An Approach at Social Relation Recognition in Egocentric Videos using Pose Estimation

MAARTEN MEIJER, University of Twente, The Netherlands

In this paper, a literature review is conducted and a framework is analysed for the automatic recognition of social relations in egocentric videos. Background research is conducted, where different methods of recognising social relations in videos are compared. An approach is proposed where the supervised machine learning models *Nearest Neighbors*, *Linear SVM* [14], *RBF SVM*, *Gaussian Process*, *Decision Tree*, *Random Forest* [8], *Neural Net*, *AdaBoost* [11], *Naive Bayes* and *QDA* are applied to data generated by human pose estimation and tracking [15]. Finally, various experiments are done to validate the proposed features and classifiers.

CCS Concepts: • **Computing methodologies** → **Object detection**; **Activity recognition and understanding**.

Additional Key Words and Phrases: deep learning, computer vision, egocentric videos, pose estimation, social relation recognition, supervised machine learning

1 INTRODUCTION

Digital technologies are getting involved in our social life more and more every day. As the world is getting more digitised, enormous amounts of multimedia data are becoming available to use in technologies. Wearable head-mounted cameras have become more popular, providing the possibility to use video data from a first-person, egocentric perspective. This rise of egocentric videos and the need for a better understanding of our social relations and interactions [7] has motivated many different studies in the automated analysis of egocentric videos using computer vision [5]. With this rise of new data, the possibility emerges to recognise social relations. This possibility has been the point of focus in many pieces of research in the last decade [1]. The identification of different types of social relations proves to be crucial in multiple domains. According to the studies [4], [39] and [12] mobile technologies that provide a precise assessment of human behaviour lead to an improved mental health. There have been different approaches to recognise social relations, most methods use deep learning to identify certain features of a video and then use these features to predict a type of social relation.

This paper will include - in order - a problem statement, background research, methodologies on how the research was conducted, the results from these experiments, a conclusion and lastly, recommendations for future work built on this research.

2 PROBLEM STATEMENT

With different kinds of research already being done on recognising social relations, there has still lacked an approach using both pose estimation [15] and supervised machine learning models. To research if this approach could perform as well as, or better than, current state-of-the-art approaches [16], or if the approach could help improve these approaches, the following research question is proposed.

> **RQ:** *Can supervised machine learning models help us in the recognition of social relations in egocentric videos when supplied with human pose estimation and tracking data?*

When trying to answer this question, it has been useful to first answer the following sub-questions and afterwards concluding an answer to the research question.

> **SQ1:** *Can we use pose estimation and pose tracking to create useful features?*
>
> **SQ2:** *Which extracted features are the most important?*
>
> **SQ3:** *Can we optimise the performance of supervised machine learning models for our specific features?*

These sub-questions have been used to guide the research in a structural manner and finally answer the main research question.

3 RELATED WORK

3.1 Dataset preparation

A comprehensive dataset is needed to recognise social relations using computer vision. This dataset does not only need to contain enough normalised data, but also well-defined labels. Different studies have not agreed upon the most correct definition of labels in the social relations field. While on the one hand, Aghaei *et al.* [1] provided an approach to only recognise the labels of formal and informal meetings, the approach of Sun *et al.* [34] was able to recognise 16 different labels, among which grandpa-grandchild, teacher-student and sports team members. These 16 labels were grouped in 5 different domains based on Bugental's social psychology theory [9]. Another study by Aimar *et al.* [2] based their labels on these same groups but discarded seven insufficiently represented labels and left nine well-presented labels. Several studies regarding social relation recognition in movies also concern the same labels as in the egocentric domain. The Video Multiple-Relation dataset, proposed by Liu *et al.* in [23], groups 105 relationships into nine classes, these nine showing some overlap with [2], but also differences. The relationships proposed by [23] are shown in Fig. 1. The ViSR dataset [22] defines eight types of social relations, also grouped in the five suggested domains. Most studies base their labels on the five social domains defined by Bugental in [9]. These five social domains are *attachment*, *coalitional group*, *mating*, *reciprocity* and *hierarchical power*.

3.2 Social relation recognition

With a labelled dataset, there have been many approaches to create a framework to detect social relations. One particularly seminal

TScIT 37, July 8, 2022, Enschede, The Netherlands

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

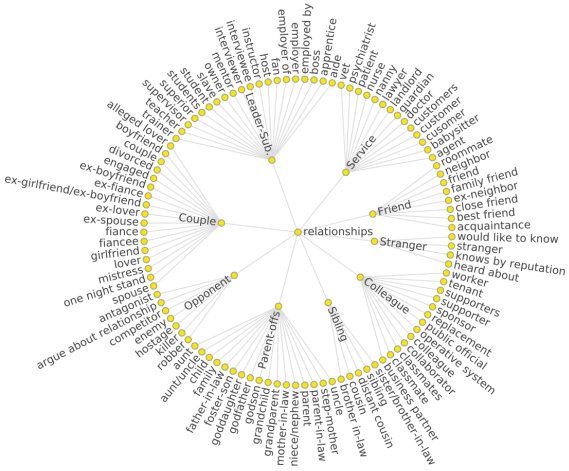

Fig. 1. The nine relationships as defined by Liu *et al.* in [23]

study is the study by Sun *et al.* [34], where two different types of approaches to models were compared and experimented on.

