# Group-Level Emotion Recognition Using a Unimodal Privacy-Safe Non-Individual Approach

Anastasia Petrova, Dominique Vaufreydaz, Philippe Dessus

# GROUP-LEVEL EMOTION RECOGNITION USING A UNIMODAL PRIVACY-SAFE NON-INDIVIDUAL APPROACH

**Anastasia Petrova**
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France
Anastasia.Petrova@inria.fr


**Dominique Vaufreydaz**
Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France
Dominique.Vaufreydaz@univ-grenoble-alpes.fr


**Philippe Dessus**
Univ. Grenoble Alpes, LaRAC, 38000 Grenoble, France
Philippe.Dessus@univ-grenoble-alpes.fr

## ABSTRACT

This article presents our unimodal privacy-safe and non-individual proposal for the audio-video group emotion recognition subtask at the Emotion Recognition in the Wild (EmotiW) Challenge 2020[1]. This sub challenge aims to classify *in the wild* videos into three categories: Positive, Neutral and Negative. Recent deep learning models have shown tremendous advances in analyzing interactions between people, predicting human behavior and affective evaluation. Nonetheless, their performance comes from individual-based analysis, which means summing up and averaging scores from individual detections, which inevitably leads to some privacy issues. In this research, we investigated a frugal approach towards a model able to capture the global moods from the whole image without using face or pose detection, or any individual-based feature as input. The proposed methodology mixes state-of-the-art and dedicated synthetic corpora as training sources. With an in-depth exploration of neural network architectures for group-level emotion recognition, we built a VGG-based model achieving 59.13% accuracy on the VGAF test set (eleventh place of the challenge). Given that the analysis is unimodal based only on global features and that the performance is evaluated on a real-world dataset, these results are promising and let us envision extending this model to multimodality for classroom ambiance evaluation, our final target application.

***Keywords*** EmotiW 2020, audio-video group emotion recognition, Deep Learning, affective computing, privacy

## 1 Introduction

Since the emergence of the "*affective computing*" research trend at the beginning of the 2000s [1], more and more features describing people's affective states are calculable using machine learning on remote sensor signals. Deep learning has attracted significant attention of the research community thanks to its faculty to automatically extract features from images and videos, determining the best representations from raw data. Deep learning shows high performance in multiple domains, notably in computer vision and speech recognition. The literature presents a wide variety of Convolutional Neural Networks (CNNs) algorithms with extensive applications in many areas, such as human-computer interaction, image retrieval, surveillance, virtual reality, etc. These achievements impact other research areas, such as affective computing and behavioral modeling to detect and predict human actions, being one of the main multimodal perception instruments for analyzing interactions between people and groups [2]. However, most of them cannot reach the human-level understanding of *in the wild* emotions [3].

---

[1]https://sites.google.com/view/emotiw2020/, last accessed 08/2020.

In the *Teaching Lab* project, we aim at developing a smart pervasive classroom to analyze teacher–students social relationships from still images or video sequences coupled with acoustic features: namely current teacher activity, current teaching episodes, whole-class engagement, students' attention or engagement, classroom ambiance, etc. The system's goal is to analyze the underlying teaching processes (e.g., teacher–students interaction, misbehavior management), not to monitor individual behaviors *per se*, even if they are inadequate. The multimodal perception system will thus monitor the whole classroom at a glance to help teachers enhance their pedagogical practices afterward. The underlying research question is: is it possible to use deep learning to capture global moods only from the whole scene?

While perceiving groups, current state-of-the-art perception systems focus on people as individuals, i.e., each individual is detected and processed as one entity. Group level perception is then computed by iterating over the group. For instance, to detect which emotion is carried by a photograph, systems typically try to locate faces accurately and then identify emotions from facial expressions with techniques such as facial action coding. Averaging over results leads to the final estimation [4]. However, this approach can raise problems for privacy and ethical concerns. An ambient classroom has "ears and eyes" and any attendee's behavior is subject to be recorded and further analyzed [5, 6] for monitoring or surveillance purposes, thus hampering their so-called "*ambient privacy*". Since the overall goal of *Teaching Lab* systems is to obtain global information about pedagogical episodes, a question arises about the possibility to calculate it without individual sub-calculations. This kind of approach exists using audio processing [7], but, as far as we know, no research on pure global group-level emotions features on images/videos, again without including individual computations, is reported in the literature.

This paper addresses our first investigation about perceiving the mood of groups of people with an ethical and privacy-safe frugal approach, i.e. an ethical approach excluding individual-based features. The EmotiW 2020 challenge [8] and its Audio-video Group Emotion Recognition sub-challenge are an opportunity for the challenge team (*OnlyGlobalFeatures*) to confront this problem using data gathered "*in the wild*" and sustain its ethical proposal using deep neural network using only global features. The current proposed unimodal model achieved 59.13% of accuracy on the VGAF test set [9] and the eleventh place in the challenge. Using only the video stream, this promising preliminary result is an opening on future evolutions of such perception systems and will be further extended to audio signal processing.

