Automatic Recognition of Emotional Subgroups in Images

Emmeke Veltmeijer , **Charlotte Gerritsen** and **Koen Hindriks**
Department of Computer Science, Vrije Universiteit Amsterdam
e.a.veltmeijer@vu.nl

Abstract

Both social group detection and group emotion recognition in images are growing fields of interest, but never before have they been combined. In this work we aim to detect emotional subgroups in images, which can be of great importance for crowd surveillance or event analysis. To this end, human annotators are instructed to label a set of 171 images, and their recognition strategies are analysed. Three main strategies for labeling images are identified, with each strategy assigning either 1) more weight to emotions (emotion-based fusion), 2) more weight to spatial structures (group-based fusion), or 3) equal weight to both (summation strategy). Based on these strategies, algorithms are developed to automatically recognize emotional subgroups. In particular, K-means and hierarchical clustering are used with location and emotion features derived from a fine-tuned VGG network. Additionally, we experiment with face size and gaze direction as extra input features. The best performance comes from hierarchical clustering with emotion, location and gaze direction as input.

1 Introduction

Recent years has seen an increased interest in the automatic recognition of group emotions from visual data [Veltmeijer *et al.*, 2021]. This serves many potential future applications, such as crowd surveillance and event detection [Guo *et al.*, 2018]. Research has also been done on the automatic detection of social subgroups [Japar *et al.*, 2021; Yücel *et al.*, 2013]. The intersection of two fields of study, recognizing emotional subgroups, has not been explored yet [Veltmeijer *et al.*, 2021]. Bridging this gap can serve several purposes. Current methods consider the people present in an image as one group and assign one shared emotion label [Tan *et al.*, 2017]. In a large crowd, it is not likely that all individuals will experience similar emotions. It is useful to recognize groups of people within the crowd that for example are feeling less happy (customer satisfaction) or even aggressive (crowd surveillance). However, tracking each individual is not an efficient way of doing so: especially in large crowds, the output would get cluttered. Additionally, an individual's emotion can be better predicted when incorporating emotions from others in their social group, while at the same time people tend to be part of social groups that feel and act in a similar manner [Frank, 1995; Mou *et al.*, 2019]. Recognizing emotional subgroups is therefore a more efficient way of detecting emotion or behaviour within a crowd. Simply combining the tasks of group and emotion recognition is not likely to suffice, since emotional subgroups can either split up or combine social groups (for example when fighting), complicating the task.

We therefore identify the strategies, features and techniques necessary for recognizing emotional subgroups. A new dataset annotation is created with human annotators, from which strategies for emotional subgroup recognition are extracted. We use several clustering methods to automatically perform emotional subgroup recognition. Based on the annotator's strategies, multiple features are implemented and tested. The contribution of this paper is three-fold: 1) describing three distinct strategies of emotional subgroup recognition by humans, 2) providing an emotional subgroup annotation for a test set of images, and 3) comparing and fine-tuning several clustering methods to automatically recognize emotional subgroups.

The remainder of this paper is structured as follows: Section 2 describes related work. Thereafter, the methods, results, and discussion of the emotional subgroup labeling (Section 3) and the automatic recognition module (Section 4) are described. Section 5 concludes the paper.

2 Related Work

2.1 Subgroup Detection

The recognition of social subgroups has been a research topic for decades. Studies proposing automatic ways of detecting social groups can roughly be divided into four categories: the detection of pedestrian groups [Wang *et al.*, 2018], conversational group and more specifically F-formation detection [Vascon *et al.*, 2016], social relationship detection [Zhang *et al.*, 2019], and automatic recognition of social groups (rather than pairs) from still images [Japar *et al.*, 2021]. These studies differ from the present work because they either work with

Code and supplementary material will be made available at https://github.com/Emmekea/emotional-subgroup-recognition.

video data rather than still images, consider subgroups based on conversation only rather than other types of interaction (conversational group detection), or include analysis of pairs rather than groups of varying sizes (social relationship detection). None of them include emotion information, which we incorporate in this study in addition to subgroup information.

