BCM19: A Multi-View Dataset for Gesture Recognition in Social Conflict Situations

Karam Tomotaki-Dawoud[‡1], Birgit Nierula[‡1], Farelle Toumaleu Siewe[1], Thomas Koch[1], Daniel Johannes Meyer[1], Andreas Bock[3], Marianne Heinze[3], Daniela Knuth[3], Denis Martin[3], Julia Schander[3], Anna Hilsmann[1], Peter Eisert[1,2] and Sebastian Bosse[1]

[1]Fraunhofer HHI, Berlin, Germany
[2]Humboldt University, Berlin, Germany
[3]Akkon Hochschule, Berlin, Germany

***Abstract*—In conflict situations, non-verbal communication plays a crucial role in escalation or de-escalation. We present steps towards an algorithm for classifying gesture-related communication markers and introduce BCM19, a novel multi-view RGB dataset containing 19 body gestures relevant to law enforcement interactions. To address the multi-view challenge, we propose embedding 2D- and 3D-skeleton estimations into pairwise distance feature vectors. Our comprehensive comparison of traditional and modern classifiers demonstrates high accuracy across various feature representations. This approach represents a significant step towards integrating social signal processing into our understanding of communication in conflict situations, with potential applications in law enforcement training and conflict de-escalation strategies.**

***Index Terms*—social signal processing, multi-view gesture recognition, body communication markers**

I. Introduction

In social situations, verbal and non-verbal cues can contribute to whether a situation will escalate or deescalate. Non-verbal cues may include facial expressions, gestures, tone of voice [1] as well as body movements and interpersonal distance [2]–[4]. They can convey meaning, emotions, and intentions in addition to verbal information [1], [5], [6].

While we might be more aware of our spoken word, non-verbal cues are often used unconsciously and their interpretation can vary significantly between and within individuals over time and circumstances. Especially communication markers like respect [5], [7]–[9], empathy [5], [10] or displaying dominance and authority [5], [11] and emotions like happiness or aggression can be open to interpretation and may vary in their non-verbal expressions with respect to background, experiences and other factors like mental disorders [12] or intoxication levels [5].

Police forces often face conflict situations, which can be associated with high levels of stress and can jeopardize the safety of those involved. As an authority enforcing norms and laws, they play a special role in solving conflicts, especially through their own strategies of de-escalating behavior. One way to enhance these de-escalation skills is to foster awareness of verbal and non-verbal cues in Extended Reality (XR) training. Such XR-trainings have been previously suggested for the training of law enforcement personnel due to their ability to induce similar stress levels as real live scenarios [13] but also due to their ability to address non-conscious and implicit beliefs [14].

In the present study, we focus specifically on body gestures as an aspect of non-verbal communication and aim at developing an automated detector for a subset of body gestures that were identified by experts from the social sciences and police as cues that may be involved in leading towards an escalation or de-escalation of a conflict situation.

Gesture recognition is an active research field [15], [16], [17], driven by advancements in deep learning models and the availability of large-scale datasets. Key limitations in existing datasets include a limited number of subjects, unspecific gesture categories, constrained frontal camera views, and limited environments [18]. Additionally, several methods have been developed that utilize a single view approach, employing either solely RGB image information, RGB image and depth information (RGB+D), 2D- or 3D-skeleton data, or a combination of those options.

However, for many real-world applications, using a single camera view may limit recognition in occlusion cases or when the user is not facing the camera.

In this paper, we address these issues by introducing the Body Communication Markers dataset (BCM19), which includes multiview RGB+Skeleton data of 19 gestures. To the best of our knowledge, there is no similar dataset containing the specific gestures we have chosen.

We propose a multi-view approach to skeleton-based gesture recognition using handcrafted features. In a multi-view gesture recognition experiment, we investigated how different skeleton-based feature representation sets influence the performance of both traditional and modern machine learning models. Our results demonstrate that the various models utilizing these features perform similarly to using raw skeleton data for recognition, while also providing better interpretability.

The paper is structured as follows: Sec. II gives an overview over prior work related to the gesture recognition pipeline. In Sec. III we introduce and describe the novel BCM19

‡ These authors contributed equally to this work.