This article is organized as follows. The related work section provides insights into current researches on group-level perception. Then, the proposed approach and the system pipeline are depicted. The next sections describe information about the training datasets and conducted experiments. Results are discussed and conclusions are drawn in the last sections.

## 2   Related work

A significant contribution to automatic computer recognition of human emotions was made by the previous EmotiW challenges participants. Audio-video Group Emotion Recognition is a new problem introduced this year, aiming to detect and classify the emotion of a group of people among three classes: positive, neutral and negative. This new task is the combination of two sub-challenges that were held in previous years. Despite the considerable progress that has been made in the field of computer vision, predicting the emotions of a group of people in the wild is still a difficult task, due to the number of faces in the image, various illumination conditions, face deformations, occlusions, non-frontal faces, etc. For this challenge, the organizers proposed a baseline based on the Inception-V3 architecture [10]. Their work includes the analysis of visual and audio cues [9].

One of these sub-challenges is Group-level emotion recognition (GReco) 2018, which consists in predicting the mood of a group of people from images. Participants combined the top-down (the usage of the global tone of a photo for analysis) and bottom-up (performing analysis for each of the extracted faces, and then the subsequent combination of the results) approaches to design their prediction algorithm. The winner proposed a hybrid deep learning network [11] obtained by fusing four individually trained models for faces, scenes, skeletons and salient regions detection using VGGNet [12], ResNet [13], Inception [14] and LSTM [15] networks. The adjustment of each model's prediction weights by grid-search strategy allowed achieving a 68.08% classification accuracy on the testing set. The second-place system proposed a cascade attention network for the face cues and CNNs (VGG19 [12], ResNet101 [13], SE-Net154 [16]) for body, face and image cues [17]. Participants used the cascade attention network to solve the partial discrepancy of some faces to the photograph's general atmosphere, which allowed them to capture the importance of each face in the image and achieve 67.48% classification accuracy. For face and body detection, the researchers used MTCNN [18] and OPENPOSE [19], respectively. The third-place proposed a four-stream hybrid network, consisting of the face-location aware global stream, the multi-scale face stream, the global blurred stream and the global stream, which achieved 65.59% accuracy on the test set [20]. For face detection, this team also used MTCNN [18]. The face stream consisted of

two networks designed for high and low-resolution face images. To learn features from the whole image, they fine-tuned a pre-trained VGG16 [12] network on the ImageNet dataset [21]. The face-location aware global stream was specially implemented to capture features from the whole image together with face-relative location information. To learn only scene related features, the participants proposed a global blurred stream, where original images with blurred faces were used as input. Finally, all four streams were fused by tuning fusion weights.

The second sub-challenge, Audio-video Emotion Recognition task (VReco), aims to predict six basic emotions plus neutral from videos, was organized seven times since 2013. In the 2019 EmotiW challenge, Li et al. [22] proposed a bimodality fusion framework containing two models. The first one is the facial image model, where four different types of neural networks (VGG-Face, Restnet18, Densenet121, VGG16) were used for feature extraction, and the obtained features were integrated into a Bi-Directional Long Short Term Memory (Bi-LSTM) model [23] to capture dynamical temporal information. The second one is the audio model based on two methods for feature extraction and classification, using deep learning feature extraction and the openSMILE toolbox [24] to compute LLD features. The fusion of the two models achieved 62.78% accuracy on the test set and ranked 2nd. Zhou et al. [25] explored emotion features (revealing the most suitable CNNs) and three types of fusion strategies (self-attention, relation-attention, and transformer), that are usually used to highlight the essential emotional features. Their model obtained 62.48% accuracy and ranked at the 3rd place.
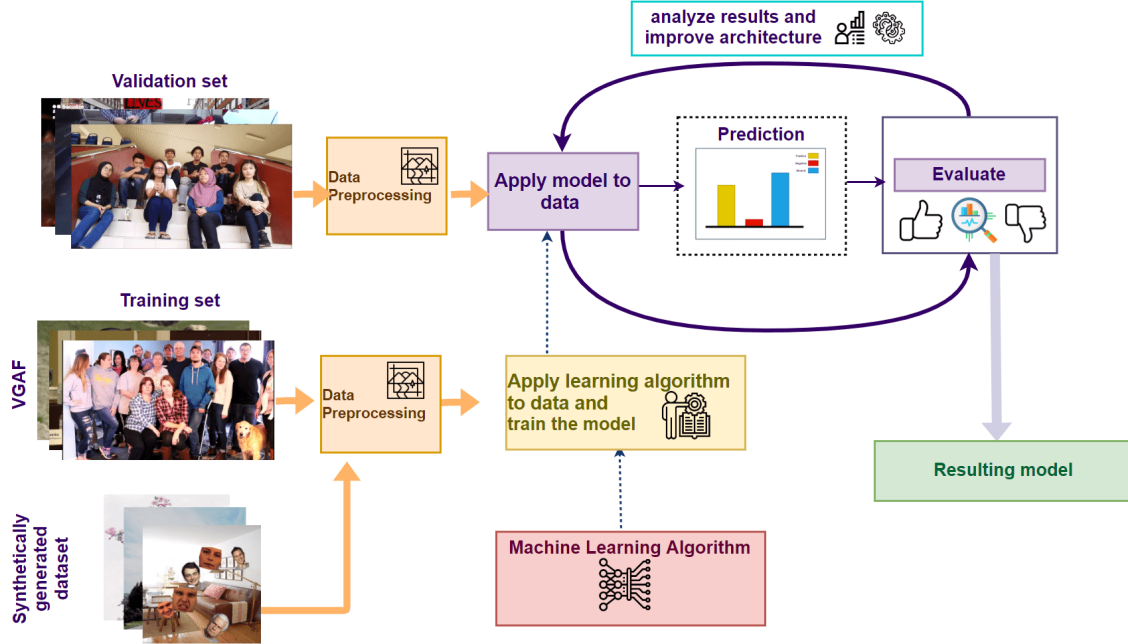


