A. GroupWalk

A.1. Annotation Procedure

We present the human annotated GroupWalk data set which consists of 45 videos captured using stationary cameras in 8 real-world setting including a hospital entrance, an institutional building, a bus stop, a train station, and a marketplace, a tourist attraction, a shopping place and more. 10 annotators annotated 3544 agents with clearly visible faces and gaits across all videos. They were allowed to view the videos as many times as they wanted and had to categorise the emotion they perceived looking at the agent into 7 categories - ”Somewhat Happy”, ”Extremely Happy”, ”Somewhat Sad”, Extremely Sad”, ”Somewhat Angry”, ”Extremely Angry”, ”Neutral”. In addition to perceived emotions, the annotators were also asked to annotate the agents in terms of dominance (5 categories- ”Somewhat Submissive”, ”Extremely Submissive”, ”Somewhat Dominant”, ”Extremely Dominant”, ”Neutral”) and friendliness (5 categories- ”Somewhat Friendly”, ”Extremely Friendly”, ”Somewhat Unfriendly”, ”Extremely Unfriendly”, ”Neutral”). Attempts to build the dataset are still ongoing.

For the sake of completeness, just like our analysis in Section **??**, we show the friendliness label distribution and dominance label distribution for every annotator in Figure 7 and Figure 8 respectively.

A.2. Labels Processing

4 major labels that have been considered are Angry, Happy, Neutral and Sad. As described in Section A.1, one can observe that the annotations are either ”Extreme” or ”Somewhat” variants of these major labels (except Neutral). Target labels were now generated for each agent. Each of them are of the size 1 x 4 with the 4 columns representing the 4 emotions being considered and are initially all 0. For a particular agent id, if the annotation by an annotator was an ”Extreme” variant of Happy, Sad or Angry, 2 was added to the number in the column representing the corresponding major label. Otherwise for all the other cases, 1 was added to the number in the column representing the corresponding major label. Once we have gone through the entire dataset, we normalize the target label vector so that vector is a combination of only 1s and 0s.

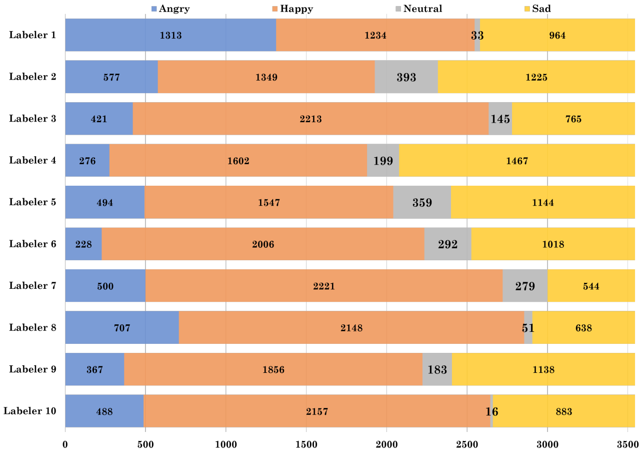

Figure 5: **Annotator Annotations of GroupWalkDataset:** We depict the emotion class labels for GroupWalk by 10 annotators. A total of 3544 agents were annotated from 45 videos.

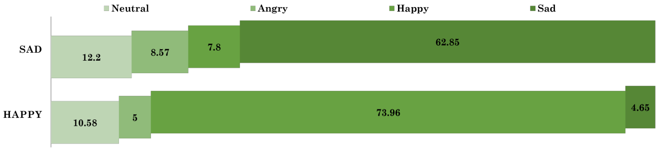

Figure 6: **Annotator Agreement/Disagreement:** For two emotion classes (Happy and Sad), we depict the trend of annotator disagreement.

B. EmotiCon on IEMOCAP Dataset

To validate that EmotiCon can be generalised for any number of modalities, we report our performance on IEMOCAP [9] in Table 4. IEMOCAP dataset consists of speech, text and face modalities of 10 actors recorded in the form of conversations (both spontaneous and scripted) using a Motion Capture Camera. The labeled annotations consist of 4 emotions – angry, happy, neutral, and sad. This is a single-label classification as opposed to multi-label classification we reported for EMOTIC and GroupWalk. Because of this we choose to report mean classification accuracies rather than AP scores. Most prior work which have shown results on IEMOCAP dataset, report mean classification accuracies too.

Labels	Kosti et al.[27]	Zhang et al.[58]	Lee et al.[30]	**EmotiCon**	
				GCN-Based	Depth-Based
Anger	80.7%	-	77.3%	87.2%	**88.2%**
Happy	78.9%	-	72.4%	82.4%	**83.4%**
Neutral	73.5%	-	62.8%	75.5%	**77.5%**
Sad	81.3%	-	68.7%	88.2%	**88.9%**
mAP	78.6%	-	70.3%	83.4%	**84.5%**

Table 4: **IEMOCAP Experiments:** Mean Classification Accuracies for IEMOCAP Dataset.

As can be seen from the Table 4, there is not a significant improvement in the accuracy, 84.5% as SOTA works, not essentially based on context have reported an accuracy of

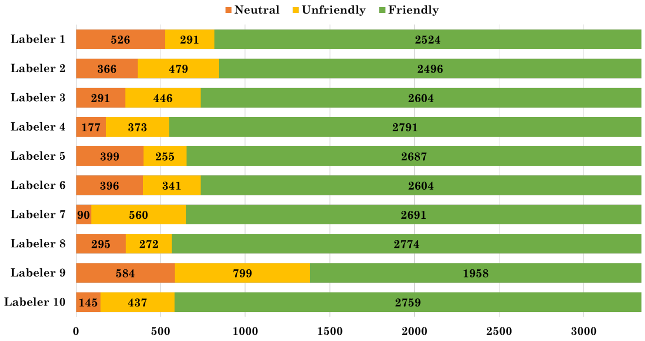

Figure 7: **Friendliness Labeler Annotations:** We depict the friendliness labels for by 10 labelers. A total of 3341 agents were annotated from 45 videos.

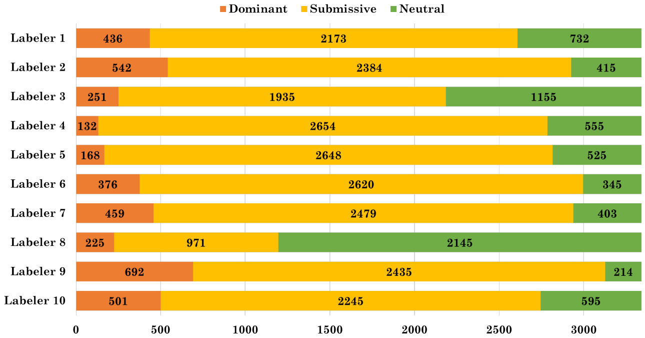

Figure 8: **Dominance Labeler Annotations:** We depict the dominance labels for by 10 labelers. A total of 3341 agents were annotated from 45 videos.

82.7%. We believe that the controlled settings in which the dataset is collected, with minimal context information results in not huge improvements. Moreover we also see that prior works in context, Kosti et al. [27] and Lee et al. [58] sort of do not get any context to learn from and hence do not perform so well. Even EmotiCon's performance is a result of incorporating modalities, with small contribution from context.