2.2 Group Emotion Recognition

Group-level emotion analysis can be divided into three methods: bottom-up, top-down, and hybrid analysis [Veltmeijer *et al.*, 2021]. Bottom-up analysis works with features of individuals (such as facial expressions) [Tarasov and Savchenko, 2018], top-down analysis with contextual (e.g. full image) features [Dhall *et al.*, 2016], and hybrid analysis combines both of these [Li *et al.*, 2016; Tan *et al.*, 2017]. For an extensive overview of recent studies on group emotion recognition, we refer to [Veltmeijer *et al.*, 2021]. In this study we will work with bottom-up features, since the global nature of top-down features makes them less appropriate for the recognition of several emotions within an image. Our work differs from the aforementioned studies by recognizing multiple emotional subgroups within an image, rather than a single group emotion.

3 Emotional Subgroup Labeling

Data Acquisition

For this study, images are selected from three different datasets: EMOTIC [Kosti *et al.*, 2019], GAFF 3.0 [Dhall *et al.*, 2018], and HAPPEI [Dhall *et al.*, 2012]. A small subset is selected, with each image containing 1) at least four detectable faces, 2) at least one person that is not part of a static, standing group that poses for the picture, and 3) heterogeneous emotional expressions.

This results in the selection of 171 images in total: 18 from EMOTIC, 78 from GAF 3.0, and 75 from HAPPEI. For a complete list of exclusion criteria, see Appendix A[1].

Data Annotation

The images that are collected are annotated in three different tasks by three different annotators, who each label all data. These annotators are laymen on the topics of social group and emotion recognition (female, age 22-27). Annotation is done using Labelbox [Labelbox, 2021]. Next to the task of emotional subgroup recognition (grouping people that belong to the same social group that have the same emotion), subgroup recognition (grouping people that belong to the same social group) and individual emotion recognition (assigning a negative, neutral or positive label to each individual), are also annotated. For more information on the instructions and information the labelers are provided with, see Appendix B[1]. The tasks are performed in a fixed order: subgroup recognition first, followed by emotional subgroup recognition and then individual emotion recognition. The subgroup recognition task is put first to avoid the emotion tasks (individual emotions and emotional subgroups) influencing the subgroup perception. The third task, individual emotion recognition, is put last with the assumption that the subgroup annotations are less likely to influence the perception of individual emotions than the individual emotions would influence the (emotional) subgroups.

[1]https://github.com/Emmekea/emotional-subgroup-recognition

	No. of faces	No. of groups
Neg	53 (4.55%)	3 (1.12%)
Neu	611 (52.40%)	157 (58.58%)
Pos	502 (43.05%)	108 (40.30%)

Table 1: The number of faces and groups labeled for each emotion.

3.1 Results

Quantitative Results

Table 1 shows the number of faces per emotion category for the emotion annotation task in the left column and the number of groups per emotion category for the emotional subgroup task in the right column. It can be seen that most faces and most groups are labeled as neutral or positive.

Agreement Among Annotators. For the emotion task, there is full agreement on 81.8% of the faces, for the other 18.2% of faces two out of three annotators agree on the emotion, with an inter-rater reliability (IRR) of κ=0.77 (Light's kappa [Light, 1971]). For the subgroup task there is full agreement on 96.7% of the images and partial agreement (two out of three) on 3.1% of the images. In the emotional subgroup task, 15.8% of the images reach full agreement and 41.5% partial agreement. This is likely to be a result of the different strategies employed by different annotators, as will be discussed next. The latter two tasks are not suited for an IRR analysis, since the connections between faces (whether they are in the same group or not) are labeled rather than only the faces themselves.

Self-report. For each labeling instance, the annotators indicate which factors are relevant for their labeling. Multiple answers per image are possible. For individual emotion recognition, 'facial expressions' is the most important factor (used for all 171 images). For social subgroup recognition, facial expressions (used for 157/171 images), distance (64/171), and interaction (33/171) are most influential. Inspecting the emotional subgroup labeling shows that facial expressions (93/171) and distance are named most often (51/171).

Ground Truth Acquisition

In each task, 171 images are labeled by three different annotators. In the emotion task, majority voting is used for establishing a ground truth emotion for each face. In the subgroup task, pairwise majority voting for all face pairs is used to establish a ground truth for each image. This means that in each image, a face pair is considered to be in the same group if the majority of labelers places them together in a group.