The first type was Convolutional Neural Networks (CNN), [35] with images as input and a social relation decision as an output. These end-to-end models learn about social relations solely by finding patterns in data using deep learning. Various improvements have been made with these types of state-of-the-art techniques over the past years. Sun *et al.* [34] have investigated a double-stream CaffeNet model, extracting data from both heads and bodies in images.

A much better approach to building a social relation recognition model is first predicting intermediate semantic attributes, after which concluding a social relation decision based on the former. According to Sun *et al.* [34], this method is attractive because the researcher can derive which semantic attributes are essential and include these into the model. These semantic attributes can be concluded by using social domain theory. Many studies investigating this second approach use CNN for the first stage, but there is no general agreement on what to use for the later stage of the model. Aimar *et al.* [2] used CNN models to extract intermediate semantic attributes. With these attributes as input, multiple strategies to classify social relations were explored, namely the *Single-task* (based on LSTM [19]), the *Multi-task Top-down* (inspired by the hierarchical approach as proposed in [10]) and the *Multi-task Independent* strategy (based on multi-task learning [30]). Sun *et al.* [34] chose semantic attributes based on the attribute categories mentioned in the definitions of social domains; age, gender, location, head appearance, head pose, clothing, proximity and activity. A novel approach was the approach from Felicioni *et al.* [16], who researched building a graph with CNN data and from that inferring the category of a social relation. The research from Liu *et al.* [23] also shows particularly state-of-the-art methods to extract semantic features and fuse them into a single decision using *Multi-Conv Attention* modules and *Multilayer Perceptrons*.

4 METHODOLOGIES

To properly research and analyse the technology to recognise social relations, a refined methodology has been constructed. In this section, this methodology is explained and validated.

4.1 Dataset analysis

This research builds upon work from various others. A dataset of egocentric videos is needed to experiment with constructing a technical framework. For this, the Ego4D dataset [17] was used. In the research, the videos from the Ego4D dataset have been refined, restructured and labelled so that they could be used effectively in the later stages of the research. Mainly the same labels have been used as proposed in [23], but three labels have been excluded due to insignificant proportions in the Ego4D dataset. The labels *leader-sub*, *sibling* and *opponent* were left out, leaving only the labels *couple*, *service*, *friend*, *stranger*, *colleague* and *parents-offspring*.

4.2 Feature extraction

The pose estimation framework AlphaPose [15] was used to recognise human bodies in videos. The framework works on video frame inputs and detects where certain limbs of individuals are and how they connect to each other. With this framework, published models have been used that are trained on the exhaustive and complete COCO dataset [21]. The AlphaPose framework allows different methods to track the human poses. In the research, an approach using Human-ReID [41] [33] and an approach using PoseFlow [40] were compared. Both tracking modules use information from the pose estimation of previous frames to infer to which person different data points belong. With this information, features have been constructed based on individual persons.

4.3 Social relation prediction

Next in the pipeline, a component was created to recognise social relations in videos. To do this, multiple descriptors - or features - have been extracted from the data that the deep learning model described above provides. The descriptors are based on the position and movement of individuals in the egocentric videos. They are generated by accumulating individuals' body data points over a video and calculating the mean and standard deviation for each person. With these descriptors, experiments have been done to measure which combination of descriptors performs the best when applying supervised machine learning on them. The different supervised machine learning models [36, 38] that are compared in this study are *Nearest Neighbors*, *Linear SVM* [14], *RBF SVM*, *Gaussian Process*, *Decision Tree*, *Random Forest* [8], *Neural Net*, *AdaBoost* [11], *Naive Bayes* and *QDA*.

4.4 Validation

To understand different models, it is helpful to visualise the feature space using the dimensionality reduction algorithms PCA and t-SNE [13]. First, PCA was used to reduce the dimensionality to 50, and afterwards, t-SNE was used to visualise the feature space in two dimensions. This technique was chosen because of the complexity of the t-SNE algorithm and the linearity of PCA [28].

Next to analysing the feature space, the different models were trained and compared. A well-used metric when comparing the effectiveness of a supervised machine learning model is the accuracy of the model. A model's accuracy is defined by the number of correct predictions divided by the number of total predictions. However, when you have an imbalanced dataset, the metric does not reflect the effectiveness of the model as well since it might be very good at one of the classes but not the others. A better metric is the F1 score [37], which was used in the research to help determine the most effective models. The definition of the F1 score is shown in Eq. 1.

$$F1 = \frac{\#\text{ true positive}}{\#\text{true positive} + \frac{1}{2} \cdot (\#\text{ false positive} + \#\text{ false negative})} \quad (1)$$

4.5 Feature importance

Finding the most imoprtant features in a dataset can be useful, since it means that a more efficient model could be constructed when only using the important features. To find out the most important features, this research uses two different methods. One method only works on Random Forests and is based on mean decrease in impurity [24]. It only works on Random Forests because it relies on measuring the impurity among different trees [6, 31]. The other method is based on feature permutation and works on all models [3]. This method was applied to the model with the best results.

