CrossMark

Beyond Temporal Pooling: Recurrence and Temporal Convolutions for Gesture Recognition in Video

Lionel Pigou[1] · **Aäron van den Oord[1]** · **Sander Dieleman[1]** ·
Mieke Van Herreweghe[2] · **Joni Dambre[1]**

Abstract Recent studies have demonstrated the power of recurrent neural networks for machine translation, image captioning and speech recognition. For the task of capturing temporal structure in video, however, there still remain numerous open research questions. Current research suggests using a simple temporal feature pooling strategy to take into account the temporal aspect of video. We demonstrate that this method is not sufficient for gesture recognition, where temporal information is more discriminative compared to general video classification tasks. We explore deep architectures for gesture recognition in video and propose a new end-to-end trainable neural network architecture incorporating temporal convolutions and bidirectional recurrence. Our main contributions are twofold; first, we show that recurrence is crucial for this task; second, we show that adding temporal convolutions leads to significant improvements. We evaluate the different approaches on the Montalbano gesture recognition dataset, where we achieve state-of-the-art results.

Keywords Gesture recognition · Deep neural networks

1 Introduction

Gesture recognition is one of the core components in the thriving research field of human–computer interaction. The recognition of distinct hand and arm motions is becoming increasingly important, as it enables smart interactions with electronic devices. Furthermore, gesture identification in video can be seen as a first step towards sign language recognition, where even subtle differences in motion can play an important role. Some examples that complicate the identification of gestures are changes in background and lighting due to the varying environment, variations in the performance and speed of the gestures, different clothes worn by the performers and different positioning relative to the camera. Moreover, regular hand motion or out-of-vocabulary gestures should not be confused with one of the target gestures.

Convolutional neural networks (CNNs) (LeCun et al. 1998) are the de facto standard approach in computer vision. CNNs have the ability to learn complex hierarchies with increasing levels of abstraction while being end-to-end trainable. Their success has had a huge impact on vision based applications like image classification (Krizhevsky et al. 2012), object detection (Sermanet et al. 2013), human pose estimation (Toshev and Szegedy 2014) and many more. A video can be seen as an ordered collection of images. Classifying a video frame by frame with a CNN is bound to ignore motion characteristics, as there is no integration of temporal information. Depending on the task at hand, aggregating the spatial features produced by the CNN with temporal pool-

✉ Lionel Pigou
lionel.pigou@ugent.be

Aäron van den Oord
aaron.vandenoord@ugent.be

Sander Dieleman
sander.dieleman@ugent.be

Mieke Van Herreweghe
mieke.vanherreweghe@ugent.be

Joni Dambre
joni.dambre@ugent.be

[1] Data Science Lab, ELIS, Ghent University, Ghent, Belgium

[2] Department of Linguistics, Ghent University, Ghent, Belgium

ing can be a viable strategy (Karpathy et al. 2014; Ng et al. 2015). As we will show in this paper, however, this method is of limited use for gesture recognition.

Apart from a collection of frames, a video can also be seen as a time series. Some of the most successful models for time series classification are recurrent neural networks (RNNs) with either standard cells or long short-term memory (LSTM) cells (Hochreiter and Schmidhuber 1997). Their ability to learn dynamic temporal dependencies has allowed researchers to achieve breakthrough results in e.g. speech recognition (Graves et al. 2013), machine translation (Sutskever et al. 2014) and image captioning (Vinyals et al. 2015). Before feeding video to recurrent models, we need to incorporate some form of spatial or spatiotemporal feature extraction. This motivates the concept of combining CNNs with RNNs. CNNs have unparalleled spatial (and spatiotemporal with added temporal convolutions) feature extraction capabilities, while adding recurrence ensures the modeling of feature evolution over time.

For general video classification datasets like UCF-101 (Soomro et al. 2012), Sports-1M (Karpathy et al. 2014) or HMDB-51 (Kuehne et al. 2011), the temporal aspect is of less importance compared to a gesture recognition dataset. For example, the appearance of a violin almost certainly suggests the target class is "playing violin", as no other class involves a violin. The model has no need to capture motion information for this particular example. That being said, there are some categories where modeling motion in some way or another is always beneficial. In the case of gesture recognition, however, motion plays a more critical role. Many gestures are not only defined by their spatial hand and/or arm placement, but also by their motion pattern.

In this work, we explore a variety of end-to-end trainable deep networks for video classification applied to frame-wise gesture recognition with the Montalbano dataset that was introduced in the ChaLearn LAP 2014 Challenge (Escalera et al. 2014). We study two ways of capturing the temporal structure of these videos. The first method involves temporal convolutions to enable the learning of motion features. The second method introduces recurrence to our networks, which allows the modeling of temporal dynamics, which plays an essential role in gesture recognition.

2 Related Work

An extensive evaluation of CNNs on general video classification is provided by Karpathy et al. (2014) using the Sports-1M dataset. They compare different frame fusion methods to a baseline single-frame architecture and conclude that their best fusion strategy only modestly improves the accuracy of the baseline. Their work is extended by Ng et al. (2015), who show that LSTMs achieve no improvements

over a temporal feature pooling scheme on the UCF-101 dataset for human action classification and only marginal improvements on the Sports-1M dataset. For this reason, the single-frame and the temporal pooling architectures are important baseline models.

