Crowd emotional sounds: spectrogram-based analysis using convolutional neural networks

Valentina Franzoni[1]
Department of Mathematics and Computer Science
University of Perugia
Perugia, Italy
valentina.franzoni@dmi.unipg.it

Giulio Biondi
Department of Mathematics and Computer Science
University of Florence
Florence, Italy
giulio.biondi@unifi.it

Alfredo Milani
Department of Mathematics and Computer Science
University of Perugia
Perugia, Italy
milani@unipg.it

Abstract— **In this work, we introduce a methodology for the recognition of crowd emotions from crowd speech and sound in mass events. Different emotional categories can be encoded via frequency-amplitude features of emotional crowd speech. The proposed technique uses visual transfer learning applied to the input sound spectrograms. Spectrogram images are generated starting from snippets of fixed length taken from the original sound clip. The plots are then filtered and normalized concerning frequency and magnitude and then fed to a pre-trained Convolutional Neural Network (CNN) for images (*AlexNet*) integrated with domain-specific categorical layers. The integrated CNN is re-trained with the labeled spectrograms of crowd emotion sounds in order to adapt and fine-tune the recognition of the crowd emotional categories. Preliminary experiments have been held on a dataset collecting publicly-available sound clips of different mass events for each class, including *Joy*, *Anger* and *Neutral*. While transfer learning has been applied in existing literature to music and speech processing, to the best of our knowledge, this is the first application to crowd-sound emotion recognition.**

Keywords—emotion recognition, image recognition, crowd computing, CNN, transfer learning, crowd emotions

I. Introduction and Related Work

Research on sound emotion recognition has mostly focused on emotions elicited by music, [1] or emotions of individual speakers, [2], [3],[4] expressed by fine-tuning of different shades of vocal features.[5], [6] Recognizing sound emotions from the crowd is of great interest for applications which focus on detecting emotional content of mass events, e.g., alerting in emergencies, mass panic, riots, or automated video annotation for sports, concerts, political meetings. Crowd emotion recognition is also useful for providing applications with contextual information from the sound background and the environment in which the user activity is taking place. A challenging issue of sound-based analysis of the emotions embedded in the screaming of the crowd, is that they are not simply the summation of the individual emotional sounds, as they would be expressed in a single-person conversation. When people scream in the crowd, they mostly use short words or single modulated utterances, together with the other people, and they also use special sounds, e.g., booing, or whistling in approval, or they produce clatter sounds by clapping, hitting tables or shaking objects. In other words, they behave and emit sounds as a crowd collective subject, e.g., a chorus although without centralized control. Moreover, in real situations, such as a sports match, emotions of different crowds can mix up, e.g., scoring team and losing team supporters shouting and clattering together. We, therefore, define *crowd speech* the whole set of simultaneous sounds, both voice-based and clatter-based. Emotional crowd speech recognition has, therefore, specific characteristics and thus demands a special treatment. [2]

In this work, we hypothesize that emotions in crowd sounds are characterized by frequency-amplitude features which are less dependent from individuals; in other words, all the crowds are similar, and are, in some sense, the same crowd. The method we propose here is to exploit the sound analysis and evaluation through a visual transfer learning technique based on *Convolutional Neural Networks* (CNNs).

Starting from a labeled data set of crowd speech sounds from real events obtained from *YouTube* videos, the visual plots, representing *frequency/magnitude* spectrograms *over time*, are generated from snippets sampled with sliding windows over the whole original clips [7] of the crowd sound. The visual plots are then filtered and standardized, as detailed in the next paragraphs, in order to make them homogeneous in scaling and encoding.

Since the low-level visual analysis needs high computational capabilities in terms of time, memory, and training data, while the specific features of the data set are related to high-level details of the sound, we chose a *transfer learning* approach. Transfer learning allows to leverage a pre-trained network for the analysis of the low-level features and fine-tune the network on a limited amount of specific clips. The spectrogram images are therefore fed to a pre-trained Convolutional Neural Network (CNN) for images [8]–[11] based on AlexNet, [12] which has been integrated with additional classification layers for crowd emotion categories. This supervised domain-specific training phase [13] has the purpose of fine-tuning, adapting, and enabling the additional

[1] *Corresponding author*

CNN layers to crowd emotion recognition. Finally, the completely trained CNN is used to recognize emotions from a test set of spectrograms (see architecture flow in Fig.1).