5 RESULTS

5.1 Dataset analysis

The Ego4D dataset [17] is a remarkably complete dataset, providing enough data for all kinds of applications. It has 3,670 hours of video material in total, 47.7 hours of which have been released that show social interaction. One downside of the dataset is that it does not include any annotations relevant to this research. The dataset consists of videos on many different social interactions, showing people with different social relations to the camera-wielding individual. The annotations are made by hand, where a relation was chosen for a specific duration of the video. These particular durations can be viewed as sub-clips of the whole video, where a certain relation can be recognised. A few example frames from videos in the Ego4D dataset have been shown in Fig. 2, annotated with six of the nine categories defined by Liu *et al.* [22]. The other three social relations were found to be represented insignificantly in the Ego4D dataset and have thus not been considered in this study. In Table 1, the composition of this dataset can be seen. The dataset has shown to be somewhat imbalanced, which needs to be considered when analysing the trained classifiers.

5.2 Human pose extraction

AlphaPose [15] runs using PyTorch [27] and produces good results compared to other state-of-the-art solutions [15]. When running Deep Learning models, GPU accelerated calculations speed up the process significantly [27], creating the need for a powerful computer. To account for this need, Google Colab [1] was used, where Tesla K80 GPU's and 13GB RAM are provided for free.

[1] https://colab.research.google.com/

Social Relation	Amount of videos
Friend	484
Stranger	109
Service	135
Colleague	103
Parent-Offspring	54
Couple	106

Table 1. Composition of the 996 sub-clips made from videos in the Ego4D dataset regarding social relations

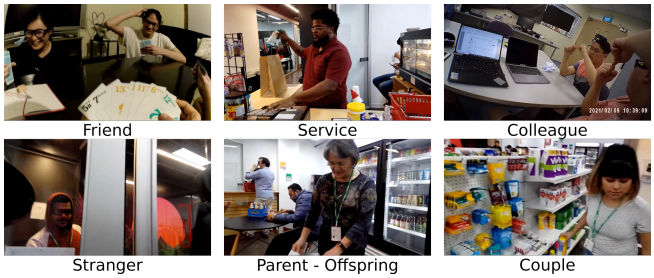

Fig. 2. Example frames from videos in the Ego4D dataset

Fig. 3. Examples of pose estimation and tracking on frames from the Ego4D dataset

The chosen model was created by Fang *et al.* , and is based on ResNet50 [18] and YOLOv3 [29]. The model achieves an average precision of 72.0 on the COCO [21] dataset. Together with this pose estimation model, a tracking module based on Human-ReID techniques [33, 41] was used. The model used for this tracking is based on OSNet, which has shown good accuracy [42]. Another tracking module - PoseFlow [40] - was tried, but it performed less accurately since it provides a MOTP of 67.8. In Fig. 3, a threefold examples can be observed, which show the output from the pose estimation and tracking used. You can see that every person has a unique colour, indicating the difference detected between persons through multiple frames.

5.3 Feature extraction

Having acquired the information about different people's movements through videos, this data can be used to train different machine learning models. The extracted features can be viewed in Table 3. In the research, the data collected by the Deep Learning models, as explained in the previous section, have been transformed and restructured to function as the input data for supervised machine

Classifier	Scaler	Accuracy score	Balanced accuracy score	F1 score (weighted)
3 Nearest Neighbors	Standard	0.371 ±0.028	0.210 ±0.016	0.339 ±0.022
Linear SVM	none	0.394 ±0.034	0.217 ±0.030	0.348 ±0.031
RBF SVM	MinMax	0.406 ±0.029	0.192 ±0.014	0.323 ±0.020
Gaussian Process	MinMax	0.402 ±0.033	0.195 ±0.015	0.320 ±0.015
Decision Tree	MinMax	0.360 ±0.036	0.175 ±0.016	0.292 ±0.020
Random Forest	Standard	**0.423 ±0.007**	0.176 ±0.008	0.281 ±0.016
Neural Net	Standard	0.413 ±0.040	**0.250 ±0.033**	**0.388 ±0.035**
AdaBoost	none	0.230 ±0.045	0.174 ±0.030	0.271 ±0.030
Naive Bayes	none	0.184 ±0.017	0.194 ±0.045	0.202 ±0.019
QDA	none	0.392 ±0.059	0.216 ±0.038	0.331 ±0.047

Table 2. The accuracy, balanced accuracy and F1 score of different classifiers using SMOTE and different types of scalers

learning models. For every body joint feature from the pose estimation and tracking model, the mean and standard deviation (both for the x and y coordinate) were calculated for a single person. Also, the confidence score of a person throughout a video is accumulated, and the mean and standard deviation are used as features. In total, this creates 70 features. These features are the input attributes for the supervised machine learning models discussed in the next section. After this rearrangement of the data, 1428 relations were extracted from sub-clips. Due to time constraints, not all videos have been used in this process, but it is expected that the released part of the social domain of the Ego4D dataset could provide five times as many relations in total. The research aims to analyse this minor part of the dataset to draw conclusions that also hold for the rest of the dataset. In Table 4, the composition of these extracted relations can be examined.

