Human Activity Recognition from Audio through CNN

Matteo Rusconi

Università degli studi di Milano

***Abstract*—This work presents some experiments on the Human Activity Recognition from audio field, by employing Convolutional Neural Networks as learning algorithm. The subject of this study are home related everyday life activities like cooking and taking showers. Relevant features like mel-frequency cepstral coefficient and chromagrams are combined with statistical features in order to improve the generalization capability of the model. It is also shown a comparison between the performances obtained when considering features regarding shorter and longer audio tracks as training data. In conclusion, there is a general comparison with other recent researches in the same topic, showing that CNNs are effectively a valid solution to the underlying problem.**

I. Introduction

Human activity recognition (HAR) is a branch of machine learning which aims to build systems capable of recognizing activities being performed by human agents by observing data coming from one or more sensors. This discipline has numerous applications in many fields such as health care and assistance, smart surveillance, natural interaction, self driving cars, entertainment and more. In this paper, HAR will be applied to audio data regarding home activities, like cooking and showering. Given that nowadays microphones are integrated in lots of mobile devices (smartphones, tablets..) HAR from audio is gaining high potential in becoming strongly integrated in everyday life, leading to the production of enormous quantities of data to be processed and used by many parties. This paper will be analyzing and discussing the application of modern deep learning techniques on these HAR from audio tasks, focusing on the Convolutional Neural Networks (CNN) realm.

Regarding the state of the art, there are other approaches based on the methodology and theories of social psychology, collecting audio data that can be tagged with Essential Social Interaction Predicates (ESIP) [2]. Binary silhouettes have also been used to represent the different human activities. Uddin et al. proposed a system based on Generalized Discriminant Analysis and Hidden Markov Models for training and recognition [3]. Also, principal component analysis (PCA) [4], [5], [6] and independent component analysis (ICA) [7] have also been applied to this purpose.

In this work CNNs are trained and evaluated on the combination of both statistical and audio features, exploring the difference of two main approaches; short and long audio tracks. In fact, one of the principal topics of this experiment is to discuss whether HAR with CNNs works better with shorter (5-10s) or longer (50-60s) audio clips.

II. Methodologies

This section will be presenting the methodologies utilized in this project, discussing data-related topics and showcasing the main properties of the employed machine learning model.

A. *Datasets description*

As stated in the previous section, HAR is going to be explored on the home activities class of sounds. The dataset with audio tracks is made available by the AmiDaMi research group page and it contains a total of eight different activity noises, which are: brewing coffee, cooking, using the microwave oven, taking a shower, dish washing, hand washing, teeth brushing, and no activity noises. The recordings are obtained using smartphone microphones, in particular with iPhone, HTC, LG and Lanix devices, which use different audio sample rates going from 8,000 Hz to 44,100 Hz with both mono and stereo recordings. The tracks have been trimmed into 10s and 60s audio clips, leading to the creation of two different datasets to be used separately, composed by 1226 and 240 audio clips respectively. At last, the training part of these datasets is class balanced with the oversampling technique in order to improve performances of the models.

B. *Feature Extraction and Normalization*

Starting from the audio clips, several features have been extracted and aggregated to form the model training datasets. The first and most important extracted feature are Mel-frequency cepstral coefficients (MFCC) which have been shown in [1], [8] to be of importance in solving similar problems. In particular, 12 MFCC have been extracted from each audio clip. Next, chromagrams have been generated and added to the features set. It has already been shown that chromagrams also provide good performance in general on audio classification tasks, and together with MFCCs are expected to be good in discriminating between the different activities. At last, as discussed in [1], [10] features that statistically describe sound waves are proven to be useful on similar machine learning problems; in particular the following set of features is taken into account (extracted from the audio sample arrays):

- mean of the integer array
- median of the integer array
- standard deviation of the integer array
- variance of the integer array
- kurtosis of the probability distribution of the integer array
- skewness of the probability distribution of the integer array