In order to build a *crowd speech* data set, we chose to select sound clips from real events taken from the internet (e.g., ambiance sound from football stadium during goal or no-goal attacking phases, cheering or booing audience in large events, and crowd sounds from relevant riots in big cities). We excluded any sound generated on purpose, e.g., actors' performances (e.g., movies, sound effects for video editing) because they are not representative of real emotions, but of simulated emotions. Our data set includes a total of 678 snippets, divided into three emotional categories, i.e., *joy*, *anger*, and *neutral*, [14] respectively corresponding to the three crowd emotions obtained by cheering, rioting, and neutral background noise in mass events.

A further step of transfer learning is done when applied to objects which are not images, i.e., *Heterogeneous Transfer Learning* (HTL). The key point is that sound can be transformed in an image encoding all the relevant original sound features, i.e., a spectrogram in the domain of frequency-amplitude plots of our crowd emotional sounds data set. The HTL methodology is not new to sound recognition, [5], [15] but to the best of our knowledge, this is the first application to crowd emotion recognition.

The substantial differences between the large set of natural images included in the network pre-training and our sound-plot images may advise against visual transfer learning because data are in the same feature space, but with different distributions. Moreover, concerning other sounds input, e.g., individual speech, crowd emotional sounds often present small differences between a category and the other, and they are strongly affected by environmental noise. However, the promising results of previous research on transfer learning applied to emotional speech [4], [16] encourage to use transfer learning for crowd sounds provided a sufficient amount of training images for the fine-tuning phase. On the other hand, using CNNs with a consistent pre-training, e.g., *GoogleNet*, *AlexNet*, [17] should provide the advantage of a performant recognition of the image low-level features, e.g., shapes, edges, color distribution, without the computational cost of training from scratch.

This work aims to investigate whether such an approach characterized by *heterogeneous visual transfer learning* can be adequately applied to CNNs operating on sound spectrograms in order to realize crowd sounds emotion recognition, which seems positively supported by our preliminary experimental results. As a further result, we hope to stimulate the discussion on the problem of crowd emotion recognition, related models, and applications. In the following paragraphs, the characteristics and architectural workflow of the heterogeneous transfer learning systems applied to the Alex-Net CNN for crowd sound emotional recognition, the data set, the experimental settings, and the obtained results are described and discussed.

II. The System Architecture Workflow

The organization of information flow of the *Heterogeneous Transfer Learning (HTL)* in the proposed system includes two main phases (depicted in the architectural scheme in Figure 1): *sound to spectrograms transformation*, and *knowledge transfer training*.

The *sound to spectrograms transformation* consists in starting with a labeled sound clip of varying duration, then sampling it by blocks of 2 seconds, normalizing the sound parameters, and finally generating a standardized spectrogram from each block, labeled with the emotion of the original clip.

The *knowledge transfer training* is a quite standardized process. [18] the original CNN is modified in the last levels and then retrained by the spectrograms images to recognize the emotional crowd labels.

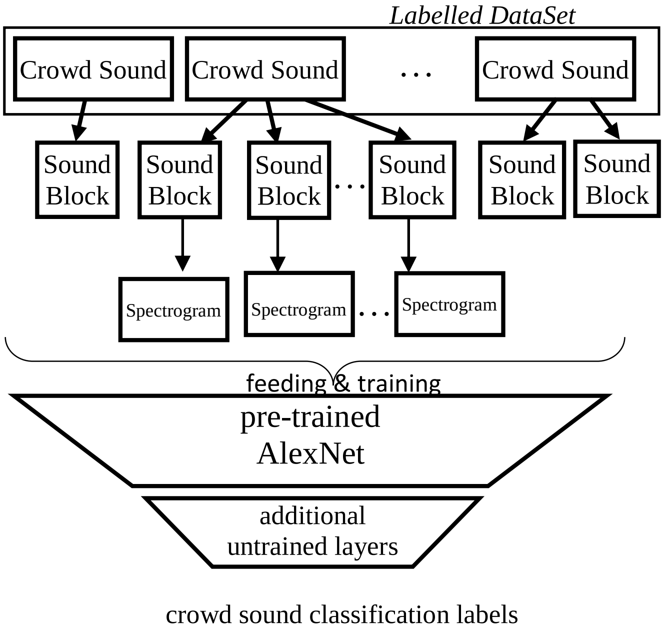

Figure 1: Heterogeneous Transfer Learning architecture flow.

A. Sound Normalization

In order to prepare images of uniform size, each original sound clip has been sliced in sound blocks of $t_b=2$ seconds using a time sliding window with $t_s=1$ second slide and $t_b–t_s=1$ second overlap. A smaller overlap has been tested with scarce results concerning accuracy. The 1-second overlap allows a faster training computation.