Figure 1: The pipeline for the proposed approach to group emotion recognition. This scheme reflects the key points of our analysis, namely the process of model selection, training and adjustment.

## 3  Methodology

As already stated, this research voluntarily positions itself in a frugal approach excluding the prior detection of faces or the body poses, and more generally, the detection of individuals. It differs from usual approaches that combine global features with individual ones. As traditional CNNs are by nature not explainable, i.e. one does not know what has actually been learned, our global proposal takes advantage of this fact to increase the privacy of monitored users. Recent studies use individualized visual cues coupled with these global visual features in classrooms, some also combining them with global acoustic features [20, 11, 17, 26]. In this first investigation about pure global group mood recognition, we will focus only on visual information, trying to reach the best performance before mixing with audio information. This study and our participation in the challenge is the first step towards a more comprehensive classroom ambiance recognition.

That said, as depicted in Figure 1, our methodology explores different convolutional neural network architectures, focusing those that have shown superior results for computer vision tasks over the past years. These systems were

trained on the training data and compared by evaluating their score on the challenge's validation set, revealing the most suitable architecture to such a global approach. We then selected the best initial model and gradually modified and refined it to increase the prediction accuracy on the validation set. The rationale is to assess the baseline performance of a global approach.

The available data to train group emotion recognition systems is limited. We decided to augment it with synthetic data generated on-the-fly during the training. This allows us to add a variety in terms of facial expressions, positions of faces and background types. This generation mechanism is described in section 4.2.

## 4 Datasets

This section presents the EmotiW 2020 dataset for the Audio-video Group Emotion Recognition sub-challenge and our synthetic dataset generation dedicated to group mood detection within images.

### 4.1 The VGAF dataset

The VGAF dataset [9] is the set of videos for group-level emotion recognition collected *in the wild*. It contains 1,004 videos collected from the Web with different real-world scenarios, faces, illuminations, occlusions. The data is distributed into three parts: Train, Validation, and Test with 587, 215 and 202 samples respectively. These videos were collected from different social events, such as protests, TV shows, interviews, parties, meetings, sport events, etc. There are three categories in this dataset: Positive, Negative, Neutral. The duration of each video is between 8 and 25 seconds. The number of people present in each video ranges from 2 to 62. Data annotation was done manually by three different persons. The final label has been agreed upon by all annotators.

Figure 2 highlights the difficulty of the task through examples of extracted frames from the VGAF [9] train dataset for each class. In most cases, the images contain non-frontal faces, occlusions, not clear backgrounds. The images contain spontaneous emotions that people express in everyday life.



Figure 2: Examples of frames from VGAF [9] dataset, from left to right: Positive, Neutral and Negative images. Three frames from VGAF dataset for Positive, Neutral and Negative classes, respectively.

### 4.2 Synthetic dataset

Besides VGAF, we created a synthetically generated dataset to enlarge the number of training samples. We used an approach similar to those described in [27]. The process consists in creating images using real faces showing different emotions and superimpose them on an arbitrary background. The underlying idea is to guide the neural network while training to focus on faces in images while ignoring background pixels.