This research was funded by the German Federal Ministry of Education and Research. Project name K3VR, funding number 13N16388.

dataset. The proposed gesture recognition pipeline, including the features extraction from the skeletons, is presented in Sec. IV. Our experimental results are presented in Sec. V. In Sec. VI we discuss the key takeaways, ethical aspects, limitations and give suggestions for future work.

II. Related work

In this section, we will briefly describe the most related work to the different modules of the pipeline - i.e. the pose estimation and the gesture recognition - described in Section IV and displayed in Fig.2. These methods encompass traditional and modern machine learning models which will be compared in our experiment.

Ever since the introduction of DeepPose in 2014 [19], human **pose estimation** has seen a surge in the development of neural network-based models [16]. These models are used for various purposes, including estimation pose for single-view images, time-sequenced images, multi-person pose estimation, and 3D pose estimation. Our choice, BlazePose GHUM [20], [21], [22] is a lightweight neural network for 2D and 3D human body landmarks and pose estimation.

With regards to the **gesture recognition** module, some common choices for pattern recognition include: i) support vector machines (SVM) [23] which are a supervised models that maximizes the margin to the decision boundary between different classes. Kernel SVM uses kernel functions to work in a high-dimensional feature space. It involves nonlinear optimization with a convex objective function, which simplifies solving the optimization problem [24].

ii) K-Nearest Neighbor (KNN) is a non-parametric supervised learning algorithm, that assigns the class or values of a novel point through a majority vote of its nearest K classified Neighbor [25] based on the assumption that similar points are approximately closer together, introduced by [26] in the 50s. It is one of the popular and simplest ML algorithms in use.

iii) Random Decision Forests (RF) [27], [28] is an ensemble learning-based algorithm, where a decision is made by averaging the output or majority vote of multiple decision trees. RF fixes the decision tree over-fitting problem by constructing a randomly chosen feature subspaces to ensure low correlation between decision trees.

iv) A multilayer perceptron (MLP) is a feedforward artificial neural network (NN) consisting of fully connected neurons with a nonlinear activation function. By utilizing nested architectures and deep learning methodologies [29], neural networks have achieved outstanding performance on various complex and challenging tasks [30].

III. Body Communication Markers 19

Nineteen relevant body communication markers were derived from literature research and interviews with experts involved in law enforcement trainings. This lead to a list of body gestures which are illustrated in Fig. 1 and described in the gesture dictionary in Table I. Note that poses 3, 4, and 5 are to some extend also neutral poses, as they represent the standard professional stance that police officers are trained to adopt during their duties because they balance approachability with readiness.

TABLE I
Gesture dictionary with descriptions of each body pose in the BCM19 dataset

1. Hands on hips: To have a frontal standing and put ones hands on the hips while facing another person.
2. Hands crossed: To have a frontal standing and cross ones arms over the chest while facing another person.
3. Hands tucked in collar: To have a frontal standing and stand facing one's counterpart while having the hands on the collar of the vest.*
4. Hands tucked in armholes of vest: To have a frontal standing faceing one's counterpart while having the hands in the armholes of the vest.*
5. Hands between nose and belly button: To have a frontal standing facing one's counterpart while gesturing with the hands between navel and nose.*
6. Face protection one hand: To shield one's face with a raised palm facing the other person.
7. Threaten use of firearm: To signal potential use of a firearm by placing one's right hand on the holstered weapon.
8. Protective movement with both arms: To shield one's head with raised arms against an object.
9. Head tilting: To convey empathy towards another by tilting one's head.
10. Empty hands: To demonstrate to one's counterpart to be unarmed by showing one's empty open palms.
11. Opening hands: To use open hand gestures during communication with one's counterpart.
12. Making oneself taller: To make oneself appear larger to demonstrate strength to one's counterpart.
13. Making oneself taller while moving towards someone: To make oneself appear larger and move towards the other person.
14. Hands on back: To keep one's hands behind one's back while communicating with another person.
15. Hands in pockets: To keep one's hands in one's pockets while communicating with another person.
16. Beckoning someone: To beckon or wave someone over.
17. Stop: To signal to another person to stop.
18. Pointing: To point at another person.
19. Threatening violence: To raise a fist and threaten someone with violence.
* This is a common posture for law enforcement personnel.