For the emotional subgroup task, an image is only assigned a ground truth if at least two of the three annotators have identified exactly the same subgroups and labeled those subgroups with the same emotion. This results in a total of 98 images with a ground truth. We focus on these 98 images in the remainder of the analysis as they will help focus on those images that elicit a clear and distinct emotional subgroup recognition.

Recognition Strategies

To analyse the human approach to emotional subgroup recognition in relation to individual emotions and social subgroups separately, we compare the way in which each annotator performs on the three tasks. For the 98 images on which there is agreement (full or partial), 62.24% of the emotional subgroup annotations are a straightforward combination of the individual emotions and the social subgroups as indicated by the same annotator. We refer to this as the *summation* strategy. An example is shown in Fig 1(a). The emotional subgroup annotation is exactly the same as the combination of the individual emotions and social subgroups. In case of multiple individual emotions within a social group, this strategy splits up the group to subgroups with the same emotion. The second strategy fuses multiple social groups due to a similar emotion (*emotion-based fusion*). The third main strategy changes individual emotions to the emotions of their surroundings by fusing them (*group-based fusion*). Examples of emotion-based fusion and group-based fusion are shown in part (b) and (c) of Fig 1, respectively. These three strategies together account for a total of 93.54% of the annotations.

Inspecting the difference between full and partial agreement images, we find that 62.96% of full agreement images and 34.55% of partial agreement images that can be explained by the core three strategies are annotated only with the summation strategy, the rest being explained by a combination of emotion-based fusion and group-based fusion. This indicates that when labelers agree on an annotation, the summation strategy is more likely to be used than when there is disagreement on the annotation. This suggests that more challenging images lead to either different interpretations or more uncertainty.

To explore whether these strategies are also found with different annotators, we perform a validation experiment. Out of the 171 original images, 22 images are selected as a representative sample for having elicited all different strategies in the original annotation. These 22 images are then labeled by 10 labelers different from the original 3 (3 female, 7 male, age 22-27). Annotation is done using Labelbox [Labelbox, 2021]. Analysing the validation results, including all annotation data (full agreement, partial agreement, and no agreement), reveals that the summation strategy is again the most popular strategy by explaining 44.55% of the emotional subgroup annotations. Adding the emotion-based fusion and group-based fusion to this brings the percentage to 70.91%.

4 Automatic Recognition

Individual Emotion Recognition

In each of the selected images, faces are detected with the face_recognition module[2], which uses dlib face recognition[3]. The individual emotion recognition module is trained on the RAF-DB database [Li *et al.*, 2017]. For the purpose of our study, where we divide emotions into the valence categories positive, neutral, and negative, the emotion labels of the dataset are mapped to these three categories at the decision-level (after network output). Fear, disgust, sadness and anger are mapped to 'negative', happiness to 'positive', and neutral remains as is. Since experiments suggest that surprise has a mildly negative valence [Noordewier and Breugelmans, 2013], it is excluded from the dataset to only include emotions with either a clear strong negative or positive valence.

Training is done with the VGG-Face network[4], which is a network pre-trained on the VGG-Face dataset [Parkhi *et al.*, 2015]. We fine-tune the network by replacing the final three fully connected layers with two new fully connected layers. The output is a probability for each class (fear, disgust, happiness, sadness, anger, neutral), which is summed per valence category and serves as a feature for clustering. For training we use a Stochastic Gradient Descent optimizer (initial lr=0.001, exponential decay rate=0.96, 100,000 decay steps) and categorical cross entropy as loss function.

Clustering Methods

We implement two clustering techniques, K-means [MacQueen and others, 1967] and hierarchical clustering [Ward Jr, 1963]. K-means forms clusters with convex, spherical shapes, which is similar to the expected shape of social subgroups and annotations thereof. The varying number of people in a group, with multiple groups fusing or groups disassembling into smaller ones, is represented in the structure of hierarchical clustering. Additionally, hierarchical clustering is often used with geographical data [Badr *et al.*, 2015]. Spatial information is important here as well, since two people should be in each other's proximity in order to be part of the same group. Since we are working with small datasets to cluster (each image forming a 'dataset', where the faces should be clustered) both K-means and hierarchical clustering are suited for the task.