Another way to capture motion is to convert a video stream to a dense optical flow. This is a way to represent motion spatially by estimating displacement vectors of each pixel. It is a core component in the two-stream architecture described by Simonyan and Zisserman (2014) and is used for human pose estimation (Jain et al. 2014), for global video descriptor learning (Ng et al. 2015) and for video captioning (Venugopalan et al. 2015). A disadvantage of this technique is the greater computational preprocessing complexity. However, we show that our models implicitly learn to infer motion features without the need for optical flow calculations.

Neverova et al. (2014) present an extended overview of their winning solution for the ChaLearn LAP 2014 gesture recognition challenge and achieve a state-of-the-art score on the Montalbano dataset. They propose a multi-modal 'Mod-Drop' network operating at three temporal scales and use an ensemble method to merge the features at different scales. They also developed a new training strategy, ModDrop, that makes the network's predictions robust to missing or corrupted channels.

Most of the constituent parts in our architectures have been used before in other work for different purposes. Learning motion features with three-dimensional convolution layers has been studied by Ji et al. (2013) and Taylor et al. (2010) to classify short clips of human actions on the KTH dataset. Baccouche et al. (2011) proposed including a two-step scheme to model the temporal evolution of learned features with an LSTM. Finally, the combination of a CNN with an RNN has been used for speech recognition (Hannun et al. 2014), image captioning (Vinyals et al. 2015) and video narration (Donahue et al. 2015).

3 Architectures

In this section, we briefly describe the different architectures we investigate for gesture recognition in video. An overview of the models is depicted in Fig. 1. Note that we pay close attention to the comparability of the network structures. The number of units in the fully connected layers and the number of cells in the recurrent models are optimized based on validation results for each network individually. All other hyper-parameters mentioned in this section and in Sect. 4.2 are optimized for the temporal pooling architecture. As a result, improvements over our baseline models are caused by architectural differences rather than better optimization, other hyper-parameters or preprocessing.
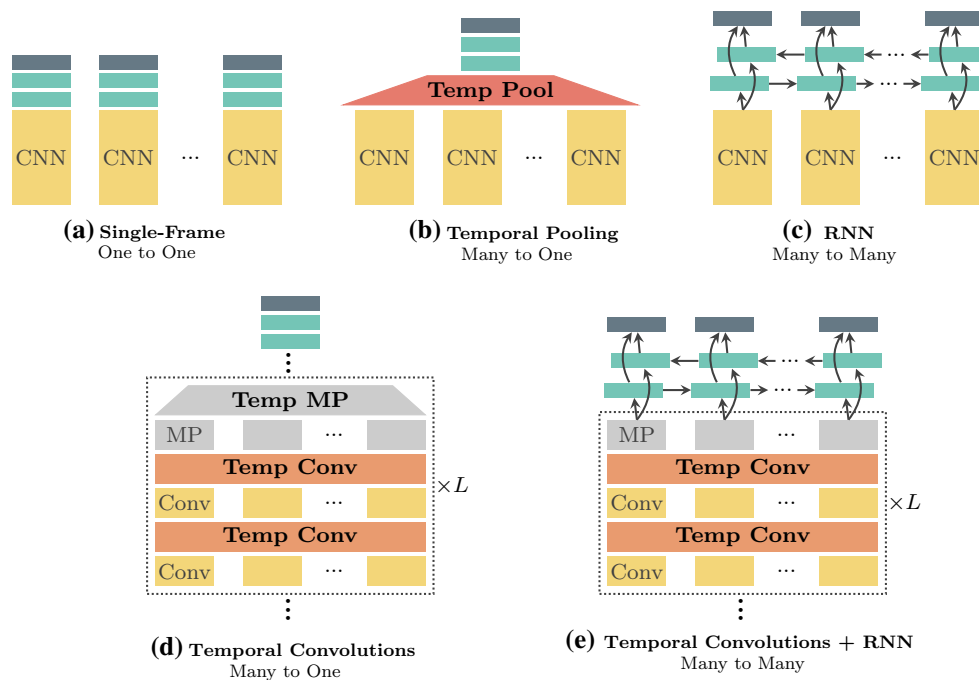
Fig. 1 Overview. **a** Single-frame CNN architecture. **b** Temporal feature pooling network (max- or mean-pooling), spanning multiple video frames. **c** Model with bidirectional recurrence. **d** Adding temporal convolutions and three-dimensional max-pooling (MP refers to max-pooling). **e** Architecture with added temporal convolutions and bidirectional recurrence

3.1 Baseline Models

3.1.1 Single-Frame

The single-frame architecture (Fig. 1a) worked well for general video classification (Karpathy et al. 2014), but is not a very fitting solution for our frame-wise gesture recognition setting. Nevertheless, this will give us an indication on how much static images contribute to the recognition. It has 3×3 convolution kernels in every layer. Two convolutional layers are stacked before performing max-pooling on non-overlapping 2×2 spatial regions. The shorthand notation of the full architecture is as follows: $C(16)–C(16)–P–C(32)–C(32)–P–C(64)–C(64)–P–C(128)–C(128)–P–D(2048)–D(2048)–S$, where $C(n_c)$ denotes a convolutional layer with n_c feature maps, P a max-pooling layer, $D(n_d)$ a fully connected layer with n_d units and S a softmax classifier. We deploy leaky rectified linear units (leaky ReLUs) in every layer. Their activation function is defined as $a : x \mapsto \max(\alpha x, x)$, where $\alpha = 0.3$. Leaky ReLUs seemed to work better than conventional ReLUs and showed promising results in other work (Maas et al. 2013; Graham 2014; Dieleman et al. 2015; Xu et al. 2015).