The dataset was collected from volunteers who enacted the gestures while being captured by two synchronized stereo cameras[1]. The cameras were placed 520 cm apart, and participants stood approximately 350 cm from each camera.

Twenty-six volunteers (13 females and 13 males, aged 23–42 years) participated in the data recording. They were informed about the study and the data handling procedures, and gave written consent before their participation.

Participants were asked to stand in front of the two cameras at a marked location. Before making a gesture, they viewed a video of that gesture performed by a police officer, along with a brief description of the gesture. Then the screen went blank, and they were asked to put themselves into the shoes of a police officer and perform the presented gesture five times with short breaks between repetitions.

This effort resulted in the Body Communication Markers 19 (BCM19), a skeleton (2D-3D) and image-based dataset[2].

[1]ZED 2i cameras (Stereolabs, San Francisco, USA)

[2]The dataset will be available on request for research purposes to researchers only [link to be added in the camera-ready version and on reviewer's request].

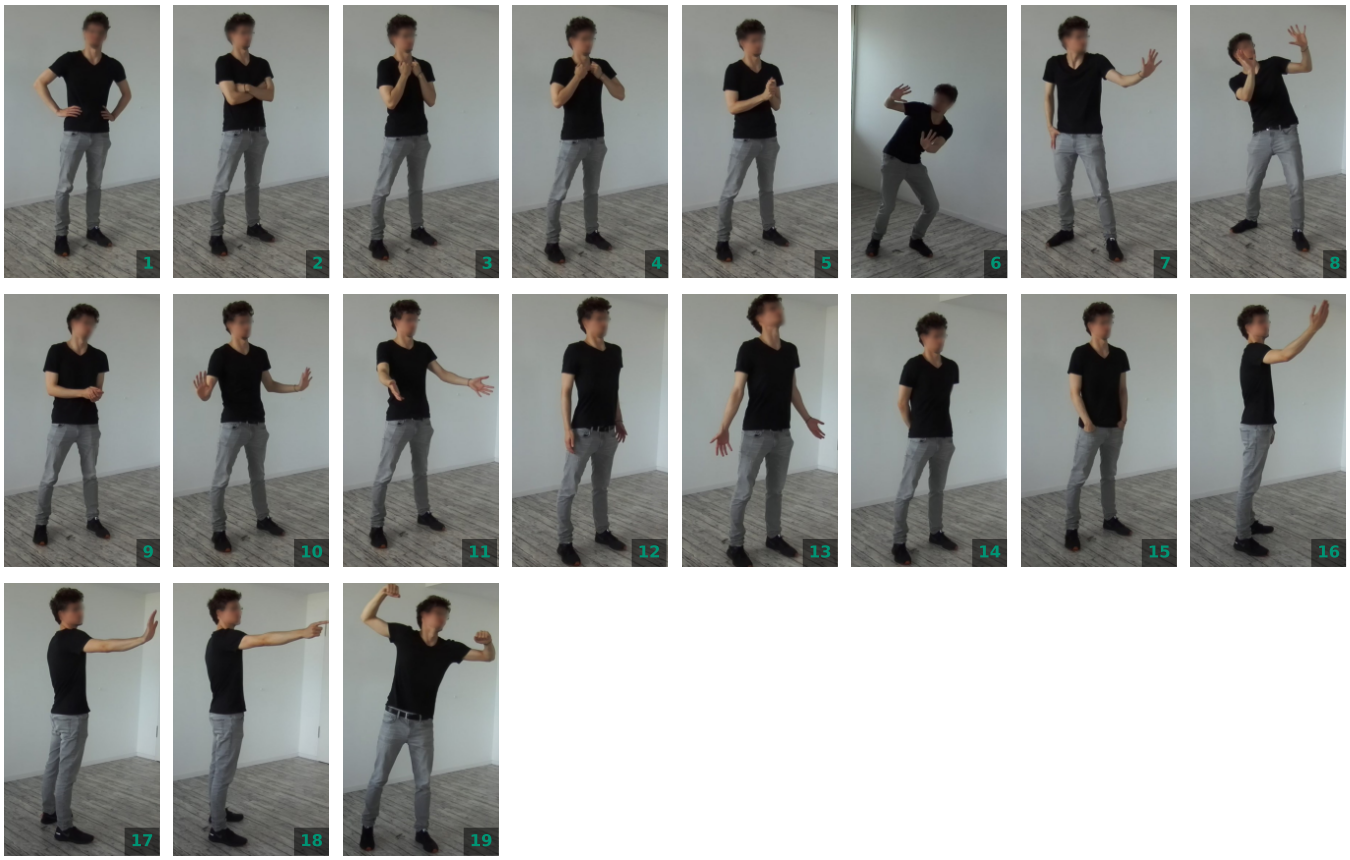

Fig. 1. BCM19 classes: 1) hands on hips, 2) hands crossed, 3) hands tucked in collar, 4) hands tucked in armholes of vest, 5) hands between nose and belly button, 6) face protection one hand, 7) threaten use of firearm, 8) protective movement with both arms, 9) head tilting, 10) empty hands, 11) opening hands, 12) making oneself taller, 13) making oneself taller while moving towards someone, 14) hands on back, 15) hands in pockets, 16) beckoning someone, 17) stop, 18) pointing, 19) threatening violence

IV. Gesture recognition pipeline

Our gesture recognition pipeline roughly follows the one outlined in [16], where two cameras capture the person from two angles, followed by a body pose 2D and 3D skeleton estimation. These skeletons are used to extract two types of suggested feature vectors, which are fed to pattern recognition ML classifiers, as shown in Figure 2.