Input for both baseline clustering methods are the emotion probabilities and coordinates of each face, represented by a feature vector with the following five elements: $[P_{neg}, P_{neutral}, P_{pos}, x, y]$, with P_{neg}, $P_{neutral}$, and P_{pos} the probabilities for the negative, neutral, and positive emotion, respectively, and x and y the center coordinates of the face. The emotion probabilities add up to one and both coordinates are divided by the size of the image in their respective dimension, ensuring that each element is normalized ($[0, 1]$). The emotion information (three feature elements) has an influence that is 1.5 times as large as the location information (two feature elements). Therefore, for the baseline experiments, we multiply both coordinate elements by 1.5 to ensure an equal contribution to the feature vector. An overview of the pipeline from input image to emotional subgroup prediction and performance is shown in Fig 2.

K-means. For each image, we use the Elbow method [Thorndike, 1953] to determine the ideal value of K in an automated fashion. The k-means algorithm, using the implementation of scikit-learn [Pedregosa *et al.*, 2011], then runs until convergence. Each run starts with the same initialization with a fixed random state.

Hierarchical Clustering. For each image, the optimal number of clusters is automatically determined from the re-

[2]https://pypi.org/project/face-recognition/

[3]http://dlib.net/

[4]https://github.com/rcmalli/keras-vggface/blob/master/keras_vggface/vggface.py

Figure 1: Examples of the different annotation methods. The left images show the individual emotion labels, the middle images the social subgroup labels, the right images the emotional subgroup labels. (a) Summation strategy, where emotions and social subgroups are combined, (b) Emotion-based fusion strategy, where two social subgroups with a similar emotion label are fused, (c) Group-based fusion strategy, where a social subgroup with different emotions is fused to an emotional subgroup with the majority emotion. Positive labels are in green (G), neutral labels in yellow (Y), negative labels in red (not shown), social groups in blue. Images retrieved from GAFF 3.0 [Dhall *et al.*, 2018].

sulting dendrogram. This is done by choosing the number of clusters just before the clustering step with the largest distance. The cluster output with the number of clusters found optimal is then used as final output.

Extension Baseline Model

We also investigate the addition of other meaningful features next to emotion and location data, to help the clustering algorithms reach a level closer to human performance. From the different strategies that were deciphered in Section 3.1, it becomes clear that both location and emotion are indeed important. To add spatial information, we include the size of each face (represented by its width) to the feature vector, to indicate distance from the camera. This feature is commonly employed in image analysis to add spatial information [Cerekovic, 2016; Tarasov and Savchenko, 2018]. Additionally, from the self-report study described in Section 3.1, it becomes clear that next to facial expressions and distance, interaction also plays a role when labeling the images. People who share a goal or look in the same direction because of a common object or person of interest, and are therefore part of the same group, are also likely to have a similar gaze direction. We therefore include gaze direction in the feature vector. This approach resembles the visual attention in [Japar *et al.*, 2021] and head pose orientation in [Leach *et al.*, 2014]. We implement this by feeding each face to the Hopenet-Lite implementation[5] of Hopenet [Ruiz *et al.*, 2018], thereby inferring the yaw of the face. The yaw is normalized (divided by 45° to get a value in the interval [0,1]) and added as gaze direction to the feature vector. We experiment with a weight of 1 and 2 for both face size and gaze direction, where multiplication of a feature by this weight alters its relative contribution to the full feature vector.

[5]https://github.com/OverEuro/deep-head-pose-lite

Evaluation Metrics

We employ two separate metrics to decompose the emotional subgroup task into subgroup performance and emotion performance, for a suited analysis of both. These are then combined to form one error function that is comparable across experiments. For comparing the subgroups between the ground truth and the clustering algorithms, we calculate the Hamming distance between two matrices. Let $G_k(V, E_k)$ be the face graph of an image with $k \in \{1, 2\}$, 1 indicating the ground truth annotation and 2 the clustering output to compare it to. V are the nodes, each node representing a face in the image. E_k are the undirected edges $(i - j) \in E_k$ with $i, j \in V$, each edge between two nodes representing those two faces being in the same group. G_k has adjacency matrix $A^{(k)} = [a_{ij}^{(k)}]$ with

$$a_{ij}^{(k)} = \begin{cases} 1, & \text{if } E_k(i-j) \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The Hamming distance between the two matrices is then defined as

$$H_{A^{(1)},A^{(2)}} = \frac{\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} |a_{ij}^{(1)} - a_{ij}^{(2)}|}{N^2}. \tag{2}$$

Where N is the number of vertices per graph, i.e. faces per image, and $A^{(1)}$ and $A^{(2)}$ are the two adjacency matrices to be compared.