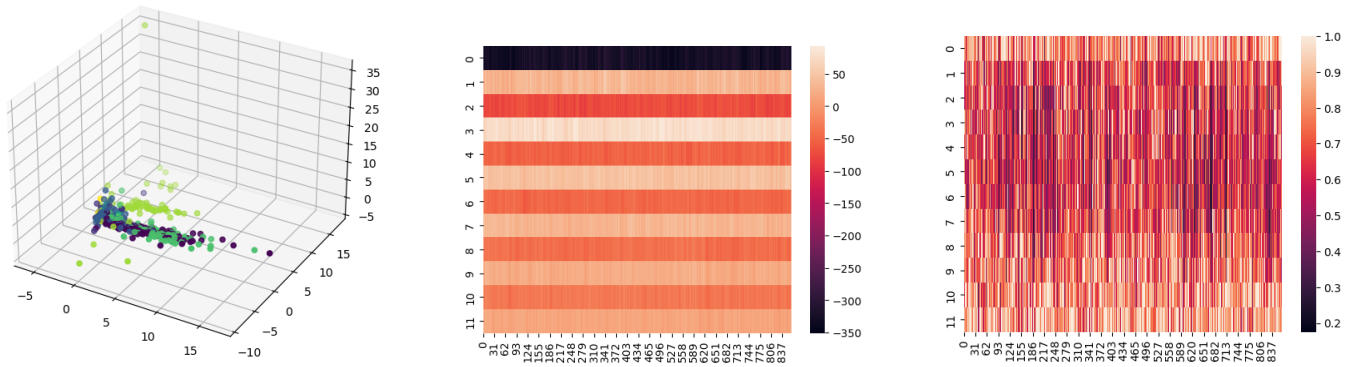

Fig. 1: Visualization of the feature space: 3D PCA projection of the statistical features with three principal components (left), MFCC and chromagram extracted features regarding the ten seconds clips dataset (middle, right)

- coefficient of variation (CV) of the probability distribution of the integer array,
- inverse CV
- 1st, 5th, 25th, 50th, 75th, 95th, and 99th percentile of the probability distribution of the integer array
- mean of the integer array after trimming the bottom and top 5% elements

Figure 1 gives a brief visualization of the feature space obtained in the discussed manner.

Finally, each feature has been normalized by subtracting mean and dividing by standard deviation, thus obtaining the datasets ready to be processed by the machine learning algorithm.

C. *Model Definition*

The main purpose of this project is to experiment how CNNs perform on audio data in terms of prediction power and generalization. Convolutional Neural Network are already proven to be among the bests regarding image classification, but it is interesting to see if their good capability of recognizing shapes and patterns could also be applied on audio extracted images like chormagrams and MFCC shown in Figure 1. Figure 2 gives an overview of the structure regarding the CNN designed for the short clips dataset. Each training input is composed by two 12x862 matrices (MFCC and chromagram) which are fed into the convolutions, and a 16 elements array (statistical features) which is directly appended to the dense input layer. The output is a 8 element vector (one neuron per class) generated by a softmax activation function, hence giving each class a probability of being the correct one. In order to deal with the long clips dataset the structure had to be tweaked to be able to handle bigger MFFCs and chromagrams, which grow up in size up to 12x5168 matrices, becoming six times bigger than the latter. In particular, a much deeper network was crucial to achieve a good dimensionality reduction going from matrices with tens of thousands of elements to flattened representations of a couple hundred elements, as shown in Figure 3. The two CNNs reach respectively a total of 1,813,968 and 2,627,068 trainable parameters.

III. Experiments and Results

The datasets are split into train and test with an 80%-20% fashion. CNNs are trained over the train dataset with a total of 200 epochs, keeping the bests as resulting models. Performances are evaluated on the test set as depicted in Table I.

model	accuracy	F1 score
CNN on short clips dataset	0.873	0.788
CNN on long clips dataset	0.854	0.716

TABLE I: Evaluation of the models

Due to the nature of the dataset, the test parts which were employed for the evaluation are noticeably class unbalanced, hence the F1 score metric comes in handy giving a more realistic measure for the quality of the trained models. By comparing the performances of the two models it seems like using shorter audio clips leads to slightly better results. Figure 4 and 5 represent the confusion matrices of the evaluation on the test sets, respectively for the short and long clips datasets. It can be noticed that the no activity sounds were the most difficult to correctly classify by the models as they were often being confused with other sounds like washing hands or cooking. Also, it seems like the deeper CNN does a better job in recognizing the taking a bath and cooking activities.
Table II shows a comparison against different approaches explained in Galvàn-Tejada et al. [1] which uses Random Forests to address the HAR problem on the same dataset, and Kabir et al. [9] which relies on Hidden Markov Models.