One of the main steps, following the creation of the BCM19 dataset, is extracting meaningful feature representations for later processing with the recognition model. In traditional machine learning (ML) systems, experts used to create filters and functions based on their domain knowledge to pre-process input data. This approach aimed to make the data more understandable for those familiar with the field. By simplifying the raw data and focusing on the most relevant information, these features helped ensure that the model considered important aspects of the problem domain and avoided biases. Furthermore, involving domain expertise in feature design could enhance model performance, particularly when the data is limited or noisy.

Alternatively, an end-to-end approach to automatically learn features may enhance performance while sacrificing interpretability. Such a model might learn some undesired behaviors, known as the "Clever Hans" effect [31], and shortcuts caused by artifacts in the dataset or the pre-processing procedure [32].

Similar to the traditional Bag of Visual Words (BoVW) method, where a visual word dictionary is constructed using handcrafted filter representations, gestures can be defined through the relationships between body parts [16]. Using BlazePose, we extract both the 2D skeleton in the image pixel space and the 3D skeleton, shown in Fig. 3. We propose two sets of feature representations as alternatives to raw data:

1) The pairwise distances of the 3D skeleton joints set involve computing the Euclidean distances between all pairs of 3D joints, excluding self-comparisons. This method represents the relationships between joints in the 3D space of the individual user:

$$Dataset_{\text{world}} = \left\{ \left\| \mathbf{u}_w^i - \mathbf{v}_w^j \right\| \mid i, j \in S, i \neq j \right\} \quad (1)$$

Where $S = \{1, 2, \ldots, n\}$ is the set of all chosen joints. $\mathbf{u}_w^i = (u_{x_w}^i, u_{y_w}^i, u_{z_w}^i)$ and $\mathbf{v}_w^j = (v_{x_w}^j, v_{y_w}^j, v_{z_w}^j)$ are the world coordinates of joints i and j in 3D space.