3.2 Temporal Feature Pooling

The second baseline model exploits a temporal feature pooling strategy. As suggested by Ng et al. (2015), we position the temporal pooling layer right before the first fully connected layer as illustrated in Fig. 1b. This layer performs either mean-pooling or max-pooling across all video frames. The structure of the CNN-component is identical to the single-frame model. This network is able to collect all the spatial features in a given time window. However, the order of the temporal events is lost due to the nature of pooling across frames.

3.3 Bidirectional Recurrent Models

The core idea of RNNs is to create internal memory to learn the temporal dynamics in sequential data. An issue (in our case) with conventional recurrent networks is that their states are built up from *previous* time steps. A gesture, however, generally becomes recognizable only after a few time steps, while the frame-wise nature of the problem requires predictions from the very first frame. This is why we use *bidirectional* recurrence, which enables us to process sequences in both temporal directions.

Describing the proposed model (Fig. 1c) formally, we start with the CNN (identical to the single-frame model) transforming an input frame x_t to a more compact vector representation v_t:

$$v_t = \text{CNN}(x_t). \tag{1}$$

A bidirectional RNN computes two hidden sequences: the forward hidden sequence $h^{(f)}$ and the backward hidden sequence $h^{(b)}$:

$$h_t^{(f)} = \mathcal{H}_f\left(v_t, h_{t-1}^{(f)}\right) \quad \text{and} \tag{2}$$

$$h_t^{(b)} = \mathcal{H}_b\left(v_t, h_{t+1}^{(b)}\right), \tag{3}$$

where $\mathcal{H}$ represents a recurrent layer and depends on the type of memory cell. There are two different cell types in widespread use: standard cells and LSTM cells (Hochreiter and Schmidhuber 1997) [we use the modern LSTM cell structure with peephole connections (Gers et al. 2003)]. Both cell types will be compared in this work.

Standard cells weight the input vector v_t with trainable parameters W_{vh} and summates with the previous hidden units h_{t-1}, weighted by W_{vh}, and a bias b_h. Standard cells are defined by

$$h_t = a(W_{vh}v_t + W_{hh}h_{t-1} + b_h), \tag{4}$$

where W_{vh}, W_{hh} and b_h are trainable parameters and a is the same leaky rectified linear nonlinearity as used in the CNN.

LSTMs cells are more complex, but their structure allows them to hold memory for much longer, hence the name. This enables them to capture long-range temporal dependencies. The cells can be described as follows:

$$i_t = \sigma(W_{vi}v_t + W_{hi}h_{t-1} + w_{ci} \odot c_{t-1} + b_i), \tag{5}$$

$$f_t = \sigma(W_{vf}v_t + W_{hf}h_{t-1} + w_{cf} \odot c_{t-1} + b_f), \tag{6}$$

$$o_t = \sigma(W_{vo}v_t + W_{ho}h_{t-1} + w_{co} \odot c_{t-1} + b_o), \tag{7}$$

$$g_t = \tanh(W_{vg}v_t + W_{hg}h_{t-1} + b_g), \tag{8}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t, \tag{9}$$

$$h_t = o_t \odot \tanh(c_t), \tag{10}$$

where $\odot$ denotes the point-wise multiplication of two vectors and all parameters referred by $W.$, $w.$ or $b.$ are trainable.

Finally, the output predictions y_t are computed with a softmax classifier which takes the sum of the forward and backward hidden states as input:

$$y_t = \text{softmax}\left(W_y(h_t^{(f)} + h_t^{(b)}) + b_y\right). \tag{11}$$

3.4 Adding Temporal Convolutions

Our final set of architectures extends the CNN layers with temporal convolutions (convolutions over time). This enables the extraction of hierarchies of motion features and thus the capturing of temporal information from the first layer, instead of depending on higher layers to form spatiotemporal features. Performing three-dimensional convolutions is one approach to achieve this. However, this leads to a significant increase in the number of parameters in every layer, making this method more prone to overfitting. Therefore, we decide to factorize this operation into two-dimensional spatial convolutions and one-dimensional temporal convolutions. This leads to fewer parameters and optionally more nonlinearity if one decides to activate both operations. We opt to not include a bias or another nonlinearity in the spatial convolution step to maintain the comparability between architectures.

First, we compute spatial feature maps s_t for every frame x_t. A pixel at position (i, j) of the k-th feature map is determined as follows:

$$s_{tij}^{(k)} = \sum_{n=1}^{N} \left(W_{\text{spat}}^{(kn)} * x_t^{(n)}\right)_{ij}, \tag{12}$$

where N is the number of input channels and W_{spat} are trainable parameters. Finally, we convolve across the time dimension for every position (i, j), add the bias $b^{(k)}$ and apply the activation function a:

$$v_{tij}^{(k)} = a\left(b^{(k)} + \sum_{m=1}^{M} \left(W_{\text{temp}}^{(km)} * s_{ij}^{(m)}\right)_t\right), \tag{13}$$

where the variables W_{temp} and b are trainable parameters and M is the number of spatial feature maps.