Feature
Nose
Left eye
Right eye
Left ear
Right ear
Left shoulder
Right shoulder
Left elbow
Right elbow
Left wrist
Right wrist
Left hip
Right hip
Left knee
Right knee
Left ankle
Right ankle
Confidence score of person

Table 3. Features that are extracted from pose estimation

5.4 Supervised machine learning

When using supervised machine learning on this transformed dataset, it is essential to critically analyse different types of models and compare the results they provide. In this research, the ten classifier types *Nearest Neighbors*, *Linear SVM*, *RBF SVM*, *Gaussian Process*, *Decision Tree*, *Random Forest*, *Neural Net*, *AdaBoost*, *Naive Bayes* and *QDA* were chosen to compare. This is because these are the most promising and most used classifier types at the time [26, 32]. The models were implemented in the Python library Scikit-learn [28]. Because some models perform better with standardised data [25], the data was normalised with two different methods. For some models, a StandardScaler (scaling the data in a normal distribution from -1 to 1) performed better, and for others, a MinMaxScaler (scaling the data in a range from 0 to 1) performed better. Because of the imbalance that was shown to be in the dataset, two different over-sampler methods were tried, namely the popular SMOTE and ADASYN [20]. No difference was found between the two over-sampler methods. Therefore, SMOTE was chosen to use for all metrics. In Table 2, the metrics accuracy, balanced accuracy and F1 score of the classifiers can be compared. For some models, the StandardScaler performed better. For others, the MinMaxScaler and for others, neither. The metrics are shown for the classifier with the scaler that performed best for that classifier. All metrics are calculated with 5-fold cross-validation, which means that the metric is calculated five times on different subsets of the whole dataset. This is done to create a more reliable metric and show the deviation of the metric. The *Neural Net*, *Linear SVM* and *3 Nearest Neighbors* classifiers showed the most promising results in these metrics, proving to have the highest F1 scores compared to the other models. To compare the different results for all the researched classifiers, confusion matrices have been constructed and are shown in Appendix A. These matrices provide a way of visualising which classes are predicted when certain classes are the truth.

5.5 Model analysis

The extracted features been analysed using the dimensionality reduction algorithms PCA and t-SNE [13]. To better visualise the feature space, SMOTE was [20] applied on the dataset beforehand. The analysis has been done to visualise the separability of different data points in a high-dimensional space. In Fig. 4, the result can be perceived. Firstly, PCA was used to reduce the dimensionality to 50.

Social Relation	Amount
Friend	604
Stranger	272
Service	192
Colleague	117
Parent-Offspring	12
Couple	231

Table 4. Composition of relations extracted from sub-clips in the Ego4D dataset

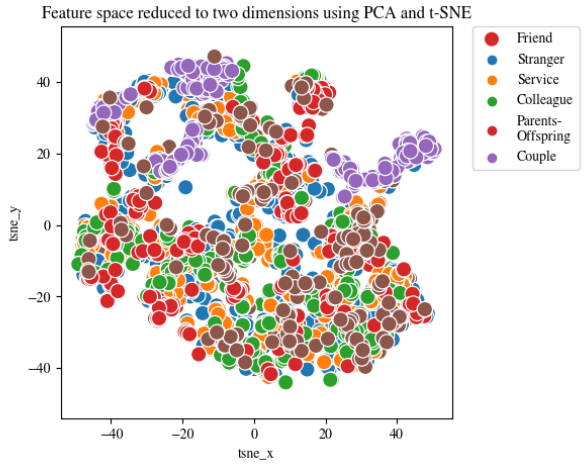

Fig. 4. Dimensionality reduction on the feature space using PCA and t-SNE combined with SMOTE

After this, t-SNE was used to further reduce the dimensionality to 2. Different perplexity values for t-SNE have been tried, where the value 50 showed the best results. It can be concluded that the model does not show apparent distinctions between relations since the dimensionality-reduced features also do not show clear distinctions between each other. No groups were formed in the t-SNE graphs, but most relations are scattered around the two-dimensional space. This indicates that there is no clear correlation between the features implemented and the corresponding relations.

5.6 Feature importance

As discussed in the section 4, two different methods have been applied to find out if some of the extracted features are more important than others. The results to the MDI [24] and the feature permutation [3] techniques are presented in Appendix B. The first technique uses the *Random Forest* classifier, since that is required from the technique. The second technique was applied to the *Neural Net* classifier, since it proved to have the best results of the examined classifiers. Among all features, no features clearly stood out to be more important.

6 CONCLUSION AND DISCUSSION

Having analysed the data and various supervised machine learning models, there are many things to conclude. First of all, the dataset shows a clear imbalance where the 'Friend' relation is represented excessively compared to the other relations. This imbalance is also represented in the fact that most supervised machine learning models analysed predicted some relations falsely as being a 'Friend' relation. One technique has been tried to remove this bias in models by over-sampling the dataset. This technique did not show any improvements in the models, except for the *Neural Net* classifier. Even though the models showed less bias towards the 'Friend' label, the overall F1 score and balanced accuracy did not improve. Furthermore, metrics were used that take into account this inherent imbalance in the dataset, like balanced accuracy and F1 score.

With the results from the feature importance techniques, no clear conclusion could be drawn. Both of the techniques showed similar results. No features showed to be significantly more important than others.

We can conclude that of all examined models, the *Neural Net*, *Linear SVM* and *3 Nearest Neighbors* showed the most promising results of all. These models had the best F1 score compared to the other models using the same features, and a clear diagonal can be seen in the confusion matrix of these classifiers.