#### 4.2.1 Generation algorithm

The generation process is depicted in Figure 3. It consists of several key steps. First, the generator selects a background and $N$ faces randomly. Analyzing statistics from the VGAF dataset, the number of people present on each video ranges from 2 to 62, while the mode equals 9. Most videos contain 4 to 18 faces. Knowing that, we decided to set $N \in [1, 9]$. After data augmentation on selected faces, the generation algorithm places them on the image at random positions but, using a binary mask, the algorithm prevents occlusions between faces. Then, it computes the image label. All negative emotions (*anger*, *fear*, *disgust*, *sadness*) were combined into the class 'Negative', *happiness* and *neutral* respectively into 'Positive' and 'Neutral' classes. Each of the generated images contains emotions from different classes. The emotion label is set as the most common class presented in the image. If none, it is set to *neutral*. Figure 6 shows examples of generated images for each class. The next sections detail these generation steps.
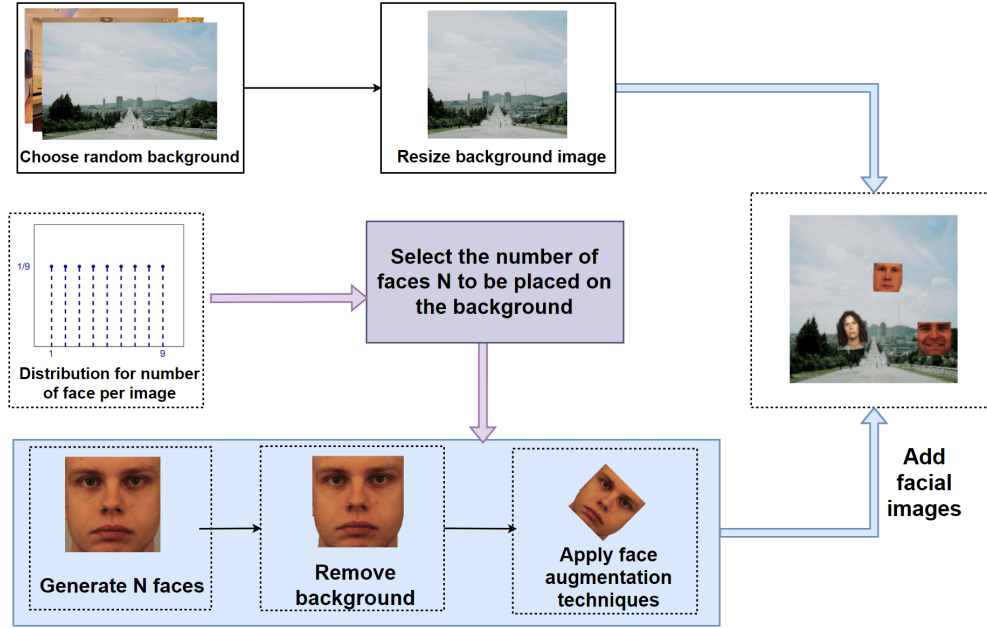
Figure 3: Synthetic Dataset Generation Scheme. This scheme represents the key steps of data generation process



Figure 4: Background images from LSUN [28] dataset, from left to right: tower, restaurant, classroom. Three examples of background images used for synthetic data generation.

### 4.2.2 Background images

The first step in creating synthetic images is the background selection. The goal of the generation process is to feed the neural network with many different environments to learn to ignore them: extracting only emotions, regardless of background. Our expectation is that the system trained with such images would learn to focus on people. The dataset LSUN [28] is a good background provider as it contains 10 different scene categories: classroom, bedroom, bridge, church outdoor, conference room, dining room, kitchen, living room, restaurant, tower. Figure 4 shows some examples of background images.

### 4.2.3 Face images augmentations

To create synthetic images related to group emotion, the generator needs photographs of faces showing different emotions. For this, two facial expression databases are available: FACES [29] and KDEF [30]. These datasets contain frontal facial photos of young, middle-aged, and old people that bring an essential variety to the image collection. As one can see in Figure 5, the FACES image contains the entire head, the neck and part of the shoulders. In our generation process, these images are mixed with a cropped version of the KDEF frontal views[2] focused on the face[31, 32], from the chin to the top of the forehead (see Figure 3). The emotion annotation in both datasets is available in discrete emotion labels, namely the universal facial expressions plus neutral. Due to the presence of the background color on the face images, before overlaying them on the background image, the algorithm gets rid of it using color filtering with OpenCV [33].

---

[2]Downloaded from https://www.ugent.be/pp/ekgp/en/research/research-groups/panlab/kdef, last accessed 08/2020.
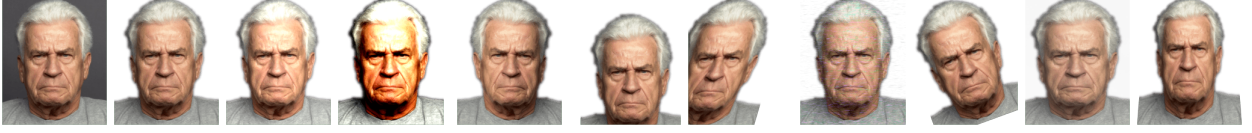
Figure 5: Examples of face augmentations, from top to bottom, from left to right: source image from [29], source image without background, brightness change, contrast change, horizontal flip, translate, shear, elastic distortion, rotate, scale, perspective transformation. Examples of data transformations used for facial images.



Figure 6: Examples of synthetically generated data. Labels from left to right: Positive, Negative, Neutral. These images shows three examples of generated data: facial images on the arbitrary background.

To enrich the synthetic corpus, data augmentation applies to face images. Wang and al. [34] collected an extensive number of state-of-the-art approaches. From the proposed set of augmentation methods, we identified the most suitable, i.e. the augmentations that distort the initial data in a non-destructive way. Geometric and photometric data transformations are considered. The first type includes transformations that alter the pixels' position in the original image, and the second type encompasses changes in the RGB channels. Figure 5 shows examples of the retain photometric and geometric transformations used for face augmentation.