Two different architectures are proposed using this new layer. In the first model (Fig. 1d), we replace the convolutional layers of the single-frame CNN with the spatiotemporal layer defined above. Furthermore, we apply three-dimensional max-pooling to reduce spatial as well as temporal dimensions while introducing slight translational invariance in time. Note that this architecture implies a sliding window approach for frame-wise classification, which is computationally intensive. In the second model, illustrated in Fig. 1e, the time dimensionality is retained throughout the network. That means we only carry out spatial max-pooling. To this end, we are able to stack a bidirectional RNN with LSTM cells, responding to high-level temporal dependencies. It also incidentally resolves the need for a sliding window approach to implement frame-wise video classification.

4 Experiments

4.1 Montalbano Gesture Recognition Dataset

The ChaLearn Looking At People (LAP) 2014 Challenge (Escalera et al. 2014) consists of three tracks: human pose recovery, human action/interaction recognition and gesture recognition. The dataset accompanying the gesture recognition challenge, called the Montalbano dataset, will be used throughout this work. The dataset is multi-modal, because the gestures are captured with a Microsoft Kinect that has a depth sensor. In all sequences, a single user is recorded in front of the camera, performing natural communicative Italian gestures. Each data file contains an RGB-D (where "D" stands for depth) image sequence and a skeletal pose stream provided by the Microsoft Kinect API. The gesture vocabulary contains 20 Italian cultural/anthropological signs. The gestures are not segmented, which means that sequences typically contain several gestures. Gesture performances appear randomly within the sequence without a prearranged rest pose. Moreover, several unannotated out-of-vocabulary gestures are present.

It is the largest publicly available gesture dataset of its kind. There are 1, 720, 800 labeled frames across 13, 858 video fragments of about 1 to 2 minutes sampled at 20Hz with a resolution of 640×480. The gestures are performed by 27 different individuals under diverse conditions; these include varying clothes, positions, backgrounds and lighting. The training set contains 11, 116 gestures and the test set contains 2742. The class imbalance is negligible. The starting and ending frames for each gesture are annotated as well as the gesture class label.

To speed up the training, we crop part of the images containing the user and rescale them to 64 by 64 pixels using the skeleton information (other than that, we do not use any pose data). However, we show in Sect. 4.4 that we even achieve good results when we do not crop the images and leave out depth information. Figure 2 illustrates the cropping of an input image. The head and the hip positions are tracked by the Microsoft Kinect API. We found these tracking points to be consistent and stable. Based on these two points we crop a square region of interest.

Lastly, we experiment with feeding the networks with dense optical flow channels. These inputs are calculated with the techniques used in Farnebäck (2003).

4.2 End-To-End Training

We train our models from scratch in an end-to-end fashion, backpropagating through time (BTT) for our recurrent architectures. The network parameters are optimized by minimizing the cross-entropy loss function using mini-batch gradient descent with the *Adam* update rule (Kingma and Ba
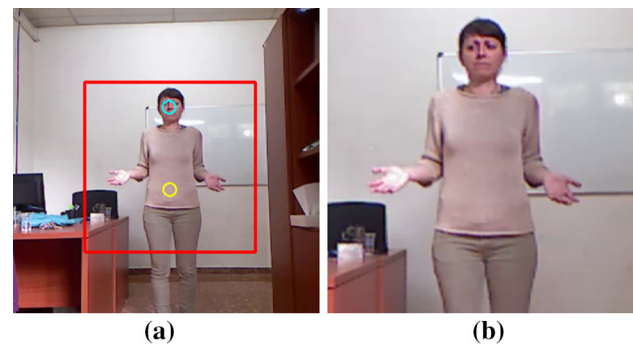

Fig. 2 Preprocessing. The *blue* and *yellow circle* indicate the head and hip position respectively. This pose information is provided by the Microsoft Kinect API. The *red square* stipulates the cropped region (Color figure online)

2015). We found that Adam works great in practice, especially when experimenting with very different layer types in the same model. All our models are trained the same way with early stopping, a mini-batch size of 32, a learning rate of 10^{-3} and an exponential learning rate decay. Before training, we initialize the weights with a random orthogonal initialization method (Saxe et al. 2013).

4.2.1 Recurrent Networks

As described in Sect. 4.1, the video files in the Montalbano dataset contain approximately 1–2 minutes of footage, consisting of multiple gestures. Recurrent models are trained on random fragments of 64 frames and produce 64 predictions, one for every frame. To summarize, a data sample has 4 channels (RGB-D), 64 frames each, with a resolution of 64 by 64 pixels; or in shorthand notation: $4@64 \times 64 \times 64$. We optimized the number of cells for each model based on validation results. For LSTM cells, we only saw a small improvement between 512 and 1024 units, so we settled at 512. For RNNs with standard cells, we used 2048 units. The location of gestures within the long sequences is not given. A gesture is generally about 20–50 frames long. If a small fraction of a gesture is located at the beginning or the end of the 64 considered frames, the model does not have enough information to label these frames correctly. That is why we allow a buildup in both forward and backward direction for evaluation; we feed 64 frames into the RNN and keep the middle 32 for evaluation.