Overall, it seems that the extracted features don't show enough significance in recognising social relations to create a state-of-the-art model since only an F1 score of 38.8% was achieved. The work by Liu *et al.* [23] resulted in an F1 score of 46.7%. These F1 scores do not necessarily relate to each other since the work by Liu *et al.* uses far more data and a different type of data, namely from movies. Also, the work by Liu *et al.* uses a multi-modal approach, meaning it uses the best parts of different models to combine into one approach. The research conducted in this study could very well be used in future studies to improve a social relation recognition model, but on its own it does not perform conforming to the state-of-the-art.

7 FUTURE WORK

Since the Scikit-learn [28] classifier *Neural Net* performed the best amongst the models tested, it makes sense that Convolutional Neural Networks should be tried as well. In the literature study, it was found that a lot of studies use CNN's to predict social relations. In a future study, different types of Neural Nets could be examined when using the same input data as this study.

Furthermore, it is believed that this input data could prove to be of significant use when combining the models discussed in this research with other research and thus creating a multi-modal approach to the problem. It is believed that the work of this study has the potential to improve current multi-modal state-of-the-art approaches.

A point where this study has lacked is in the size of the data set. Due to time constraints, only parts of the data set have been used. In future studies, it would improve performance if more data is used.

REFERENCES

[1] Maedeh Aghaei, Mariella Dimiccoli, Cristian Canton Ferrer, and Petia Radeva. 2017. Towards social pattern characterization in egocentric photo-streams. (9 2017). https://doi.org/10.48550/arxiv.1709.01424

[2] Emanuel Sanchez Aimar, Petia Radeva, and Mariella DImiccoli. 2019. Social Relation Recognition in Egocentric Photostreams. In *Proceedings - International Conference on Image Processing, ICIP*, Vol. 2019-Septe. IEEE Computer Society, 3227–3231. https://doi.org/10.1109/ICIP.2019.8803634

[3] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. 2010. Permutation importance: A corrected feature importance measure. *Bioinformatics* 26, 10 (2010). https://doi.org/10.1093/bioinformatics/btq134

[4] Min Hane Aung, Mark Matthews, and Tanzeem Choudhury. 2017. Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies. , 603–609 pages. https://doi.org/10.1002/da.22646

[5] Alejandro Betancourt, Pietro Morerio, Carlo S. Regazzoni, and Matthias Rauterberg. 2014. The Evolution of First Person Vision Methods: A Survey. (9 2014). https://doi.org/10.1109/tcsvt.2015.2409731

[6] Gérard Biau and Erwan Scornet. 2016. A random forest guided tour. *Test* 25, 2 (2016). https://doi.org/10.1007/s11749-016-0481-7

[7] Marc Bolaños, Mariella Dimiccoli, and Petia Radeva. 2015. Towards Storytelling from Visual Lifelogging: An Overview. (7 2015). https://doi.org/10.1109/thms.2016.2616296

[8] Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001). https://doi.org/10.1023/A:1010933404324

[9] Daphne Blunt Bugental. 2000. Acquisition of the Algorithms of Social Life: A Domain-Based Approach. *Psychological Bulletin* 126, 2 (2000), 187–219. https://doi.org/10.1037/0033-2909.126.2.187

[10] Ricardo Cerri, Rodrigo C. Barros, André C. P. L. F. de Carvalho, and Yaochu Jin. 2016. Reduction strategies for hierarchical multi-label classification in protein function prediction. *BMC Bioinformatics* 17, 1 (12 2016), 373. https://doi.org/10.1186/s12859-016-1232-1

[11] Tu Chengsheng, Liu Huacheng, and Xu Bing. 2017. AdaBoost typical Algorithm and its application research. *MATEC Web of Conferences* 139 (6 2017), 222. https://doi.org/10.1051/matecconf/201713900222

[12] Philip Chow, Wes Bonelli, Bethany Teachman, Haoyi Xiong, Karl Fua, Bethany A Teachman, and Laura E Barnes. 2016. *SAD: Social Anxiety and Depression Monitoring System for College Students Assessing Mental Stress Based on Smartphone Sensing Data: An Empirical Study View project Cyber-Human Systems View project SAD: Social Anxiety and Depression Monitoring System for College Students.* Technical Report. https://www.researchgate.net/publication/295918103

[13] Laurens der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[14] Theodoros Evgeniou and Massimiliano Pontil. 2001. Support Vector Machines: Theory and Applications, Vol. 2049. 249–257. https://doi.org/10.1007/3-540-44673-7{_}12

[15] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2016. RMPE: Regional Multi-person Pose Estimation. (12 2016). https://doi.org/10.48550/arxiv.1612.00137

[16] Simone Felicioni and Mariella Dimiccoli. 2021. Interaction-GCN: A Graph Convolutional Network Based Framework for Social Interaction Recognition in Egocentric Videos. In *Proceedings - International Conference on Image Processing, ICIP*, Vol. 2021-Septe. 2348–2352. https://doi.org/10.1109/ICIP42928.2021.9506690

[17] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2021. Ego4D: Around the World in 3,000 Hours of Egocentric Video. (10 2021). https://doi.org/10.48550/arxiv.2110.07058