For the emotion recognition, we report the accuracy. These two metrics are then combined into an error function as follows:

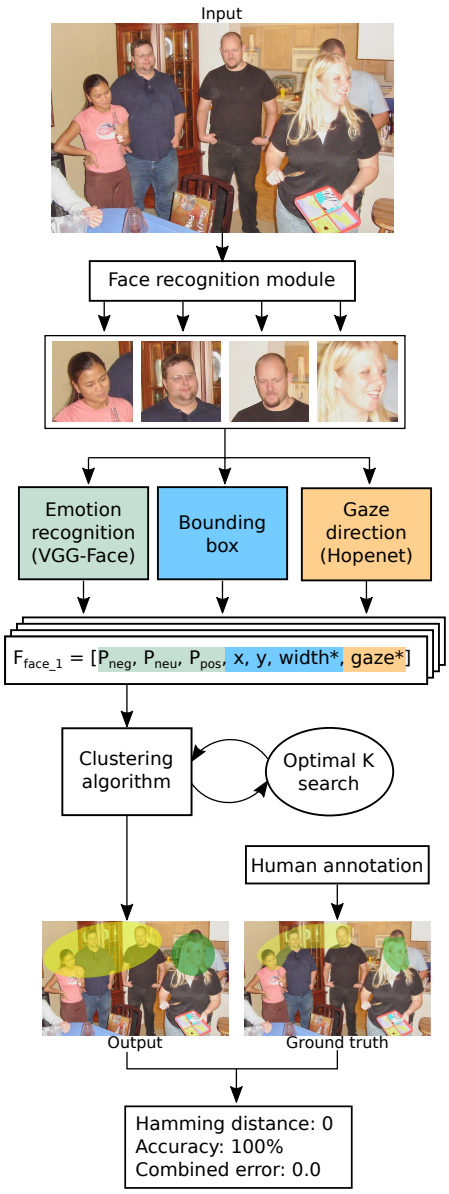

Figure 2: Shown is the pipeline from input image to emotional subgroup prediction and performance. The feature elements with an $*$ are optional elements that are not included for all experiments. Image retrieved from GAFF 3.0 [Dhall *et al.*, 2018].

$$\mathrm{S} = \frac{\left(\sum_{A^{(1)},A^{(2)}=1}^{M} H_{A^{(1)},A^{(2)}}\right)^2 + \left(\sum_{m=1}^{M}(1-\frac{\mathrm{Acc}_m}{100})\right)^2}{M^2}. \tag{3}$$

Where M is the number of images. Taking the square of both individual metrics ensures a higher penalty when one of them shows a poor performance, leading to a combined error where a poor performance can not as easily be compensated for by a good performance on the other metric.

	H	Acc	Error
Unclustered emotions	-	60.8%	-
K-means clustering	0.36	57.2%	0.32
Hierarchical clustering	0.35	56.3%	0.31

Table 2: Performance of the emotion detection model, K-means clustering algorithm and hierarchical clustering algorithm based on the Hamming distance, accuracy and combined error.

4.1 Results

The performances of both K-means and hierarchical clustering are listed in Table 2. The hierarchical clustering algorithm has a slightly lower Hamming distance and accuracy, and a combined error 0.01 lower than K-means.

The emotion accuracy for the emotional subgroups is only meaningful up to the level of the individual emotion accuracy. The emotion accuracy before clustering is 60.8%, meaning that the clustering has little (negative) impact on the emotion recognition performance given the performances in Table 2.

Relative Influence Emotion and Location

The baseline feature vector consists of five elements: $[P_{neg}, P_{neutral}, P_{pos}, x, y]$. Since each element of the vector is normalized ($[0, 1]$), multiplying either the emotion probabilities or the coordinates by a certain weight will increase the relative influence of those feature elements on the clustering output.