The audio frequency has been cut to the human hearing range, i.e., from *20* to *20000* Hz. The range has been kept wide in order to include all the clatter sounds composing the crowd speech sound.

The loudness of the whole sound blocks data set is then normalized on *-23* LU, using the *EBU R128* standard [19], which measures audio in Loudness Units, *LU* or *LUFS* (Loudness Units, referenced to Full Scale).

The range of spectral magnitude of our dataset results in a *[minimum, maximum]* range of *[-130,-22] dB*, expressed as *power/Hz.*

B. Generation of crowd-speech frequency/amplitude spectrograms

Each sound block has been finally used to generate a spectrogram in the *melodic* (*mel*) perceptual scale of pitches

[20] (see examples in Figure 2). The *mel* scale represents the sound pitch based on listener perception. A perceptual pitch of *1000* mel is assigned to a tone of *1000* Hz, *40* dB above the listener's threshold.

The *mel* spectrogram represents the short-term power spectrum of a sound, and transforms the input raw sound sequence into a bidimensional feature map where the *x*-axis represents *time*, the *y*-axis represents the *frequency* (log10 scale), and the values represent *amplitude*.

We generated magnitude spectrograms of size *257x259* for frequency and time, using the *jet* colormap of *64* colors, particularly suitable for our recognition goal, for the luminance of colors, which is not easily garbled. Then the spectrogram images have been downsized to *227x227* pixels, which are the input dimensions for the AlexNet CNN.

During the spectrogram generation, a Hamming window has been applied, to smooth the discontinuities in the original, non-integer number of periods in the signal. With this technique, we hopefully avoid the recognition of non-existing elements due to ripples with strong luminance values in the *mel* spectrogram. [21] The Hamming window size is *400* samples, with a frame increment of 4.5 milliseconds.

As a final step, we balanced the classes, reducing the data set to have about the same number of images for each class, for a total of 678 spectrograms. Such reduction has been done deleting random blocks.

Future directions will include experiments on unbalanced classes.

C. *Domain-specific training of the AlexNet CNN*

The featured crowd emotions analyzed in our experiments are *joy*, *anger*, and *neutral*, respectively corresponding to the three crowd emotions obtained by cheering, rioting, and neutral background noise in crowd events. The classes have been balanced in the amount of blocks.

The visual transfer learning technique exploits the tremendous recognition capabilities of CNNs trained with huge data sets of images, i.e., in our case, the *ImageNet* database [22] using the *AlexNet* CNN, pre-trained on *ImageNet*. AlexNet has been chosen because particularly efficient in image recognition, but also because already used in speech emotion recognition, for comparison purposes. The basic idea is that the early layers of well-trained image recognition CNNs are somewhat all similar. [5] Early layers are specialized in recognizing image features of increasing complexity from pixels with high contrasting neighbors to edges, corners, to larger areas with color distributions and more complex shapes. According to this interpretation, only the final layers are operating some composition of the previous features, thus implementing the final categorization. The actual knowledge transfer is realized by starting with a CNN pre-trained by supervision on over a million of general images and a thousand categories, and by replacing the last layers with one or more new layers and focusing them towards the recognition of few different categories. The advantages are related to the speed of this process, faster than entirely training the CNN from scratch, and to the possibility to re-train the new categories with fewer image samples than the huge number of samples needed by the original network, or a new one.

D. *Crowd sound data set*

For the two preliminary experiments presented in this paper, we collected a data set of cheering, rioting, and neutral crowd sounds extracted from *YouTube* videos of different events and duration. The visual part of the video helped us to label the extracted sound clips correctly.

We purposely avoided any clip which includes actors' performances, in order to train the network only with true crowd sounds, to be realistic in their complexity.

The chosen videos include:

- *Cheering* (crowd shouting, big crowd clapping) for the *Joy* emotion category;
- *Rioting* (crowd shouting, banging, clapping, police intervention) for the *Anger* emotion category;
- *Background noise* (people chatting, laughing) in crowded events for *Neutral* emotion category.

The different clips have been chosen to share several similar characteristics (e.g. noise, continuous or rhythmic sounds), in order to avoid a bias introduced by considering inherently different categories. The crowd sound data set is composed of 890 blocks in total from 18 original clips for the three categories, for a duration of 1711 seconds (see Table 1).