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. (12 2015). https://doi.org/10.48550/arxiv.1512.03385

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (11 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[20] Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* 18, 17 (2017), 1–5. http://jmlr.org/papers/v18/16-365

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft COCO: Common Objects in Context. (5 2014). https://doi.org/10.48550/arxiv.1405.0312

[22] Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3561–3569. https://doi.org/10.1109/CVPR.2019.00368

[23] Zihe Liu, Weiying Hou, Jiayi Zhang, Chenyu Cao, and Bin Wu. 2022. A Multimodal Approach for Multiple-Relation Extraction in Videos. *Multimedia Tools and Applications* 81, 4 (2022), 4909–4934. https://doi.org/10.1007/s11042-021-11466-y

[24] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. 2013. Understanding variable importances in Forests of randomized trees. In *Advances in Neural Information Processing Systems*.

[25] Peshawa Muhammad Ali and Rezhna Faraj. 2014. *Data Normalization and Standardization: A Technical Report.* https://doi.org/10.13140/RG.2.2.28948.04489

[26] Vladimir Nasteski. 2017. An overview of the supervised machine learning methods. *HORIZONS.B* 4 (12 2017), 51–62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, Vol. 32.

[28] 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. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 85 (2011), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

[29] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. (4 2018). https://doi.org/10.48550/arxiv.1804.02767

[30] Sebastian Ruder. 2017. An Overview of Multi-Task Learning in Deep Neural Networks. (6 2017). https://doi.org/10.48550/arxiv.1706.05098

[31] Matthias Schonlau and Rosie Yuyan Zou. 2020. The random forest algorithm for statistical learning. *Stata Journal* 20, 1 (2020). https://doi.org/10.1177/1536867X20909688

[32] Amanpreet Singh, Narina Thakur, and Aakanksha Sharma. 2016. A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. 1310–1315.

[33] Ratha Siv, Matei Mancas, Bernard Gosselin, Dona Valy, and Sokchenda Sreng. 2022. People Tracking and Re-Identifying in Distributed Contexts: Extension Study of PoseTReID. (5 2022). https://doi.org/10.48550/arxiv.2205.10086

[34] Qianru Sun, Bernt Schiele, and Mario Fritz. 2017. A Domain Based Approach to Social Relation Recognition. (4 2017). https://doi.org/10.48550/arxiv.1704.06456

[35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 07-12-June-2015. https://doi.org/10.1109/CVPR.2015.7298594

[36] J E T Akinsola, Akinsola Jet, and Hinmikaiye J O. 2017. Supervised Machine Learning Algorithms. *International Journal of Computer Trends and Technology* 48, 8 (2017).

[37] Abdel Aziz Taha and Allan Hanbury. 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging* 15, 1 (2015), 29. https://doi.org/10.1186/s12880-015-0068-x

[38] Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. 2019. Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making* 19, 1 (2019). https://doi.org/10.1186/s12911-019-1004-8

[39] Debra Umberson and Jennifer Karas Montez. 2010. Social Relationships and Health: A Flashpoint for Health Policy. *Journal of Health and Social Behavior* 51, 1_suppl (3 2010), S54–S66. https://doi.org/10.1177/0022146510383501

[40] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. 2018. Pose Flow: Efficient Online Pose Tracking. (2 2018). https://doi.org/10.48550/arxiv.1802.00977

[41] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. 2019. Learning Generalisable Omni-Scale Representations for Person Re-Identification. (10 2019). https://doi.org/10.48550/arxiv.1910.06827

[42] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. 2019. Omni-Scale Feature Learning for Person Re-Identification. (5 2019). https://doi.org/10.48550/arxiv.1905.00953

A CONFUSION MATRICES OF DIFFERENT CLASSIFIERS

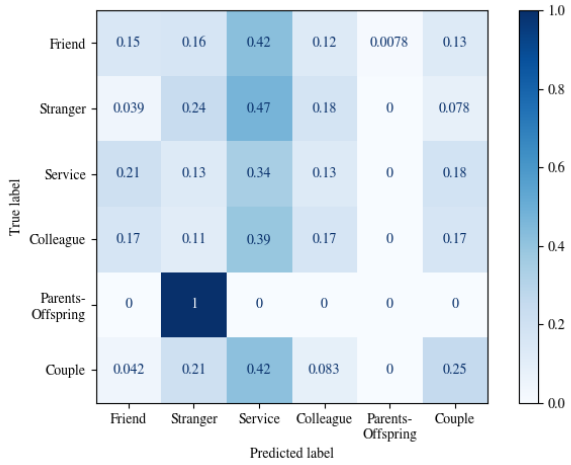

Fig. 5. Confusion matrix for AdaBoost classifier on a SMOTE over-scaled dataset

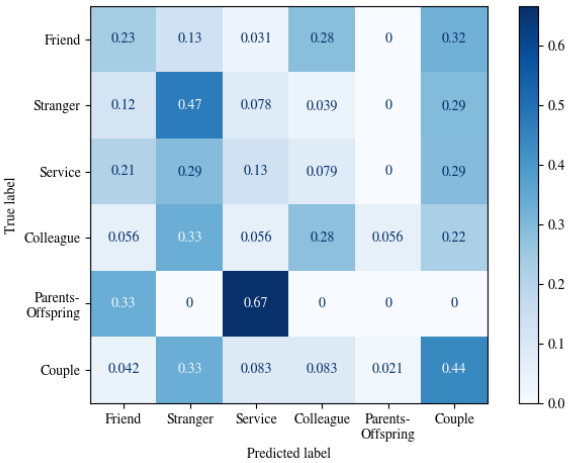

Fig. 6. Confusion matrix for Decision Tree classifier on a MinMax scaled and SMOTE over-scaled dataset