First we experiment with different weights for all images together. The relative influence of location and emotion is increased to be 2, 3, 4, or 5 times as high as the other. Performances on the Hamming distance (grouping) and the accuracy (emotions assigned to group members) are shown in Fig 3. Note that the two vertical axes have different scales. The horizontal axis indicates the relative influence of location, meaning that e.g. '2' stands for location being twice as important as emotion, while '0.5' means that location is half as important as emotion. It can be seen that when the influence of location is low, the performances are stable. When the weight of the coordinates increases, performance on the Hamming distance improves, but only slightly (a maximum difference of 0.02). Simultaneously, a decrease of the emotion importance leads to a decrease in performance on the accuracy metric (a maximum difference of 16%).

Extension Baseline Model

Table 3 includes the results on both the separate metrics (Hamming distance, accuracy) and the combined error for all resulting feature vectors. The first three columns show the results for the K-means algorithms. It can be seen that the Hamming distance improves by 0.01 when the face width with weight 2 ($w = 2$) is added to the feature vector. The accuracy is best for the face width with $w = 1$ addition. Combining these two scores indicates that the best performance is gained from the addition of face width with $w = 1$ to the baseline model. Adding the gaze direction worsens the combined error. For hierarchical clustering, the best results are achieved

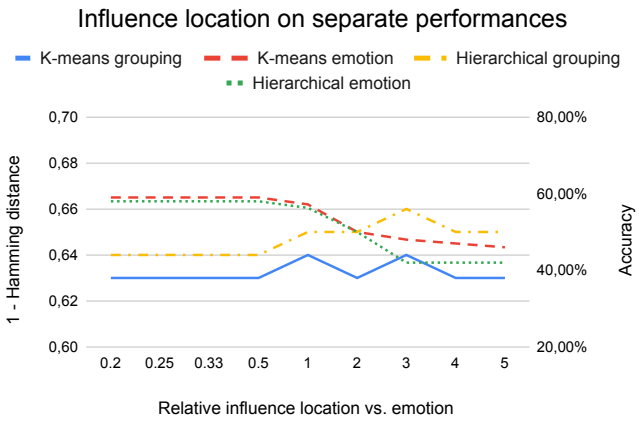

Figure 3: Performance of both K-means and hierarchical clustering, on the inverse Hamming distance for grouping (left y-axis) and the accuracy for emotion labels of the groups (right y-axis), for different weights assigned to the input features. A higher number indicates more influence for the location features.

	H_k	Acc_k	Err_k	H_h	Acc_h	Err_h
Baseline	0.36	57.2%	0.32	0.35	56.3%	0.31
Face w1	0.36	**57.4%**	**0.31**	0.35	56.3%	0.31
Face w2	**0.35**	54.6%	0.33	**0.34**	53.8%	0.33
Gaze w1	0.36	53.3%	0.35	**0.34**	**57.7%**	**0.30**
Gaze w2	0.38	49.3%	0.40	0.37	53.7%	0.35

Table 3: The Hamming distance, accuracy and combined error for K-means (k subscript) and hierarchical clustering (h subscript), for different feature vectors. Added to the baseline feature vector are the face width and gaze direction with weight 1 and 2 (Face w1, Face w2 and Gaze w1, Gaze w2).

by including the gaze direction feature with $w = 1$. This causes a small improvement on both the Hamming distance (0.01 improvement) and the accuracy (1.4% improvement).

4.2 Discussion

Relative Influence Emotion and Location

Fig 3 shows that the Hamming distance only slightly improves when the relative influence of location becomes bigger, whereas accuracy shows a clear improvement when the influence of the emotion probabilities increases. The former shows that emotion on its own is already a relatively solid predictor for emotional group formation, indicating the importance of emotions for emotional subgroup recognition. This is reflected by the self-report of the annotators, mentioning 'Facial expressions' most often to play a role in their assessment.

Extension Baseline Model

Experiments with extra features, summarized in Table 3, show that adding face size with $w = 2$ gives a minor improvement on the Hamming distance for both K-means and hierarchical clustering, while at the same time leading to a small drop in accuracy. The drop in accuracy, which is also shown for most other added features, can be explained by the increased emphasis put on the position of the face rather than the emotional expression. A similar pattern emerged from increasing the relative weight of the location features, as shown in Fig 3. For K-means the best results come from the model that adds face size to the baseline input features, albeit with an accuracy that is virtually indifferent from that of the baseline. For hierarchical clustering, including information about the gaze direction (yaw) of each face results in a better performance on both the Hamming distance and accuracy. This signifies that gaze direction is useful for clustering emotional subgroups, as is reflected in the self-report answers 'interaction' and 'body language'. A possible explanation for the fact that this improvement is not observable for K-means is its hyperspherical, convex nature, as was described in Section 4. Adding gaze direction to emotion and location information may reduce the spherical forms of the clusters, something that is not relevant for hierarchical clustering, therefore limiting the added value of gaze direction.