Table 1: Per-class blocks number and duration, in seconds

Class	**Different Clips**	**Blocks Total Duration (*s*)**
Joy	9	199
Neutral	3	84
Anger	6	1428
Total	**18**	**1711**

E. *Experimental setup*

The experiments are divided into two approaches, both using the *80%* of the sound blocks, i.e., spectrograms, for the fine-tuning of the CNN, and the remaining *20%* for test/validation, in order also to prevent and detect overfitting.

The first approach implements the training and test on sound blocks randomly extracted from the data set (see examples in Figure 2). This is a standard approach on images, also used in state-of-the-art works on speech emotion recognition. [4] Our consideration on this approach is that the accuracy of the results may be influenced by the fact that a random split will consider for training and test different blocks of the same sound clip, possibly including the same part of the clip, if pertaining to blocks generated through the sliding window from the same contiguous frames in the 1-second windows overlap. This approach is weak for overfitting detection, because we cannot prevent the algorithm to randomly extract the same overlapping frames shared by two different contiguous blocks, potentially used one for training and the other for testing. If not contiguous, two blocks from the same sound clip may include very similar characteristics (noise, rhythmic or continuous sound, e.g., clapping or screaming).

Therefore, we implemented a second experimental approach splitting the training and test subsets, choosing different original files manually, i.e., selecting blocks from different sound clips.

The network has been trained for *six* epochs, using a minibatch size of *10* images. Both the initial learning rate and the L2 regularization factor were set to $1*10^{-4}$. The model was tested on the validation data every *three* iterations; training images were shuffled at the beginning of each epoch, and validation images before each validation step.

In both the experiments, results are evaluated using validation accuracy. [23]

III. Experimental Results

Experiments using random train/test splitting show that the task can be handled effectively by the network; three training epochs are sufficient to achieve perfect accuracy, i.e., *1*, on the data set. No substantial improvement is obtained with more epochs, concerning to both training and validation error. On the other hand, after manually splitting the data set as described in section II.E, performances are only slightly lower, although more consistent, and free from overfitting.

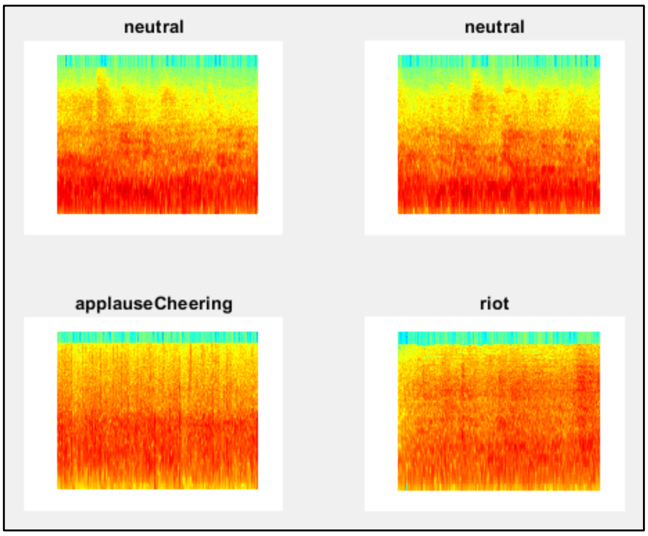

Figure 2: Randomly chosen examples of CNN classification.

The final accuracy score, in this case, is *98.54%*, and the training process reaches the optimum faster than the previous, i.e., at the beginning of the *second* epoch instead of the *third*.

Premising that our experiment is not directly comparable to speech emotion recognition in [4], which analyzes a very different emotional sound domain, and uses different settings for the sound blocks generation, we can say that our experiments show a significant performance with respect to their *80%* accuracy, even in the second experiment where we lower the performance gaining a better consistency. This result fosters to conclude that the transfer learning AlexNet-CNN-based approach is suitable for the crowd emotional sounds domain.

Future directions hint to compare the experiments using different spectrogram perceptual scales of pitches than the *mel* scale, e.g., *log*, *bark*, *erb*. Experiments can also be extended to other data sets, e.g., goal/no goal soccer matches cheering, more complex by itself, because the different crowds can mix up, e.g., the scoring team and the losing team supporters shouting and clattering together at the same time. In order to recognize goal/no goal actions, we should consider not only that supporters and opponents for the attacking team will indeed shout together at the same time, but also that both actions concluded with a goal and others leading to a fail share the same initial phase of cheering. Therefore, the relevant elements will be the final phase with the goal or fail, and the pattern of growing excitement, useful to identify such last phase. Therefore, the evolution of the action should be considered, instead of simply focusing on the separate sound blocks. Applying instead the actual algorithm to such a data set, collecting all the goal actions and some of the no goal actions of the final match Brazil-Germany of the World Cup 2014 (finished with the result of 1-7), we obtain only around *0.72* accuracy for the random approach and *0.53* accuracy for the manual approach.