### 4.2.4 Validation

After finalizing the generation process, the first stage was to validate the potential of this data as a training resource. Indeed, before using the synthetic corpus in the challenge system, we had to verify that it is suitable to learn group emotions only with global features. We conducted several tests. As a starting point, we defined a two-class experiment (happiness or not) with only one face within the images. We implemented a model similar to Smile-CNN proposed by Chen & al. [35]. The authors presented a deep learning architecture focused on detecting one of the most frequent human emotions [36], joy (aka happiness in corpora's labels). This architecture intends to detect smiles under both laboratory-controlled and wild data, on single face images. This straightforward architecture was designed to work on rather small face images ($64 \times 64$), not on faces in context. The result was equivalent to random.

We decided to evaluate deeper neural networks. Even complexifying the task with up to 9 faces on the image, modified versions of ResNet 18 [13] and Inception V3 [10] achieved respectively 83.4% and 88.5% of accuracy. These results validate that the synthetic dataset can serve as training corpus in group emotion recognition and confirm the potentiality to detect emotions from the whole image with multiple faces. We started our work on the EmotiW Challenge from this point.

## 5 Experiments

This section describes the implementation details, the evaluation of different approaches, the description of the proposed model and the challenge results. For all our experiments, we used PyTorch [37] as a deep learning framework. We compared different models by training them using VGAF and our synthetic datasets, validating them on the VGAF validation set. The challenge organizers were in charge of computing challenge results.

## 5.1 Selection of underlying architecture

Starting from validation results (section 4.2.4), we decided to explore exhaustively different state-of-the-art neural network architectures. Indeed, for some computer vision tasks, the advantages and drawbacks of these architectures have been explored. As far as we know, it is not the case for group-level emotion recognition from images. The accuracies of 13 widely employed deep learning architectures were compared on a challenge validation set. Results using these architectures as backbone, modifying them to provide a 3-class classification are presented in Table 5.1. Each architecture was modified by changing the number of neurons in the last fully-connected layer to 3, as we have a ternary classification task. All of the considered models were pretrained on the 1000-class ImageNet dataset [21], which contains more than one million high-resolution images. Since these images are significantly different from the challenge's task, these models were then fine-tuned using the VGAF training dataset.

The computed results are quite surprising. First, unlike first results on binary classification (section 4.2.4), Inception V3 has a lower accuracy than ResNet18. Second, the deeper architectures have not the best accuracies. Among all of the evaluated algorithms, VGG-19 showed the best performance (52.36% of accuracy). Therefore, this architecture is the basis of the proposed model. The implementation details, as well as challenge results, are described in the next sections.

| Underlying architecture | Validation accuracy |
|---|---|
| Wide ResNet50 [38] | 43.60% |
| ResNet34 [13] | 43.60% |
| MobileNet v2 [39] | 45.04% |
| DenseNet [40] | 45.95% |
| ResNeXt [41] | 47.39% |
| Inception V3 [10] | 47.91% |
| ResNet50 [13] | 48.17% |
| GoogLeNet [14] | 48.43% |
| VGG-11 [12] | 48.69% |
| ResNet18 [13] | 49.22% |
| VGG-16 [12] | 50.91% |
| SqueezeNet [42] | 51.44% |
| **VGG19** [12] | **52.36%** |

Table 1: Performance of different network architectures on the VGAF validation set in order of accuracy

## 5.2 Challenge model

This section presents the optimization of the challenge model, its official accuracy and ranking of the team in the Audio-video Group Emotion Recognition sub-challenge.

### 5.2.1 Model optimization

For the challenge, starting from the initial VGG19 model, we further explored several options to improve the model accuracy:

- Train the model from scratch, allowing the network to learn more specific deep features, closer to the task;
- Vary the size of linear layers;
- Add complementary linear layers;
- Fine-tune only last layer of our model;
- Combine the above options.

Comparing all of the aforementioned options, the best performance on the validation set is observed by implementing the following procedure on the model presented in Figure 7. First, some fully-connected layers, dropout layers and ReLU activation functions are added at the end of the network, finishing by a ternary tensor. As training from scratch does not provide improvement, the model is first trained on ImageNet. Using these weights, the model is then fine-tuned on both synthetic and VGAF corpora. For each epoch, we generated 10k images using our data generation algorithm and sampled 10k frames from VGAF videos, each frame having an equal probability for being selected. For all images
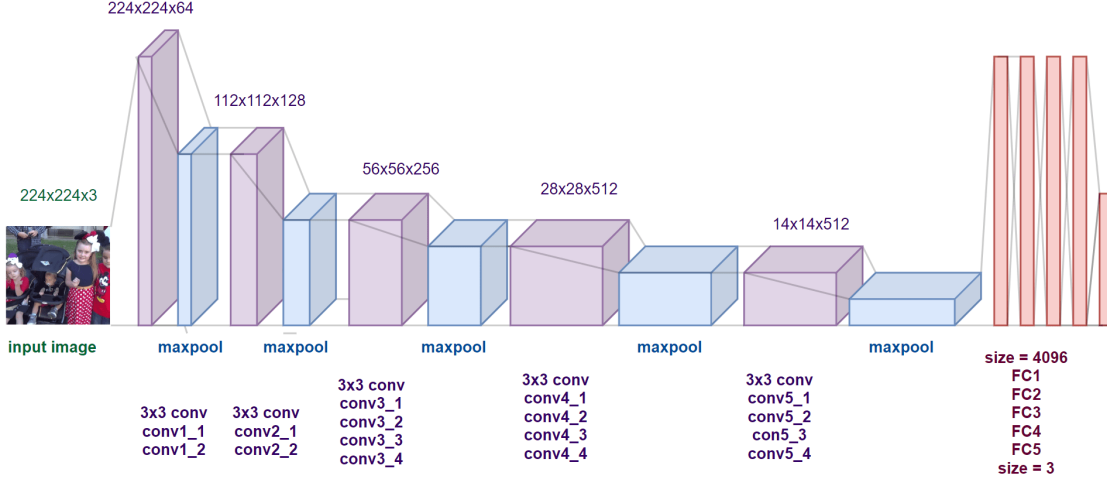
7