Throughout the experiments, K-means and hierarchical clustering show similar results as is shown in Table 3. These results tell us that both clustering methods are suited for the task. Future work should monitor if the convex nature of K-means still remains an advantage when further exploring meaningful features and investigating extended datasets, or if hierarchical clustering proves to be better for an up-scaled version of the task.

5 Conclusion

In this research we identify three clear, different strategies that are employed by human annotators to recognize emotional subgroups: the summation strategy, the emotion-based fusion strategy, and the group-based fusion strategy. Images that show agreement among annotators, are most often those that elicit the use of the summation strategy, while images with partial agreement more often elicit the use of the emotion-based fusion (putting more emphasis on emotion than social groups) or the group-based fusion (putting more emphasis on social groups than on emotion) strategy. Based on these strategies, we apply K-means and hierarchical clustering for the automatic recognition of emotional subgroups by clustering emotion features, retrieved from a neural network trained on emotion recognition, together with location features of each face. Experimenting with different additional features suggests, with a modest performance improvement, that face size and gaze direction contain meaningful information for k-means and hierarchical clustering, respectively. Both clustering algorithms show similar results. The hyperspherical, convex nature of K-means may become limiting when additional features lead to changes in the cluster shapes. The best performance, when including both full and partial agreement images, comes from the addition of gaze direction to the feature vector fed to the hierarchical clustering algorithm. This leads to a Hamming distance of 0.34, an accuracy of 57.7%, and a combined error of 0.30. This shows that the task of emotional subgroup recognition is a complex one, but also that a relatively small feature vector is already able to reasonably represent human perception.

Future Work. Future research could focus on the different strategies that are employed when people are asked to distinguish emotional subgroups in an image. To do so, it is important to increase the number of images and include different situations and emotions in them, to further explore what motivates the use of a particular strategy. Given enough data, other classification strategies that make use of more complex features could be used for the task. This could also be extended to different modalities, for example by including motion information or audio from video data.

Acknowledgements

We thank SURFsara (www.surfsara.nl) for the support in using the Lisa Compute Cluster. This work is part of the research programme Innovational Research Incentives Scheme Vidi SSH 2017 with project number 016.Vidi.185.178, which is financed by the Dutch Research Council (NWO).

References

[Badr *et al.*, 2015] Hamada Badr, Benjamin Zaitchik, and Amin Dezfuli. A tool for hierarchical climate regionalization. *Earth Sci. Inform.*, 8(4):949–958, 2015.

[Cerekovic, 2016] Aleksandra Cerekovic. A deep look into group happiness prediction from images. In *ICMI 2016 - Proc. 2016 ACM Int. Conf. Multim. Interact.*, pages 437–444, 2016.

[Dhall *et al.*, 2012] Abhinav Dhall, Jyoti Joshi, Ibrahim Radwan, and Roland Goecke. Finding happiest moments in a social context. In *Asian Conference on Computer Vision*, pages 613–626. Springer, 2012.

[Dhall *et al.*, 2016] Abhinav Dhall, Roland Goecke, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. In *ICMI 2016 - Proc. 2016 ACM Int. Conf. Multim. Interact.*, pages 427–432, 2016.

[Dhall *et al.*, 2018] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *ICMI 2018 - Proc. 2018 ACM Int. Conf. Multim. Interact.*, pages 653–656, 2018.

[Frank, 1995] Kenneth Frank. Identifying cohesive subgroups. *Social networks*, 17(1):27–56, 1995.

[Guo *et al.*, 2018] Xin Guo, Bin Zhu, Luisa Polanía, Charles Boncelet, and Kenneth Barner. Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions. In *ICMI 2018 - Proc. 2018 ACM Int. Conf. Multim. Interact.*, pages 635–639, 2018.