A final direction is to implement the approach on imbalanced classes, naturally inherent real-world applications, with the aim of developing a real-time emotional crowd sound recognizer.

IV. Acknowledgments

This work has been partially funded by the research project DDG n.123 5/16/2018 Models and Systems for e-learning process monitoring and integration of the University of Perugia. The authors thank Prof. Margaret Lech from School of Engineering, RMIT University, for the precious and valuable technical insights.

V. REFERENCES

[1] J. J. Deng, C. H. C. Leung, A. Milani, and L. Chen, “Emotional States Associated with Music,” *ACM Trans. Interact. Intell. Syst.*, 2015.

[2] M. Forsell, “Acoustic Correlates of Perceived Emotions in Speech,” *Infancy*, 2007.

[3] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017.

[4] M. Lech, M. Stolar, R. Bolia, and M. Skinner, “Amplitude-Frequency Analysis of Emotional Speech Using Transfer Learning and Classification of Spectrogram Images,” *Adv. Sci. Technol. Eng. Syst. J.*, 2018.

[5] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, “Speech Emotion Recognition Using CNN,” 2014.

[6] H. M. Fayek, M. Lech, and L. Cavedon, “Towards real-time Speech Emotion Recognition using deep neural networks,” in *2015, 9th International Conference on Signal Processing and Communication Systems, ICSPCS 2015 - Proceedings*, 2015.

[7] S. Prasomphan, “Detecting human emotion via speech recognition by using speech spectrogram,” in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2015, pp. 1–10.

[8] M. Riganelli, V. Franzoni, O. Gervasi, and S. Tasso, “EmEx, a tool for automated emotive face recognition using convolutional neural networks,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10406 LNCS, pp. 692–704, 2017.

[9] O. Gervasi, V. Franzoni, M. Riganelli, and S. Tasso, “Automating facial emotion recognition,” *Web Intell.*, 2019.

[10] A. Bonarini, “Can my robotic home cleaner be happy? Issues about emotional expression in non-bio-inspired robots,” *Adapt. Behav.*, vol. 24, no. 5, pp. 335–349, 2016.

[11] G. Biondi, V. Franzoni, O. Gervasi, and D. Perri, “An Approach for Improving Automatic Mouth Emotion Recognition BT - Computational Science and Its Applications – ICCSA 2019,” 2019, pp. 649–664.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[13] G. Biondi, V. Franzoni, and V. Poggioni, *A deep learning semantic approach to emotion recognition using the IBM watson bluemix alchemy*

language, vol. 10406 LNCS. 2017.

[14] P. Ekman, “An Argument for Basic Emotions,” *Cogn. Emot.*, 1992.

[15] T. F. Quatieri, “Energy Separation in Signal Modulations with Application to Speech Analysis,” *IEEE Trans. Signal Process.*, 1993.

[16] M. N. Stolar, M. Lech, R. S. Bolia, and M. Skinner, “Real time speech emotion recognition using RGB image classification and transfer learning,” in *2017, 11th International Conference on Signal Processing and Communication Systems, ICSPCS 2017 - Proceedings*, 2018.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton., “AlexNet,” *Adv. Neural Inf. Process. Syst.*, 2012.

[18] L. Chen, A. Zhang, and X. Lou, “Cross-subject driver status detection from physiological signals based on hybrid feature selection and transfer learning,” *Expert Syst. Appl.*, 2019.

[19] Standard “EBU R 128-2014 Loudness normalisation and permitted maximum level of audio signals.” 2014.

[20] B. C. J. Moore and B. R. Glasberg, “Suggested formulae for calculating auditory-filter bandwidths and excitation patterns,” *J. Acoust. Soc. Am.*, 1983.

[21] X. Liu, G. Cheung, X. Ji, D. Zhao, and W. Gao, “Graph-Based Joint Dequantization and Contrast Enhancement of Poorly Lit JPEG Images,” *IEEE Trans. Image Process.*, 2019.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *ImageNet Classification with Deep Convolutional Neural Networks*, 2012.

[23] H. M. Fayek, M. Lech, and L. Cavedon, “Evaluating deep learning architectures for Speech Emotion Recognition,” *Neural Networks*, vol. 92, pp. 60–68, Aug. 2017.