4.2.2 Non-Recurrent Networks

The single-frame CNN is trained frame by frame and all other non-recurrent networks are trained with the number of frames optimized for their specific architecture. The best number of frames to mean-pool across is 32, determined by validation scores with tested values in [8, 16, 32, 64]. In the

Table 1 A comparison of the results for our different architectures on the Montalbano gesture recognition dataset (RGB-D cropped images, without optical flow)

Architecture	Jaccard index	Precision (%)	Recall (%)	Error rate (%)*
Single-frame CNN (Fig. 1a)	0.465	67.86	57.57	20.68
Temp max-pooling (Fig. 1b)	0.748	85.03	82.92	8.66
Temp mean-pooling (Fig. 1b)	0.775	85.93	85.80	8.55
Temp Conv (Fig. 1d)	0.842	89.36	90.15	4.67
RNN, standard cells (Fig. 1c)	0.885	92.77	93.56	3.58
RNN, LSTM cells (Fig. 1c)	0.888	93.75	93.28	3.55
Temp Conv + RNN, standard (Fig. 1e)	0.900	93.76	94.47	2.82
Temp Conv + RNN, LSTM (Fig. 1e)	0.906	94.49	94.57	2.77

The Jaccard index indicates the mean overlap between the binary predictions and the binary ground truth across gesture categories. We also compute precision and recall scores for each gesture class and report the mean score across classes
* The error rate is based on majority voted frame-wise predictions from *isolated* gesture fragments

case of max-pooling, we find that pooling over 16 frames gives better outcomes. Also, pretraining the CNNs frame-by-frame and fine-tuning with temporal max-pooling gave slightly improved results. We observed no improvements, however, using this technique with temporal mean-pooling. The architecture with added temporal convolutions and three-dimensional max-pooling showed optimal results by considering 32 surrounding frames. The targets for all the non-recurrent networks are the labels associated with the centermost frame of the input video fragment. We evaluate these models using a sliding window with single-frame steps.

4.2.3 Regularization and Data-Augmentation

We employed many different methods to regularize the deep networks. Data augmentation has a significant impact on generalization. For all our trained models, we used the same augmentation parameters: $[-5, 5]$ pixel translations in vertical direction and $[-10, 10]$ horizontal, $[-2, 2]$ rotation degrees, $[-2, 2]$ shearing degrees, $[\frac{1}{1.1}, 1.1]$ image scaling factors and $[\frac{1}{1.2}, 1.2]$ temporal scaling factors. From each of these intervals, we sample a random value for each video fragment and apply the transformations online using the CPU. Dropout with $p = 0.5$ is used on the inputs of every fully connected layer. Furthermore, using leaky ReLUs instead of conventional ReLUs and factorizing three-dimensional convolutions into spatial and temporal convolutions also reduce overfitting.

4.3 Results

We follow the ChaLearn LAP 2014 Challenge score to measure the performance of our architectures. This way, we can compare with previous work on the Montalbano dataset. The competition score is based on the Jaccard index, which is defined as follows:

$$J_{s,n} = \frac{|A_{s,n} \cap B_{s,n}|}{|A_{s,n} \cup B_{s,n}|}. \tag{14}$$

The binary ground truth for gesture category n in sequence s is denoted as the binary vector $A_{s,n}$, whereas $B_{s,n}$ denotes the binary predictions. The Jaccard index $J_{s,n}$ can be seen as the overlap rate between $A_{s,n}$ and $B_{s,n}$. To compute the final score, the mean Jaccard index among all categories and sequences is computed:

$$J_{\text{avg}} = \frac{1}{NS} \sum_{s=1}^{S} \sum_{n=1}^{N} J_{s,n}, \tag{15}$$

where $N = 20$ is the number of categories and S the number of sequences in the test set.

An overview of the results for our different architectures is shown in Table 1. The predictions of the single-frame baseline achieve a Jaccard index below 0.5. This is to be expected as no motion features are extracted. We observe a significant improvement with temporal feature pooling (a Jaccard index of 0.775 vs. 0.465). Furthermore, mean-pooling performs better than max-pooling. Adding temporal convolutions and three-dimensional max-pooling improves the Jaccard index to 0.842.

The four last entries in Table 1 use recurrent networks. Surprisingly, the RNNs are only acting on high-level spatial features, yet are surpassing a CNN learning hierarchies of motion features (a Jaccard index of 0.842 vs. 0.888). Finally, combining the temporal convolution architecture with an RNN improves the score even more (LSTM: 0.906, standard: 0.900). This deep network not only learns multi-level spatiotemporal features, but is capable of modeling temporal dynamics within them.

The difference in performance for the two types of cells is very small and they can be considered equally capable for this type of problem where temporal dependencies are not

Table 2 Montalbano gesture recognition dataset results compared to previous work

Model	Crop	Depth	Optical flow	Pose	Jaccard index
Chang (2014) (MRF, KNN, PCA, HoG)	**Yes**	No	No	**Yes**	0.827
Monnier et al. (2014) (AdaBoost, HoG)	**Yes**	**Yes**	No	**Yes**	0.834
Neverova et al. (2014) (Multi-Scale DNN)	**Yes**	**Yes**	No	No	0.836
Neverova et al. (2014) (Multi-Scale DNN)	**Yes**	**Yes**	No	**Yes**	0.870
Temp Conv + LSTM	No	No	No	No	0.842
	Yes	No	No	No	0.876
	Yes	**Yes**	No	No	**0.906**
	Yes	**Yes**	**Yes**	No	0.895