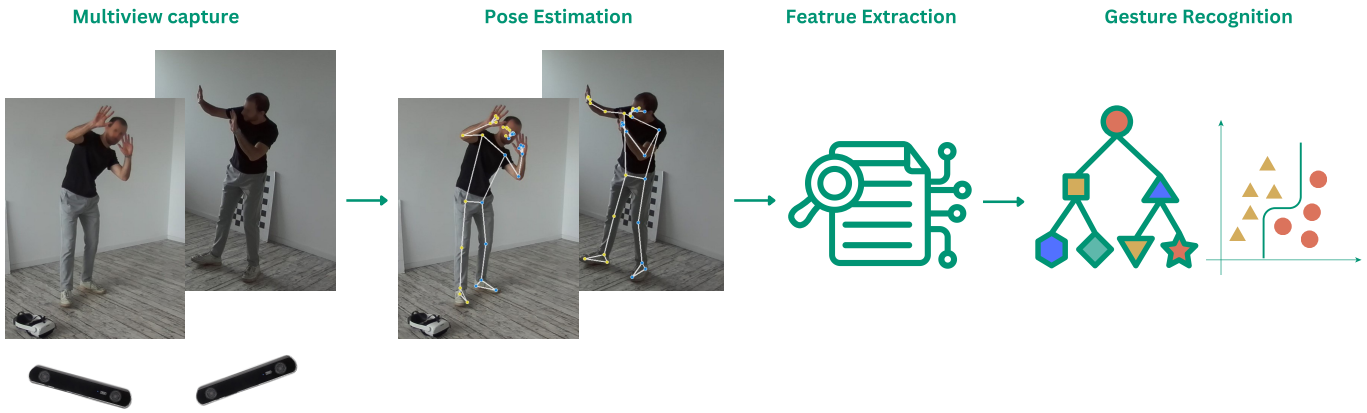

Fig. 2. Body Gesture Recognition Pipeline: 1. Multiview capture: Two cameras capture the person from 2 angles (for details see the Appendix **??**). 2. Pose estimation: Next, the 2D and 3D kinematic body pose was estimated. 3. Feature extraction: Based on the estimated 2D and 3D keypoints, a representation was extracted based on the pairwise distances between those points. 4. Gesture recognition: Finally traditional or modern ML models were used to categorize the body gestures.

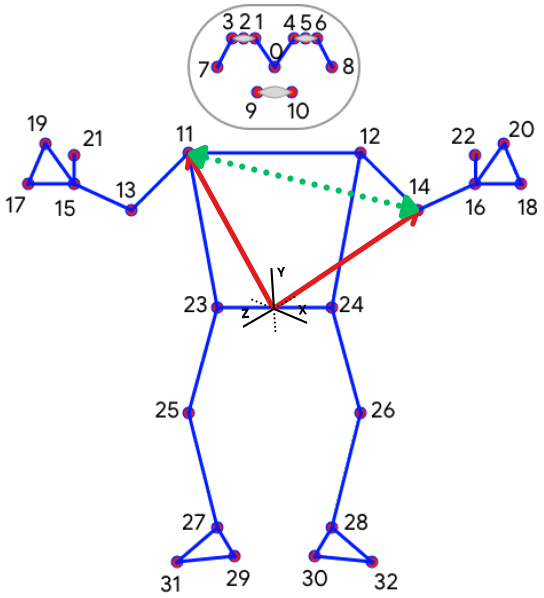

Fig. 3. The original 33 keypoints [20]. The dashed green line indicates the distance between two joints, calculated directly on the image plane. While the two red 3D vectors represent the estimated 3D coordinates in meters, with the origin at the center of the hips.

2) A set that combines both 2D and 3D information. In this set, we compute the distances between all 2D joints in the image pairs and normalize them by the respective 3D joint distances. This method captures the image plane relationships and accounts for the different body sizes of the participants, however, this might also incorporate additional view dependencies.

$$Dataset_{\text{mix}} = \left\{ \frac{\left\|\mathbf{u}_p^i - \mathbf{v}_p^j\right\|}{\left\|\mathbf{u}_w^i - \mathbf{v}_w^j\right\|} \middle| \, i, j \in S, i \neq j \right\} \quad (2)$$

in addition to the terms from (1), the terms $\mathbf{u}_p^i = (u_{x_p}^i, u_{y_p}^i)$ and $\mathbf{v}_p^j = (v_{x_p}^j, v_{y_p}^j)$: Pixel coordinates of joints i and j in the 2D image.

3) Finally, a less interpretable baseline that directly uses the raw 3D skeleton coordinates (x,y,z). This representation does not involve pairwise distances but directly utilizes the spatial coordinates of all joints.

$$Dataset_{\text{raw coordinates}} = \left\{ \left\|(x_w^i, y_w^i, z_w^i)\right\| \mid i \in S \right\} \quad (3)$$

To minimize redundancy, we have excluded the outer and inner eye keypoints and the mouth from our selection of joints. Estimating the center of the eyes, the nose, and the ears is sufficient to account for the head's orientation. Hence, the first two feature sets consist of 812 features, representing the unique combinations of 27 joint pairs (406) from both camera views (406 × 2). The raw coordinates set includes 162 features (3 coordinates × 27 joints × 2 cameras).

