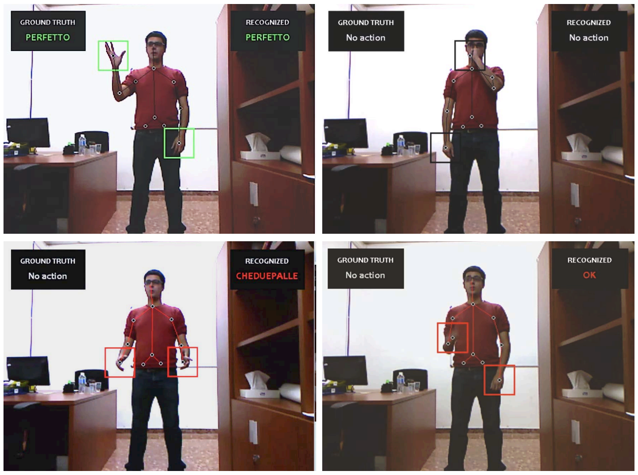

Figure 4. Examples of gestures: correctly recognized and correctly ignored (the first row), false detections due to high similarity between gesture elements (the second row).

Combining all modalities in a single framework allows the model to compensate for major weaknesses of individual modules and reduce the negative influence of noise.

5. Conclusion

We have described a generalized method for gesture and near-range action recognition from a combination of range video data, audio and articulated pose. The model can be further extended and augmented with arbitrary channels (depending on available sensors) by adding additional parallel pathways without significant changes in the general structure. Multiple spatial and temporal scales per channel can be easily introduced. As future work, we aim to reformulate and generalize the problem from gesture detection and recognition to sensing gesture parameters (such as, for example, motion amplitude in scrolling-like gestures). In this case, the video stream providing information about subtle hand movements will play a primary role.

References

[1] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Spatio-Temporal Convolutional Sparse Auto-Encoder for Sequence Classification. In *BMVC*, pages 124.1–124.12, 2012. 2, 3

[2] O. Çeliktutan, C. B. Akgül, C. Wolf, and B. Sankur. Graph-Based Analysis of Physical Exercise Actions. In *1st ACM MM Workshop on Multimedia Indexing and Information Retrieval for Healthcare (MIIRH)*, 2013. 5

[3] B. Chen, J.-A. Ting, B. Marlin, and N. de Freitas. Deep learning of invariant Spatio-Temporal Features from Video. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010. 2

[4] F. Cutugno, A. Origilia, and D. Seppi. Evalita 2011: Forced alignment task. In *Evaluation of Natural Language and Speech Tools for Italian*, pages 305–311, 2013. 6

[5] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior Recognition via Sparse Spatio-Temporal Features. In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 65–72, 2005. 2

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–45, Oct. 2010. 2

[7] G.Taylor, R.Fergus, Y.LeCun, and C.Bregler. Convolutional learning of spatio-temporal features. In *ECCV*, 2010. 2

[8] I. Guyon, V. Athitsos, P. Jangyodsuk, and B. Hamner. ChaLearn Gesture Challenge: Design and First Results. In *CVPR Workshop on Gesture Recognition and Kinect Demonstration Competition*, pages 1–6, 2012. 3

[9] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies. In *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press, 2001. 6

[10] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. 3

[11] I.Laptev, M.Marszalek, C.Schmid, and B.Rozenfeld. Learning realistic human actions from movies. *CVPR*, 2008 2

[12] H. Jaeger and H. Haas. Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *Science*, 304(78):78–80, Apr. 2004. 3

[13] S. Ji, W. Xu, M. Yang, and K. Yu. 3D Convolutional Neural Networks for Human Action Recognition. *PAMI*, 35(1):221–31, Jan. 2013. 2

[14] R. Khan and N. Ibraheem. Hand gesture recognition: a literature review. *International Journal of Artificial Intelligence & Applications (IJAIA)*, 3(4):161–174, 2012. 2

[15] A. Kläser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3D-gradients. In *BMVC*, 2008. 2

[16] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, pages 3361–3368. Ieee, June 2011. 2

