Spatio-Temporal Convolutional Sparse Auto-Encoder for Sequence Classification

Moez Baccouche[1]
moez.baccouche@orange.com
Franck Mamalet[1]
franck.mamalet@orange.com
Christian Wolf[2]
christian.wolf@liris.cnrs.fr
Christophe Garcia[2]
christophe.garcia@liris.cnrs.fr
Atilla Baskurt[2]
atilla.baskurt@liris.cnrs.fr

[1] Orange Labs R&D
4 rue du Clos Courtel
F-35510, France

[2] Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR 5205
F-69621, France

We address in this paper the problem of task-independent video sequence classification. We aim to introduce a generic model which differ from the highly problem-dependent dominant methodology that relies on so-called *hand-crafted* features. We propose a learning-based model with two main steps: the first one aims to automatically learn spatio-temporal features instead of hand-crafting them. These learned features are *sparse-overcomplete*, i.e. their dimension is larger than the input one, but only a small number of components are non-zero. The second step consists in labelizing the entire video sequence considering the temporal evolution of the learned features. The first learning step is performed in an unsupervised way, and is based on spatio-temporal convolutional sparse auto-encoders (which will be introduced hereafter), and the second consists in a supervised classification using recurrent neural networks.

The feature learning process is based on an auto-encoder scheme: an encoder which builds a non-sparse code vector representing the spatio-temporal salient information contained in the input, and a decoder which learns to reconstruct the input from a sparse version of the obtained code (see Figure 1). The model takes as input small space-time patches in order to reduce the diversity of the content to be encoded, since the patterns are locally less variable than if the full frame was considered. The encoder is a convolutional neural network with $2D+t$ convolution kernels (each one having the same size than the input patch). The decoder consists in a set of output neurons fully connected to the sparse code layer. The sparsity is obtained using the *sparsifying logistic* proposed by Ranzato *et al*. [2], placed between the encoder and the decoder. This model is associated to a global objective function, which is the sum of two terms, representing respectively the encoder prediction and the decoder reconstruction mean square errors, and which is minimized during training.

In order to handle the spatial and temporal shift-invariance of the learned representations, a "best shift search" module is introduced before the auto-encoder (see Figure 1). The idea is to represent the spatio-temporal neighbourhood of a given input patch by a single "shifted" patch, which is the one minimizing the objective function, given the current set of parameters. To that aim, an additional hidden variable is introduced, the translation vector, on which the optimization is done.

To avoid encoding non-relevant patterns (e.g. colour and texture), the model is trained only with the patches containing significant spatio-temporal information (according to a motion-based selection critirea). This plays the same role as the saliency detectors in the case of the *hand-crafted* features.

Entire sequences are finally labeled with a particular recurrent neural network classifier, namely *Long Short-Term Memory Recurrent Neural Network* (LSTM) [1], in order to take benefits of its ability to use the temporal evolution of features for classification. The LSTM classifier takes as input a sequence of feature vectors, each one corresponding to the concatenated responses of the patches placed at the grid of possible locations in each frame.

Aiming at verifying the genericity of the proposed model, experiments were carried out on two different problems: human actions and facial expression recognition. For the first experiment, we used the standard KTH human actions dataset [3]. To our knowledge, our method obtains the best results among the methods using automatically learned features, both on the two versions of the KTH dataset (95.83% for KTH1 and 93.74% for KTH2). More generally, we obtain the second best result for KTH1, and the third for KTH2, even when compared with approaches relying on hand-crafted features designed for the KTH dataset.

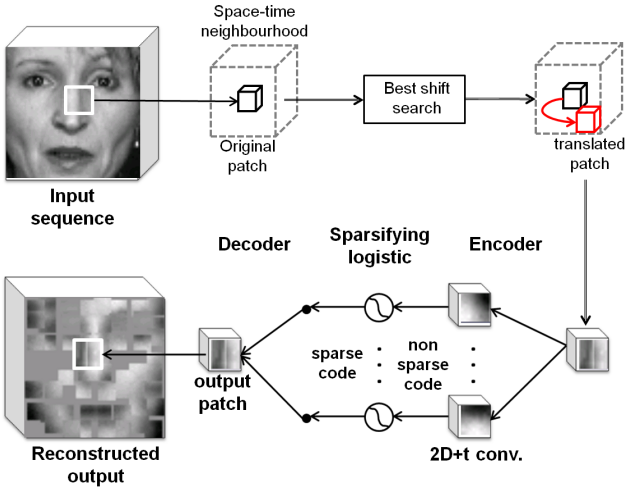

Figure 1: Overview of the proposed spatio-temporal convolutional sparse auto-encoder: Illustration on a sample from the GEMEP-FERA facial expressions dataset.

For facial expression recognition, we used the recent GEMEP-FERA dataset [4]. Obtained results are superior to the state of the art (87.57% for the overall classification rate), with a significative performance improvement particlularly for the person-independent configuration (with a recognition rate of 80.75%), which is a positive evidence of the high generalization of our approach, and eliminating the person-specific effect and capturing the facial expression salient information.

To conclude, we have proposed a neural model for video sequence classification, with a fully automated learning-based feature construction process. We have introduced the spatio-temporal convolutional sparse auto-encoder architecture, and its corresponding training procedure. We have also presented a novel approach for handling shift-invariance of the representation. Finally, we have shown how the temporal evolution of these features is used to classify the sequences, using a recurrent neural network model. Experimental results on two different problems confirms the high genericity of the model since it achieves the best results among related works. Future work will address scale invariance and applications to different problems.

[1] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[2] M.A. Ranzato, F.J. Huang, Y.L. Boureau, and Y. Lecun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[3] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *International Conference on Pattern Recognition*, volume 3, pages 32 – 36, 2004.

[4] M.F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *International Conference on Automatic Face & Gesture Recognition*, pages 921–926, 2011.