The final crucial decision for the gesture recognition pipeline is choosing the pattern recognition model. In addition to the traditional machine learning models mentioned in Sec. II, we also consider using an MLP as a modern alternative. Although there are various potential applications, we will concentrate on three basic NN model designs

1) A basic NN with 2 layers, where the first layer has half the number of neurons as our feature vector. The smaller network size aims to reduce overfitting on our dataset.
2) A deeper NN with 8 layers that includes a bottleneck in the middle, forcing the network to learn a compressed representation in the latent space [33].
3) Lastly, a VGG classifier head [34], which consists of 3 fully connected layers, each with 4096 neurons, followed by a softmax output.

V. Results

In our experiment on multi-view gesture recognition, we examined how different sets of skeleton-based feature rep-

resentations impact the performance of both traditional and modern machine learning models. We evaluated three types of feature sets: 3D relational geometric representations, a combination of 2D and 3D relational data, and raw 3D coordinates.

We tested the performance of seven machine learning models: K-nearest neighbors (KNN) with 5 neighbors ($k = 5$), linear and kernel support vector machines (SVMs) using a Radial Basis Function (RBF) kernel for the kernel SVM, random forest (RF) with 100 estimators, and three custom-designed neural networks—a simple 2-layer neural network (NN), an 8-layer bottlenecked NN, and a VGG-head based model. Each model was trained and evaluated using 5-fold cross-validation across all feature sets to measure how the choice of features impacted recognition accuracy, precision, recall, and F1-score. To mitigate overfitting, training of all neural networks included an early stopping criterion.

Our results, summarized in terms of accuracy and F1-score in Figures 4, 5, and **??**, show that the random forest model consistently matched or outperformed the neural networks, particularly when using the combined 2D and 3D feature set. Statistical analysis using adjusted (Bonferroni-corrected) paired t-tests revealed that, for the mixed 2D and 3D relational set, the random forest model significantly outperformed the neural networks, with adjusted p-values for accuracy ranging from 0.000019 to 0.028288. In contrast, for the 3D relational and raw 3D coordinate sets, the differences in performance between the random forest and the neural networks were less pronounced, with no statistically significant differences observed. Specifically, there were no significant differences in accuracy between the random forest and the simple NN (p = 1.0) or the bottlenecked NN (p = 0.077) for the 3D relational representation set. Similarly, differences in F1-score were not statistically significant after adjusting for multiple comparisons. On the raw 3D set, no significant differences were observed between the random forest and any of the neural networks.

By examining the mean performances of all models across all the feature representation sets, shown in Fig. **??**, some patterns are noticeable:

All models improved when trained on the raw coordinates compared to the other two relational sets, except for the linear SVM. The linear SVM show good accuracy (above 80 %) on the 3D relational representation set, saw a small degradation on the 2D-3D combined set, and then experienced a performance drop with the raw set.

Surprisingly, the kernel SVM performed poorly with the handcrafted features but showed notable improvement when trained on the raw data.

K-nearest neighbors delivered consistent accuracy around 80% across the different feature sets, with a small decrease for the 2D-3D set and a slight increase with the raw set.

Among the neural networks, the VGG-based model performed the worst overall (except for precision). Interestingly, while the VGG-head struggled with the handcrafted 812-feature set, it outperformed most models (except RF) when

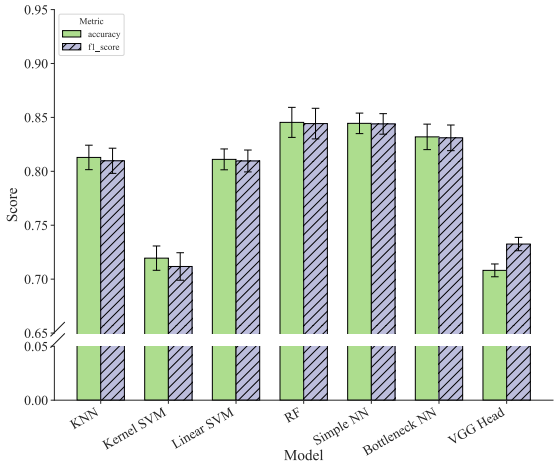

Fig. 4. Overall model comparison in terms of accuracy and F1-score on the 3D relational geometric representation dataset. The Random Forest (RF) model performs the best, slightly better than the simple and bottlenecked neural networks.