IV. Conclusion

The intent of this paper is to give a simple but meaningful example of how deep Convolutional Neural Networks can be employed on audio classification, with the utilization of powerful features like MFCCs and chromagrams. The comparison presented in Table II shows that this approach can be perfectly comparable with other modern approaches, if not even slightly better. In a possible future extension of this work it could be interesting to see how a lookalike model would perform on a bigger and more balanced dataset.

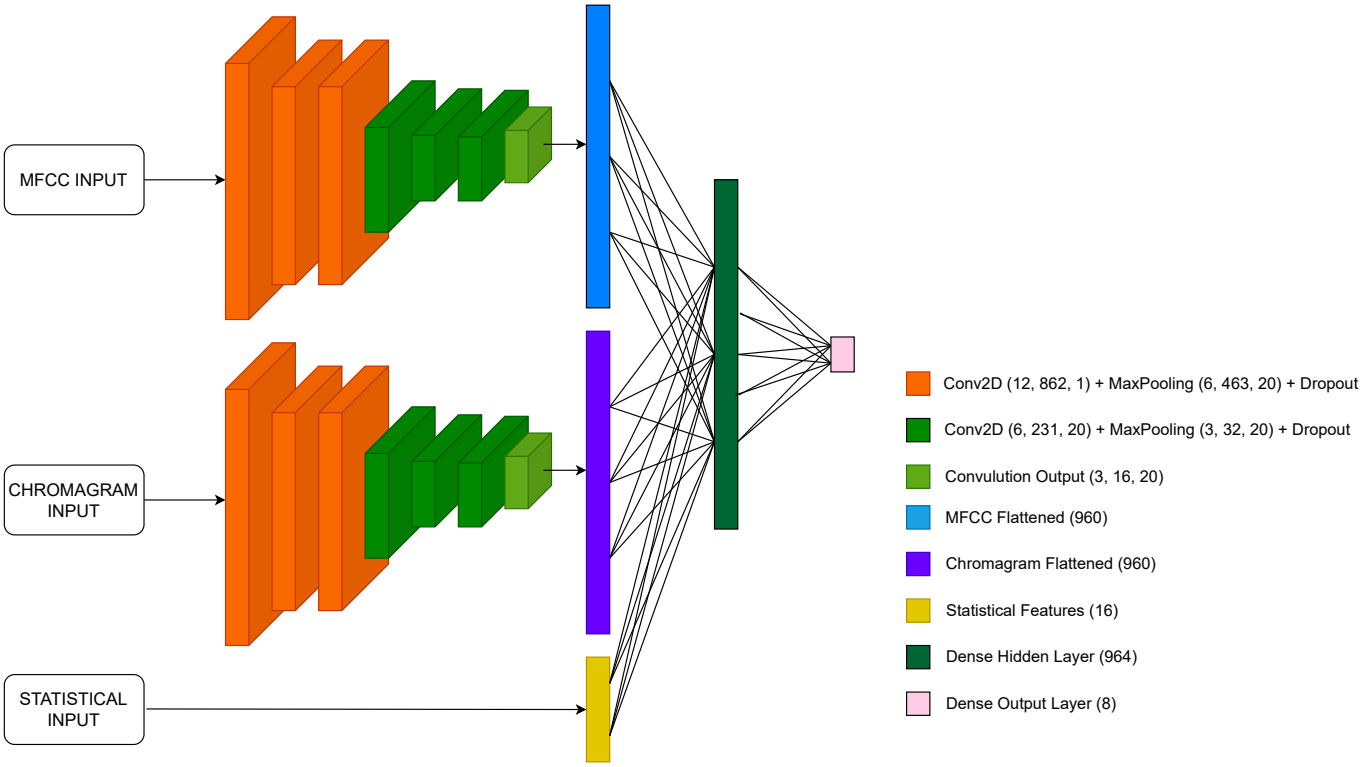