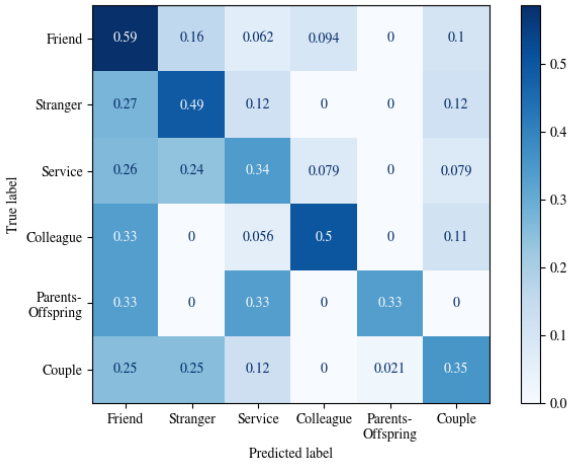

Fig. 7. Confusion matrix for Gaussian Process classifier on a MinMax scaled and SMOTE over-scaled dataset

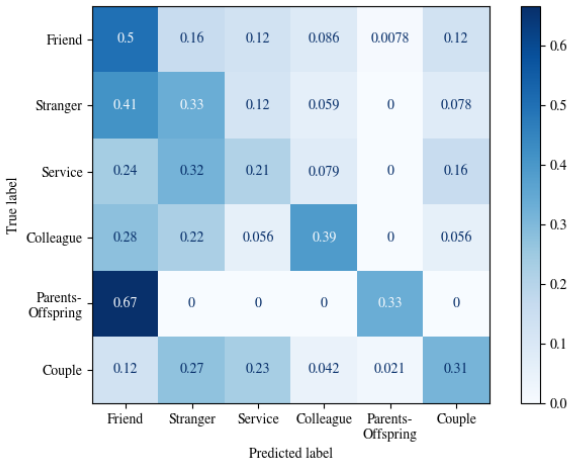

Fig. 8. Confusion matrix for Linear SVM classifier on a SMOTE over-scaled dataset

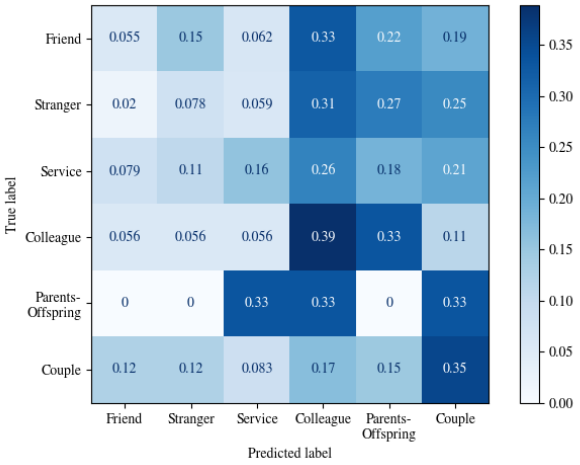

Fig. 9. Confusion matrix for Naive Bayes classifier on a SMOTE over-scaled dataset

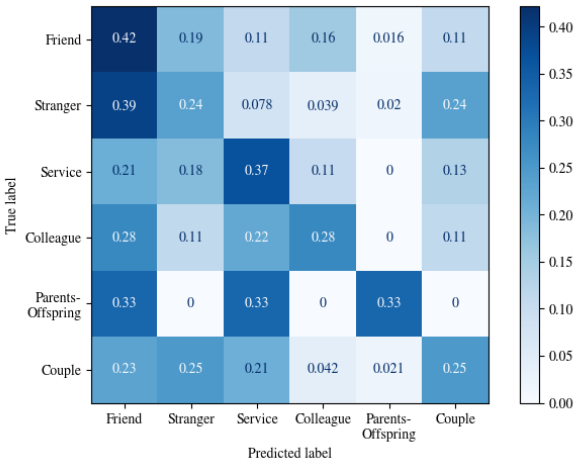

Fig. 10. Confusion matrix for Nearest Neighbors classifier on a Standard scaled and SMOTE over-scaled dataset

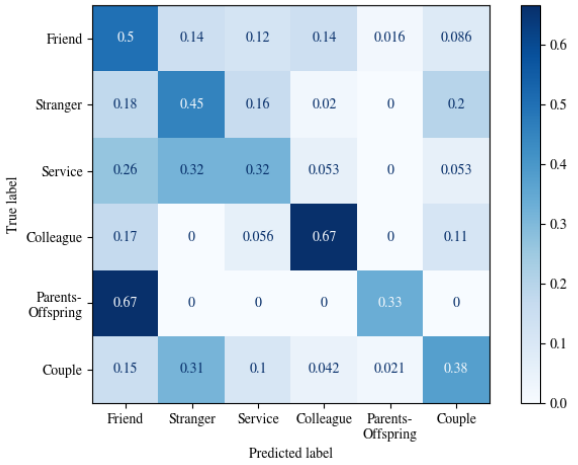

Fig. 11. Confusion matrix for Neural Net classifier on a Standard scaled and SMOTE over-scaled dataset