[Japar *et al.*, 2021] Nurul Japar, Ven Jyn Kok, and Chee Seng Chan. Coherent group detection in still image. *Multimed. Tools Appl.*, pages 1–20, 2021.

[Kosti *et al.*, 2019] Ronak Kosti, Jose Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE Trans. Pattern Analys. Mach. Intell.*, 2019.

[Labelbox, 2021] Labelbox. https://labelbox.com, 2021. Accessed: 2022-06-09.

[Leach *et al.*, 2014] Michael Leach, Rolf Baxter, Neil Robertson, and Ed Sparks. Detecting social groups in crowded surveillance videos using visual attention. In *Proc. IEEE Conf. Comput. Comput. Vis. Pattern Recognit. Workshops*, pages 461–467, 2014.

[Li *et al.*, 2016] Jianshu Li, Sujoy Roy, Jiashi Feng, and Terence Sim. Happiness level prediction with sequential inputs via multiple regressions. In *ICMI 2016 - Proc. 2016 ACM Int. Conf. Multim. Interact.*, pages 487–493, 2016.

[Li *et al.*, 2017] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2852–2861, 2017.

[Light, 1971] Richard Light. Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological bulletin*, 76(5):365, 1971.

[MacQueen and others, 1967] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symp. Math. Stat. Prob.*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[Mou *et al.*, 2019] Wenxuan Mou, Hatice Gunes, and Ioannis Patras. Your fellows matter: Affect analysis across subjects in group videos. In *2019 IEEE Int. Conf. Aut. Face Gest. Rec. FG 2019*, pages 1–5. IEEE, 2019.

[Noordewier and Breugelmans, 2013] Marret Noordewier and Seger Breugelmans. On the valence of surprise. *Cognition & emotion*, 27(7):1326–1334, 2013.

[Parkhi *et al.*, 2015] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proc. Brit. Mach. Vis. Conf. (BMVC)*, pages 41.1–41.12. BMVA Press, 2015.

[Pedregosa *et al.*, 2011] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[Ruiz *et al.*, 2018] Nataniel Ruiz, Eunji Chong, and James Rehg. Fine-grained head pose estimation without keypoints. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, June 2018.

[Tan *et al.*, 2017] Lianzhi Tan, Kaipeng Zhang, Kai Wang, Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Group emotion recognition with individual facial emotion cnns and global image based cnns. In *ICMI 2017 - Proc. 2017 ACM Int. Conf. Multim. Interact.*, pages 549–552, 2017.

[Tarasov and Savchenko, 2018] Alexander Tarasov and Andrey Savchenko. Emotion recognition of a group of people in video analytics using deep off-the-shelf image embeddings. In *Int. Conf. Analys. Images Soc. Netw. Texts*, pages 191–198. Springer, 2018.

[Thorndike, 1953] Robert Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[Vascon *et al.*, 2016] Sebastiano Vascon, Eyasu Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. Detecting conversational groups in images and sequences: A robust game-theoretic approach. *Comput. Vis. Image Underst.*, 143:11–24, 2016.

[Veltmeijer *et al.*, 2021] Emmeke Veltmeijer, Charlotte Gerritsen, and Koen Hindriks. Automatic emotion recognition for groups: a review. *IEEE Transactions on Affective Computing*, 2021.

[Wang *et al.*, 2018] Qi Wang, Mulin Chen, Feiping Nie, and Xuelong Li. Detecting coherent groups in crowd scenes by multiview clustering. *IEEE Trans. Pattern Analys. Mach. Intell.*, 42(1):46–58, 2018.

[Ward Jr, 1963] Joe Ward Jr. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, 58(301):236–244, 1963.

[Yücel *et al.*, 2013] Zeynep Yücel, Francesco Zanlungo, Tetsushi Ikeda, Takahiro Miyashita, and Norihiro Hagita. Deciphering the crowd: Modeling and identification of pedestrian group motion. *Sensors*, 13(1):875–897, 2013.

[Zhang *et al.*, 2019] Meng Zhang, Xinchen Liu, Wu Liu, Anfu Zhou, Huadong Ma, and Tao Mei. Multi-granularity reasoning for social relation recognition from images. In *2019 IEEE Int. Conf. Multim. Expo (ICME)*, pages 1618–1623. IEEE, 2019.