Figure 7: Scheme of the modified VGG-19 architecture. The scheme is a sequence of different layers: convolutional, maximum pool, dropout and fully connected layers, used in the architecture.

| Class | Precision | Recall | F1-score |
|-------|-----------|--------|----------|
| Neutral | 0.40 | 0.62 | 0.48 |
| Positive | 0.60 | 0.62 | 0.61 |
| Negative | 0.80 | 0.50 | 0.61 |
| Mean value | 0.60 | 0.58 | 0.57 |

Table 2: Precision, Recall and F1-score for Neutral, Positive and Negative Classes on the validation set

| Class | Test accuracy (%) |
|-------|-------------------|
| Positive | 59.9078 |
| Neutral | 46.6019 |
| Negative | 75.2174 |
| All classes | 59.1270 |

Table 3: Accuracy for our best challenge submission.

data transformation is performed: arbitrary $\pm 10°$ rotation, horizontal flip, random scaling and cropped $224 \times 224$ regions for frames from the VGAF dataset. After 10 epochs, the model is trained only with the VGAF dataset using 20k frames for each epoch. The training parameters can be summarized as follows. The learning rate equals 0.001. The training script uses stochastic gradient descent with 0.9 momentum as optimizer, and cross-entropy as loss function.

To obtain video-level labels from frame-level classification, we applied two techniques: score averaging and score accumulation. The best results were observed by predicting labels for all frames and then averaging the scores. The final system has a 57.18% accuracy on the validation set.

### 5.2.2 Challenge results

The challenge organizers allowed to make a total of five submissions to evaluate the method on the test set. The team made three attempts, and managed to achieve an accuracy of 59.127%, reaching the eleventh place out of nineteenth. The best team achieved 76.85%, potentially with a multimodal approach. One can consider that the achieved accuracy is reasonable, given the small size of the training set, the wild data conditions and the fact that our team employed a frugal and unimodal approach with a fairly basic temporal integration.

## 6 Discussion

In order to go deeper into the results, we did several analyses. First, the class-wise accuracy matrix is shown in Table 5.2.1. As can be seen from this table, the model shows better performance for the Positive and Negative classes on the testing set. As expected, the neutral class is the most difficult to classify. To analyze further the misclassifications, information is provided by the confusion matrix on the VGAF validation set (Figure 8) and in Table 5.2.1 that contain various metrics for each of the predicted classes. Similarly to the test set, the Negative class has the highest performance, followed by Positive and Neutral. Almost 60% of neutral videos have been misclassified, ~40% for positive and only ~20% for negative videos. As shown in the table, the trained model makes the smallest number of mistakes in predicting the class 'Negative'. It is worth noting that the 'Neutral' class is hard to classify correctly. The recall is low for all

8

classes. The F1-score for Positive and Negative classes are equivalent, while errors differ. The Negative class has the worst recall. These results are consistent with expectations, and it provides a clear way to improve the performance of our unimodal system, addressing classification behavior with regard to the different classes.
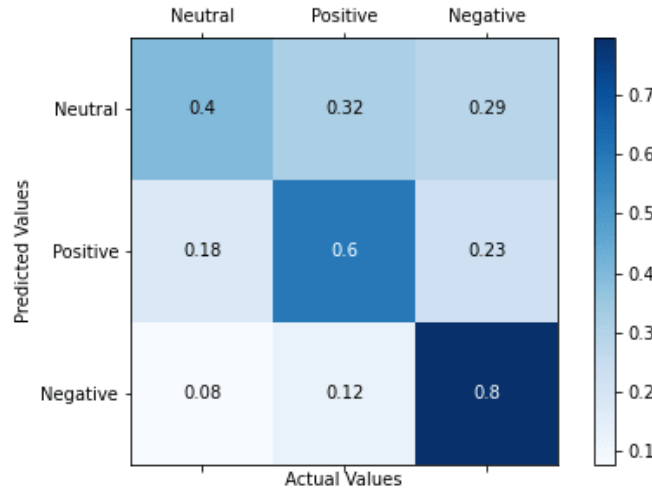