Fig. 2: CNN structure designed for the ten seconds clips dataset

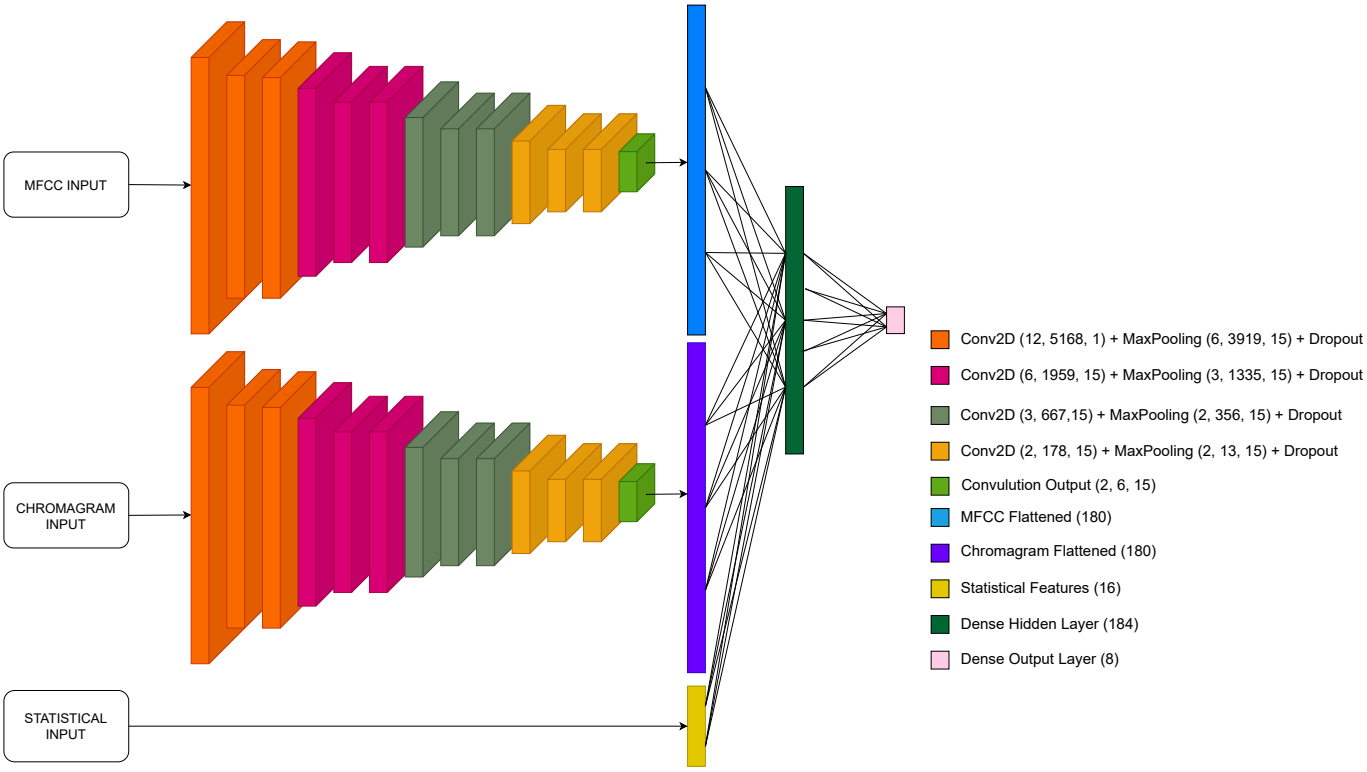

Fig. 3: CNN structure designed for the sixty seconds clips dataset

model	accuracy	number of samples	number of activities
CNN on short clips dataset	0.873	1226	8
CNN on long clips dataset	0.854	240	8
Galvàn-Tejada et al. RF	0.856	1191	8
Kabir et al. scenario A	0.748	1000	10
Kabir et al. scenario B	0.748	1300	13
Kabir et al scenario C	0.733	1600	16

TABLE II: Comparison between similar approaches in HAR from audio on home activities