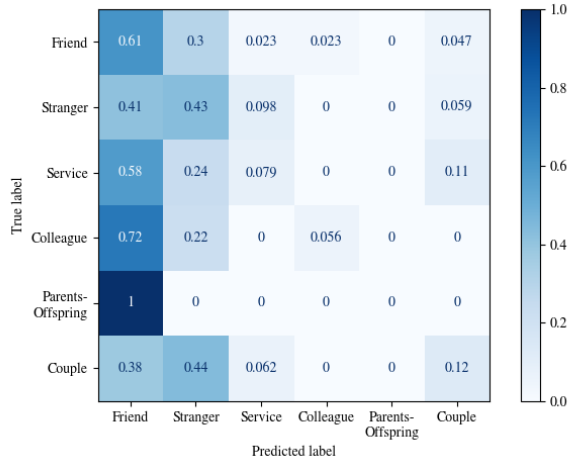

Fig. 12. Confusion matrix for QDA classifier on a SMOTE over-scaled dataset

B FEATURE IMPORTANCE

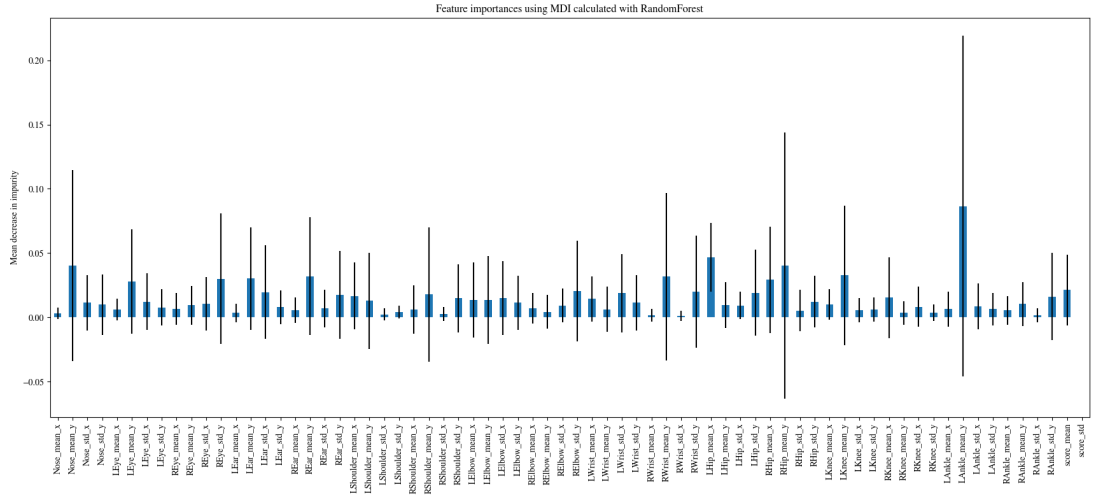

Fig. 13. Feature importance calculated with mean decrease in impurity on the *Random Forest* classifier

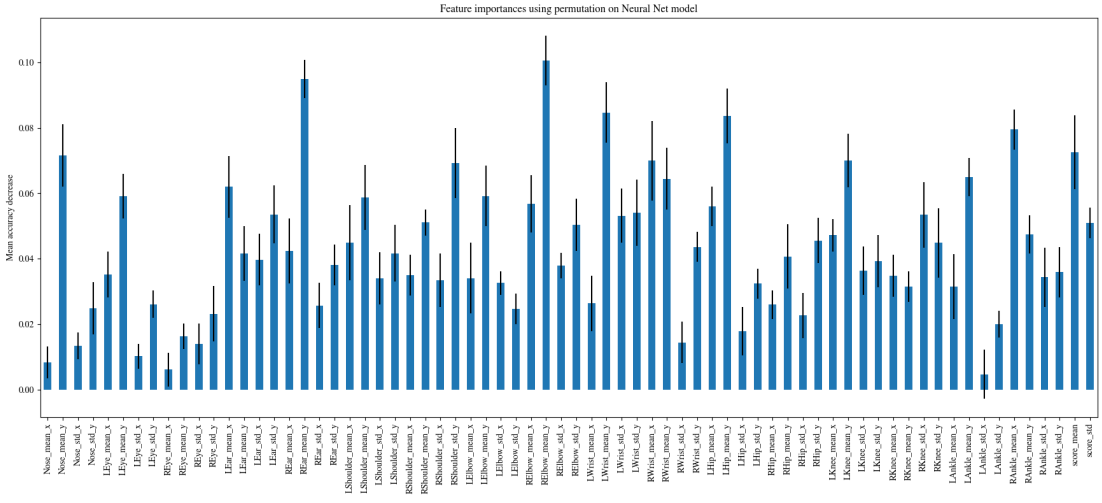

Fig. 14. Feature importance calculated with feature permutation on the *Neural Net* classifier