Figure 8: Normalized confusion matrix for the validation set for 3-class classification task for the final model. It shows the accuracy of a classification. It could be seen, that the best prediction accuracy for Negative and Positive classes.
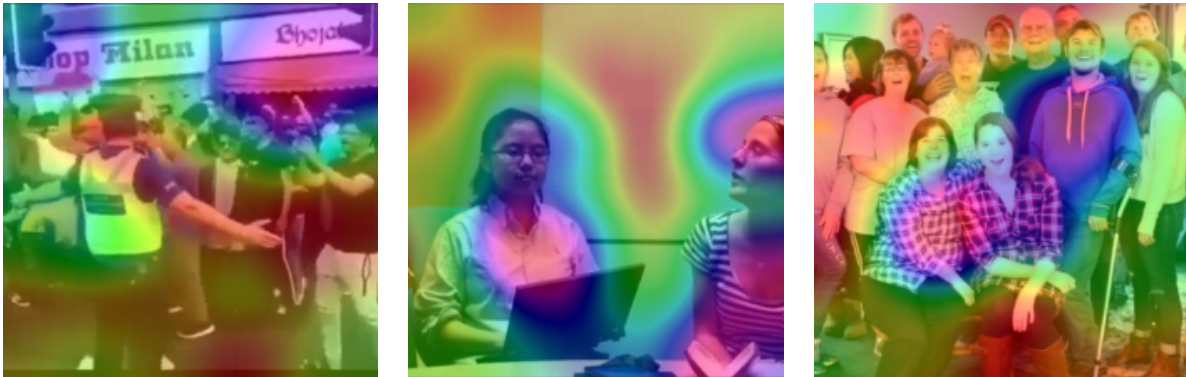


Figure 9: Negative, Neutral and Positive visualization using Score-cam algorithm [43].

Hypotheses in our methodology are two-fold. The first hypothesis is that using the image at a glance, the neural network will learn to focus on people. Moreover, using our synthetic dataset, the expectation was that it would mainly consider people's faces. The second hypothesis is that the network will ignore the background. To evaluate these aspects, we employed Class Activation Map (CAM) techniques to highlight the contribution of each part of the image to the current class classification. More precisely, we use Score-CAM [43, 44] to compute images shown in Figure 9.

These three images illustrate Score-CAM results on Negative, Neutral and Positive images. Red parts of the heatmap represent less important pixels, while green, blue and purple highlight more and more relevant ones. Looking at these representations, one can conclude that our preliminary hypotheses are not fulfilled. Among these three examples, the best heatmap is for the neutral class. The network actually focuses on both people, ignoring the background. In contrast, in the Negative heatmap, the important pixels are located in storefronts while the network overlooks most people. Last, the positive heatmap reflects that the network concentrates on some people but completely ignores others as if they belonged to the background. Unexpectedly, the neural network does not focus on all people nor their faces. It also does not totally ignore the background pixels. We need to further explore these visualizations to better understand what is relevant and modify the network training accordingly. Nevertheless, these visualizations provide a bag of pixels, but not the underlying visual features (color, shape, texture...). More investigations have to be conducted in order to identify the substance of these features.

# 7 Conclusion & future work

This paper proposes a deep learning model for group emotion detection using a privacy-safe non-individual approach. The main novelty of this research is the frugal modeling, i.e. processing only global image cues instead of including more and more individual-based features, as well as mixing available datasets with a dedicated synthetic one. We compared thirteen state-of-the-art architectures and created a pure unimodal model able to recognize the mood of groups of people. The final architecture is based on the VGG-19 model with the addition of several fully-connected layers, dropout layers and ReLU activation functions. The experiments showed the promising result of 59.13% on the EmotiW Challenge test set, in a challenging 3-class classification task *in the wild*.

In the future, some improvements need to be investigated. First, several methods can be used to enhance our synthetic corpus generation: face swapping or replacement in available datasets, facial expression generation [45, 46] or even more complex methods targeting body language. The way our proposal tackles temporal classification is relatively basic and could be improved. This could be done using Recurrent Neural Networks (RNN) or Long Short-Term Memory (LSTM) networks, for instance. Another way of improvement relies on an in-depth analysis of the proposed network to identify possible enhancements in the network architecture or in the training process. Finally, as most other participants of the challenge did, including the audio stream will be obviously of interest.

To conclude, this research on a unimodal privacy-safe non-individual approach to group-level emotion recognition is promising. The results reinforce the team's ambition to identify classroom ambiance applying ethical rules prohibiting individual computation. Within the context of the *Teaching Lab* project, the final goal of such a perception system is to provide clues to teacher trainers to help young teachers improve their pedagogical skills.

# 8 Acknowledgments

# References

[1] R. W. Picard, "Affective computing: challenges," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 55–64, 2003.

[2] S. G. Barsade and A. P. Knight, "Group affect," *Annu. Rev. Organ. Psychol. Organ. Behav.*, vol. 2, no. 1, pp. 21–46, 2015.

[3] D. Dupré, E. G. Krumhuber, D. Küster, and G. J. McKeown, "A performance comparison of eight commercially available automatic classifiers for facial affect recognition," *PLOS ONE*, vol. 15, pp. 1–17, 04 2020.