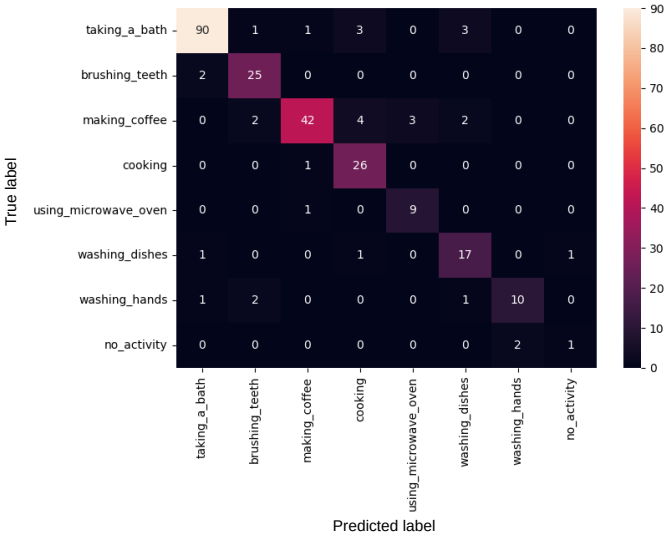

Fig. 4: Confusion matrix of the evaluation on the short clips dataset test set

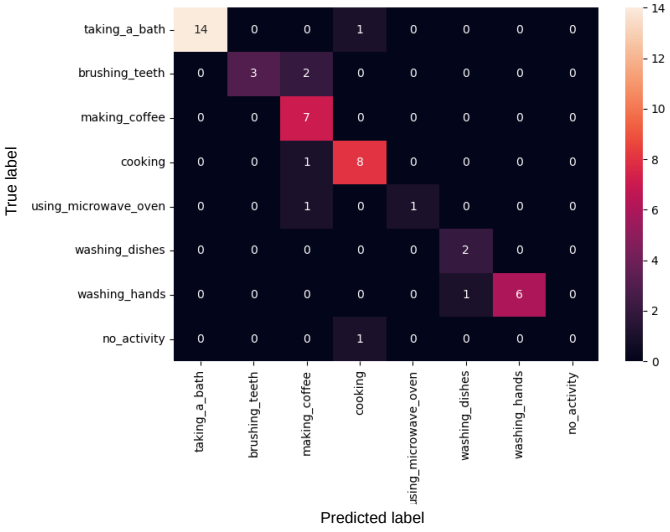

Fig. 5: Confusion matrix of the evaluation on the long clips dataset test set

REFERENCES

[1] Carlos E. Galván-Tejada et al., *An Analysis of Audio Features to Develop a Human Activity Recognition Model Using Genetic Algorithms, Random Forests, and Neural Networks*. Universidad Autonoma de Zacatecas, Mexico, 2016.

[2] J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford, "A hybrid discriminative/generative approach for modeling human activities," in Proceedings of the 19th International Joint

[3] M. Z. Uddin, D.-H. Kim, and T.-S. Kim, "A human activity recognition system using HMMs with GDA on enhanced independent component features," International Arab Journal of Information Technology, vol. 12, no. 3, pp. 304–310, 2015.

[4] I. T. Jolliffe, Principal Component Analysis, Wiley Online Library, Chichester, UK, 2002.

[5] Y. Wang, K. Huang, and T. Tan, "Human activity recognition based on R transform," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07), pp. 1–8, IEEE, June 2007.

[6] Z. He and L. Jin, "Activity recognition from acceleration data based on discrete consine transform and SVM," in Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '09), pp. 5041–5044, IEEE, San Antonio, Tex, USA, October 2009.

[7] A. Hyvarinen, J. Karhunen, and E. Oja, ¨ Independent Component Analysis, vol. 46, John Wiley & Sons, New York, NY, USA, 2004.

[8] M. Mascia, A. Canclini, F. Antonacci, M. Tagliasacchi, A. Sarti, and S. Tubaro, "Forensic and anti-forensic analysis of indoor/outdoor classifiers based on acoustic clues," in Proceedings of the 2015 23rd European Signal Processing Conference (EUSIPCO '15), pp. 2072–2076, IEEE, Nice, France, August 2015.

[9] M. H. Kabir, M. R. Hoque, K. Thapa, and S.-H. Yang, "Twolayer hidden markov model for human activity recognition in home environments," International Journal of Distributed Sensor Networks, vol. 12, Article ID 4560365, 2016.

[10] S. P. Tarzia, P. A. Dinda, R. P. Dick, and G. Memik, "Demo: indoor localization without infrastructure using the acoustic background spectrum," in Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services (MobiSys '11), pp. 385–386, ACM, July 2011.