[17] Y. LeCun and L. Bottou. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 4

[18] A. Lee, T. Kawahara, and K. Shikano. Julius - an open source real-time large vocabulary recognition engine. In *INTERSPEECH*, pages 1691–1694, 2001. 5

[19] O. Lopes, M. Pousa, S. Escalera, and J. Gonzalez. Multi Hand Pose Recognition System using Kinect Depth Sensor. In *ICPR Workshop on Gesture Recognition*, 2012. 2

[20] A. Malima, E. Özgür, and M. Çetin. A fast algorithm for vision-based hand gesture recognition for robot control. *IEEE 14th Conference on Signal Processing and Communications Applications*, 2006. 2

[21] C. M. Mateo, P. Gil, J. A. Corrales, S. T. Puente, and F. Torres. RGBD Human-Hand recognition for the Interaction with Robot-Hand. In *IROS*, 2012. 2

[22] R. Memisevic and G. Hinton. Unsupervised learning of image transformations. In *CVPR*, 2007. 2

[23] M. A. Mendoza and N. P. de la Blanca. Applying Space State Models in Human Action Recognition: A Comparative Study. In *5th International Conference on Articulated Motion and Deformable Objects*, pages 53–62, 2008. 3

[24] L. Nguyen, J.-M. Odobez, and D. Gatica-Perez. Using Self-Context for Multimodal Detection of Head Nods in Face-to-Face Interactions. In *ICMI*, 2012. 2

[25] E. Ohn-Bar and M. M. Trivedi. Joint Angles Similiarities and HOG for Action Recognition. In *CVPR Workshop on Human Activity Understanding from 3D Data (HAU3D13)*, pages 465–470, 2013. 2

[26] I. Oikonomidis, N. Kyriazis, and A. Argyros. Efficient model-based 3D tracking of hand articulations using Kinect. In *BMVC*, pages 101.1–101.11, 2011. 2

[27] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun. Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition. In *CVPR*, 2007. 2

[28] B. Rimé, L. Schiaratura, M. Hupet, and A. K. S. Ghysselinckx. Effects of relative immobilization on the speaker's nonverbal behavior and on the dialogue imagery level. *Motivation and Emotion*, 8, 1984. 1

[29] P. Scovanner, S. Ali, and M. Shah. A 3-Dimensional SIFT Descriptor and its Application to Action Recognition. In *15th International Conference on Multimedia*, 2007. 2

[30] E. Stergiopoulou and N. Papamarkos. Hand gesture recognition using a neural network shape fitting technique. *Engineering Applications of Artificial Intelligence*, 22(8):1141–1158, Dec. 2009. 2

[31] I. Sutskever and G. Hinton. Learning Multilevel Distributed Representations for High-Dimensional Sequences. In *11th International Conference on Artificial Intelligence and Statistics*, 2007. 3

[32] G. Taylor, G. Hinton, and S. Roweis. Modeling Human Motion Using Binary Latent Variables. In *NIPS*, 2006. 3

[33] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 103(1):60–79, 2013. 2

[34] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. *BMVC*, pages 124.1–124.11, 2009. 2

[35] R. Wang and J. Popović. Real-Time Hand-Tracking with a Color Glove. In *SIGGRAPH*, 2009. 2

[36] D. Weinland, R. Ronfard, and E. Boyer. A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition. Technical report, INRIA, 7212, 2010. 2

[37] G. Willems, T. Tuytelaars, and L. V. Gool. An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. In *ECCV*, pages 650–663, 2008. 2

[38] M.-H. Yang and N. Ahuja. Recognizing Hand Gesture Using Motion Trajectories. In *CVPR*, 1999. 3

[39] Y.Tian, R.Sukthankar, and M.Shah. Spatiotemporal Deformable Part Models for Action Detection. In *CVPR*, 2013. 2

[40] Y. Zhu, W. Chen, and G. Guo. Fusing Spatiotemporal Features and Joints for 3D Action Recognition. In *CVPR Workshop on Human Activity Understanding from 3D Data (HAU3D13)*, pages 486–491, 2013. 2