[4] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, "Emotiw 2016: Video and group-level emotion recognition challenges," in *Proceedings of the 18th ACM international conference on multimodal interaction*, pp. 427–432, 2016.

[5] K. Ahuja, D. Kim, F. Xhakaj, V. Varga, A. Xie, S. Zhang, J. E. Townsend, C. Harrison, A. Ogan, and Y. Agarwal, "Edusense: Practical classroom sensing at scale," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–26, 2019.

[6] Laurent, Romain, Vaufreydaz, Dominique, and Dessus, Philippe, "Ethical Teaching Analytics in a Context-Aware Classroom: A Manifesto," *ERCIM News*, pp. 39–40, Jan. 2020.

[7] A. James, Y. H. V. Chua, T. Maszczyk, A. M. Núñez, R. Bull, K. Lee, and J. Dauwels, "Automated classification of classroom climate by audio analysis," in *9th International Workshop on Spoken Dialogue System Technology*, pp. 41–49, Springer, 2019.

[8] A. Dhall, G. Sharma, R. Goecke, and T. Gedeon, "Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges," in *Proceedings of the ACM International Conference on Multimodal Interaction*, 2020.

[9] G. Sharma, S. Ghosh, and A. Dhall, "Automatic group level affect and cohesion prediction in videos," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 161–167, IEEE, 2019.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[11] X. Guo, B. Zhu, L. F. Polania, C. Boncelet, and K. E. Barner, "Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pp. 635–639, 2018.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[17] K. Wang, X. Zeng, J. Yang, D. Meng, K. Zhang, X. Peng, and Y. Qiao, "Cascade attention networks for group emotion recognition with face, body and image cues," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pp. 640–645, 2018.

[18] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[19] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

[20] A. S. Khan, Z. Li, J. Cai, Z. Meng, J. O'Reilly, and Y. Tong, "Group-level emotion recognition using deep models with a four-stream hybrid network," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pp. 623–629, 2018.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[22] S. Li, W. Zheng, Y. Zong, C. Lu, C. Tang, X. Jiang, J. Liu, and W. Xia, "Bi-modality fusion for emotion recognition in the wild," in *2019 International Conference on Multimodal Interaction*, pp. 589–594, 2019.

[23] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[24] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

[25] H. Zhou, D. Meng, Y. Zhang, X. Peng, J. Du, K. Wang, and Y. Qiao, "Exploring emotion features and fusion strategies for audio-video emotion recognition," in *2019 International Conference on Multimodal Interaction*, pp. 562–566, 2019.

[26] A. Ramakrishnan, E. Ottmar, J. LoCasale-Crouch, and J. Whitehill, "Toward automated classroom observation: Predicting positive and negative climate," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–8, IEEE, 2019.

[27] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 109–117, 2017.

[28] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[29] N. C. Ebner, M. Riediger, and U. Lindenberger, "Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation," *Behavior research methods*, vol. 42, no. 1, pp. 351–362, 2010.

[30] E. Goeleven, R. De Raedt, L. Leyman, and B. Verschuere, "The karolinska directed emotional faces: a validation study," *Cognition and emotion*, vol. 22, no. 6, pp. 1094–1118, 2008.

[31] A. Dawel, L. Wright, J. Irons, R. Dumbleton, R. Palermo, R. O'Kearney, and E. McKone, "Perceived emotion genuineness: normative ratings for popular facial expression stimuli and the development of perceived-as-genuine and perceived-as-fake sets," *Behavior Research Methods*, vol. 49, no. 4, pp. 1539–1562, 2017.

[32] A. Joseph and P. Geetha, "Facial emotion detection using modified eyemap–mouthmap algorithm on an enhanced image and classification with tensorflow," *The Visual Computer*, vol. 36, no. 3, pp. 529–539, 2020.

[33] J. Howse, *OpenCV computer vision with python*. Packt Publishing Ltd, 2013.

[34] X. Wang, K. Wang, and S. Lian, "A survey on face data augmentation," *arXiv preprint arXiv:1904.11685*, 2019.

[35] J. Chen, Q. Ou, Z. Chi, and H. Fu, "Smile detection in the wild with deep convolutional neural networks," *Machine vision and applications*, vol. 28, no. 1-2, pp. 173–183, 2017.

[36] D. Trampe, J. Quoidbach, and M. Taquet, "Emotions in everyday life," *PloS one*, vol. 10, no. 12, p. e0145450, 2015.

[37] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[38] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[39] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[41] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

[42] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[43] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," 2019.

[44] U. Ozbulak, "Pytorch cnn visualizations." `https://github.com/utkuozbulak/pytorch-cnn-visualizations`, 2019.

[45] Q. Xu, Z. Qin, and T. Wan, "Generative cooperative net for image generation and data augmentation," 2017.

[46] X. Wang, Y. Wang, and W. Li, "U-net conditional gans for photo-realistic and identity-preserving facial expression synthesis," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 3s, pp. 1–23, 2019.