Crop the cropping of specific areas in the video using the skeletal information, *Depth* the usage of depth-maps, *Optical flow* the inclusion of optical flow channels, *Pose* the usage of the skeletal stream as features. Note that even when we do not use depth images, we still achieve better results
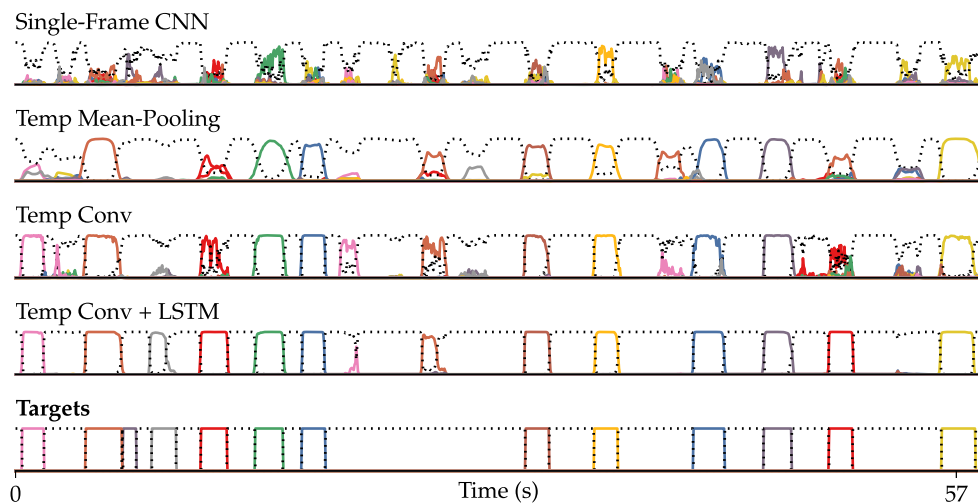

Fig. 3 The output probabilities are shown for a sequence fragment in the test set. The *dashed line* represents silences. The non-recurrent models make more mistakes and have difficulties making hard decisions to too long-ranged. However, our training phase is considerably more stable and roughly twice as fast with LSTM cells. Models with standard cells require tuning of hyperparameters to even have a converging setup, while we never encounter a diverged experiment with LSTM networks.

In Table 2, we compare our results with previous work. Our best model outperforms the method of Neverova et al. (2014) when we only consider RGB-D pixels as input features (0.906 vs. 0.836). When we remove depth information and perform no preprocessing other than rescaling the images, we still achieve better results (0.842). The previous best performing score (0.870), where the skeletal stream is used as input features, is outperformed by our model without pose information (0.906) nor depth images (0.876). We observe no improvement with the use of optical flow for this task. This suggests that the models are able to capture motion from the RGB data (see further and Fig. 4) and that the optical flow does not add useful information in our case.

where the gesture starts or ends and are unable to smooth out predictions in time. Adding recurrence enables deep networks to learn the behavior of the manual annotators with great accuracy

To illustrate the differences in output predictions of the different architectures, we show them for a randomly selected sequence in Fig. 3. We see that the single-frame CNN has trouble classifying the gestures, while the temporal pooling is significantly more accurate. However, the latter still has difficulties with boundaries. Adding temporal convolutions shows improved results, but the output contains more jagged predictions. This seems to disappear by introducing recurrence. The output of the bidirectional RNN matches the target labels strikingly well.

In Fig. 4, we show that adding temporal convolutions enables neural networks to capture motion information. When the user is standing still, the units of the feature map are inactive, while the feature map from the network without temporal convolutions has a lot of active units. When the user is moving, the feature map shows strong activations at the movement locations. This suggests that the model has learned to extract motion features.

While standing still

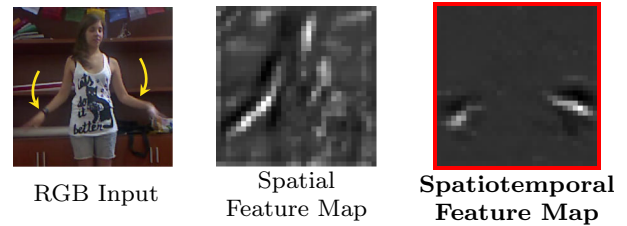RGB Input · Spatial Feature Map · **Spatiotemporal Feature Map**
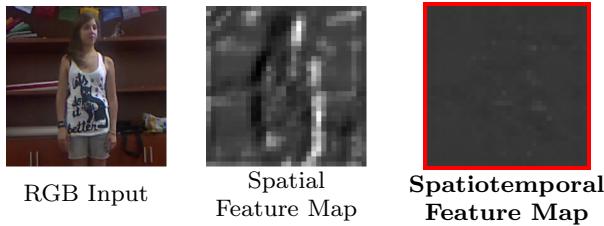
Fig. 4 Motion Features. This figure illustrates the effect of integrating temporal convolutions. The depicted spatial feature map is the most active 4-layer-deep feature map, extracted from an architecture without temporal convolutions. The spatiotemporal feature map is extracted

While moving

RGB Input · Spatial Feature Map · **Spatiotemporal Feature Map**

from a model with temporal convolutions. The strong activations in the spatiotemporal feature maps while moving indicate learned motion features
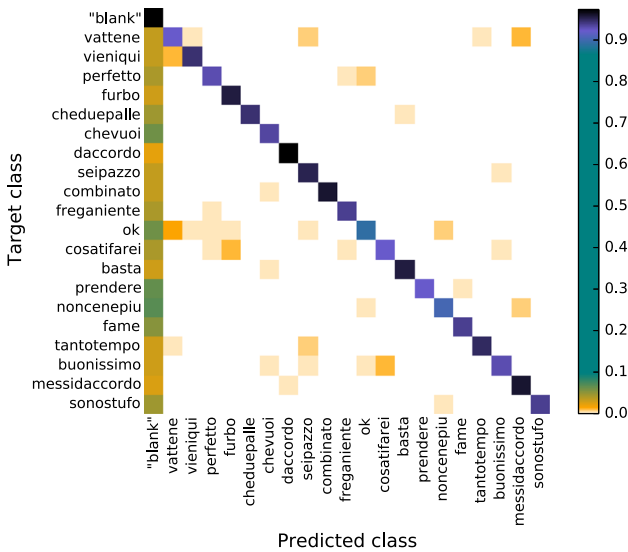

Fig. 5 The confusion matrix for the model with temporal convolutions and LSTM cells, evaluated on the test set

4.4 Failure Cases

The confusion matrix in Fig. 5 visualizes the performance of our best model (temporal convolutions + recurrence) for each gesture. The diagonal values clearly all have high values, which indicates a highly accurate classification. The most occurring error is the prediction of a silence, while the target is a particular gesture. This is due to the fact that the most common class is a silence. This imbalance causes the model to bet on a silence when the input is too confusing.

There are very few confusions *between* gestures. We depict the most common confusions in Fig. 6. The gestures "Vieni qui" (*Eng*: come here) and "Vattene"(*Eng*: begone) both raise one arm and move their hand towards or away from the user. When it is not clear in which direction the hand moves, the models confuses both gestures. The "Frega niente" and "Perfetto" gestures both start from near the mouth
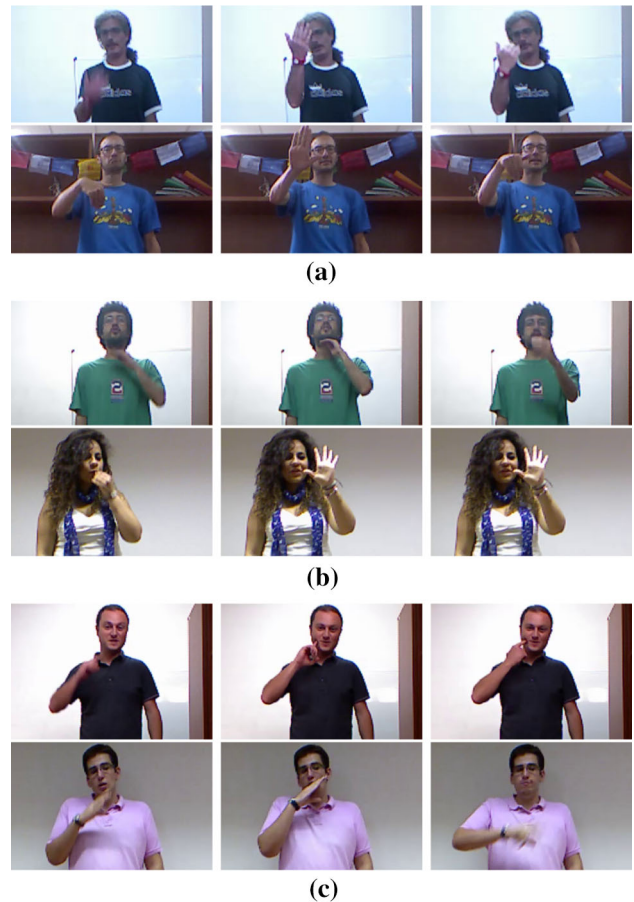

Fig. 6 Three examples to illustrate confusion between similar gestures. **a** *Top* "Vieni qui". *Bottom* "Vattene". **b** *Top* "Frega niente". *Bottom* "Perfetto". **c** *Top* "Buonissimo". *Bottom* "Cosa ti farei"

and move away, while "Buonissimo" and "Cosa ti farei" stay near the mouth for a while.

In Fig. 7, we show the video samples where the Jaccard index is the lowest. There is one outlier sample (Fig. 7a) where the recognition fails almost completely. The user is in the corner of the screen and the gestures are sometimes

Fig. 7 The four lowest scoring test set video samples are depicted. This is evaluated with the best performing model. **a** The user is almost off camera. Jaccard index = 0.378. **b** The video sample consists of noise gestures. Jaccard index = 0.652. **c** The user posture is not straight. Jaccard index = 0.698. **d** This background consistently gives low scores. Jaccard index = 0.711

performed off screen. A second form of failure involves noise (or out-of-vocabulary) gestures, e.g. there are two noise gestures in the fragment in Fig. 3. These should be classified silences, since the Montalbano dataset does not provide annotations for them. However, as they are fairly seldom, they are sometimes confused for a gesture. The video sample in Fig. 7b is packed with noise gestures, which explains the poor performance. Another difficulty is the posture of a user. Most users keep their posture straight. This causes the neural networks to not be invariant of upper body movement as in Fig. 7c. Lastly, we observe that one particular background (Fig. 7d) consistently gives lower Jaccard index scores than others. Although it is difficult to determine the cause, we assume the reason for this is the poor lighting of the environment.

5 Conclusion and Future Work

We showed in this paper that adding bidirectional recurrence and temporal convolutions improves frame-wise gesture recognition in video significantly. We observed that RNNs responding to high-level spatial features perform much better than single-frame and temporal pooling architectures, without the need to take into account the temporal aspect in the lower layers of the network. However, adding temporal convolutions in all layers of the architecture has a notable impact on the performance, as they are able to learn hierarchies of motion features, unlike RNNs. Standard cells and LSTM cells appear to be equally strong for this problem. Furthermore, we observed that RNNs outperform non-recurrent networks and are able to predict the beginning and ending frames of gestures with great accuracy, whereas other models show uncertainty at these boundaries.

In the future, we would like to build upon this work for research in the domain of sign language recognition. This is even more challenging than gesture recognition. The vocabulary is larger, the differences in finger positions and hand movements are more subtle and signs are context dependent, as they are part of a language. Sign language is not related to written or spoken language, which complicates annotation and translation. Moreover, signers communicate simultaneously with facial, manual (both hands are separate communication channels) and body expressions. This means that sign language video cannot be translated the way speech recognition can transcribe audio to written sentences.

References

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A. (2011). Sequential deep learning for human action recognition. In A. Salah & B. Lepri (Eds.), *Human behavior understanding* (pp. 29–39). Berlin Heidelberg: Springer.

Chang, J. Y. (2014). Nonparametric gesture labeling from multi-modal data. *Computer vision-ECCV 2014 workshops* (pp. 503–517). Springer.

Dieleman, S., van den Oord, A., Korshunova, I., Burms, J., Degrave, J., Pigou, L., & Buteneers, P. (2015). Classifying plankton with deep neural networks. http://benanne.github.io/2015/03/17/plankton.html. Accessed 17 Mar 2015.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634.

Escalera, S., Bar, X., Gonzlez, J., Bautista, M.A., Madadi, M., Reyes, M., Ponce, V., Escalante, H.J., Shotton, J., & Guyon, I. (2014). Chalearn looking at people challenge 2014: Dataset and results. In: ECCV workshop.

Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In J. Bigun & T. Gustavsson (Eds.), *Scandinavian conference on image analysis* (pp. 363–370). Berlin Heidelberg: Springer.

Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2003). Learning precise timing with lstm recurrent networks. *The Journal of Machine Learning Research*, *3*, 115–143.

Graham, B. (2014). Spatially-sparse convolutional neural networks. arXiv:1409.6070 (preprint).

Graves, A., Mohamed, A.R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In: Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on, IEEE, pp. 6645–6649.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., & Coates, A., et al.

(2014). Deepspeech: Scaling up end-to-end speech recognition. arXiv:1412.5567 (preprint).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

Jain, A., Tompson, J., LeCun, Y., & Bregler, C. (2014). MoDeep: A deep learning framework using motion features for human pose estimation. *Computer Vision ACCV*, *2014*, 302–315.

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1), 221–231.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In: Computer vision and pattern recognition (CVPR), 2014 IEEE conference on, IEEE, pp. 1725–1732.

Kingma, D., & Ba, J. (2015). Adam: A method for stochastic optimization. ICLR 2015.

Krizhevsky, A., Sutskever, I., & Hinton, GE, (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 1097–1105). http://papers.nips.cc/paper/4824-imagenet-classification-withdeep-convolutional-neural-networks.pdf.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). Hmdb: A large video database for human motion recognition. In: Computer vision (ICCV), 2011 IEEE international conference on, IEEE, pp. 2556–2563.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Maas, A.L., Hannun, A.Y., & Ng, A.Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In: Proc. ICML, vol. 30.

Monnier, C., German, S., & Ost, A. (2014). A multi-scale boosted detector for efficient and robust gesture recognition. In: Computer vision-ECCV 2014 workshops (pp. 491–502). Springer

Neverova, N., Wolf, C., Taylor, G.W., & Nebout, F. (2014). ModDrop: Adaptive multi-modal gesture recognition. arXiv:1501.00102 (preprint).

Ng, J.Y.H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., & Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In: Computer vision and pattern recognition (CVPR), 2015 IEEE conference on, IEEE, pp. 4694–4702.

Saxe, A.M., McClelland, J.L., & Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv:1312.6120 (preprint).

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2013). Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv:1312.6229 (preprint).

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 568–576). http://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos.pdf.

Soomro, K., Zamir, A.R.,& Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402 (preprint).

Sutskever. I., Vinyals, O., & Le, Q.V. (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 3104–3112). http://papers.nips.cc/paper/5346-sequence-to-sequence-learningwith-neural-networks.pdf.

Taylor, G. W., Fergus, R., LeCun, Y., & Bregler, C. (2010). Convolutional learning of spatio-temporal features. In K. Daniilidis, P. Maragos, & N. Paragios (Eds.), *Computer vision-ECCV 2010* (pp. 140–153). Berlin Heidelberg: Springer.

Toshev, A., & Szegedy, C. (2014). DeepPose: Human pose estimation via deep neural networks. In: Computer vision and pattern recognition (CVPR), 2014 IEEE conference on, IEEE, pp. 1653–1660.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to sequence–video to text. arXiv:1505.00487 (preprint).

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156–3164.

Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical evaluation of rectified activations in convolutional network. In: ICML deep learning workshop.