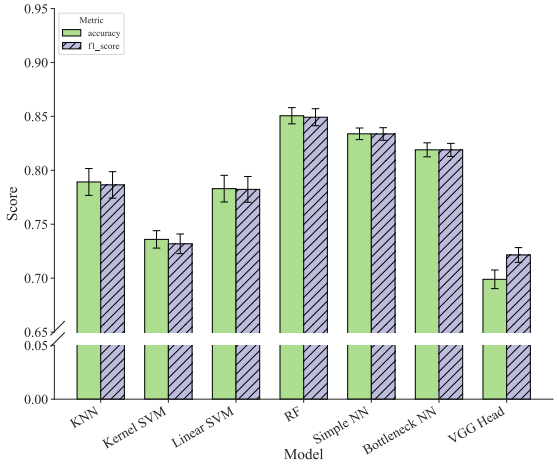

Fig. 5. Overall model comparison in terms of accuracy and F1-score on the combined 2D and 3D relational geometric representation dataset. The Random Forest (RF) model performs the best across both metrics, while the VGG-head model performs the worst.

using the 174 raw 3D coordinates. This suggests that the wide layers of the VGG model were prone to overfitting on the lower-dimensional feature space.

The simpler 2-layer and bottlenecked networks were able to learn meaningful representations on all the sets, achieving performance comparable to the random forest, with a small decrease on the 2D-3D set.

Lastly, the random forest was the best model over all the metrics and on all the feature sets. RF was the only model besides the kernel SVM to show improvement on the 2D-3D set compared to the 3D relational set.

VI. Discussion

In this paper, we addressed the limitations of existing gesture datasets, particularly their lack of gesture categories and their restriction to frontal camera views [18]. We introduced

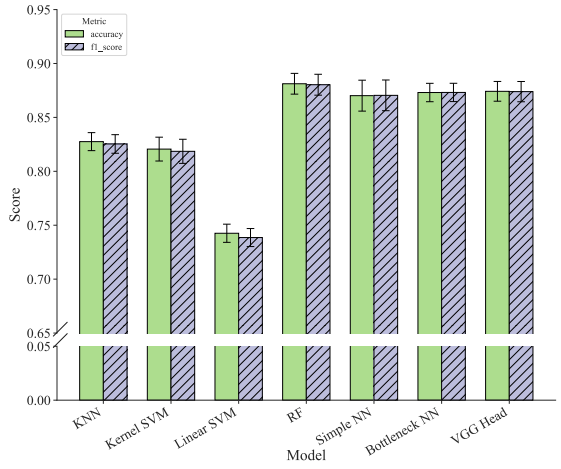

Fig. 6. Overall model comparison in terms of accuracy and F1-score on the raw 3D coordinates dataset. The Random Forest (RF) model performs slightly better than all of the neural network models. The kernel SVM manages to surpass the linear model only on this dataset.

the BCM19 dataset and proposed a multi-view approach to skeleton-based gesture recognition. BCM19 includes multi-view RGB images from two cameras, along with 2D and 3D skeleton information. Our results demonstrate good accuracy across various feature representations. Notably, using handcrafted features with Random Forest and simple and bottlenecked neural networks achieved performance close to that of models trained on raw data. Moreover, these approaches offer the advantage of better model interpretability and less overfitting.

The strength of the BCM19 dataset lies in its comprehensive approach to gesture selection. These 19 body gestures were identified through a social science literature review and further refined by experts in law enforcement training. The selected gestures serve as potential indicators—communication markers—of conflict escalation or de-escalation, making them particularly valuable for awareness training in de-escalation techniques. While these gestures have not yet undergone scientific validation, the BCM19 dataset represents an important first step toward developing a comprehensive library of communication markers in law enforcement interactions.

While some of the models presented in this paper achieved accuracies between 80% and 90%, it was noticeable that all models achieved better results on lower-dimensional input data. In the present work, we focused on extracting relevant features from the skeleton data by using the pairwise distances between key body parts, which resulted in an increase in input dimensionality. Reducing the dimensionality of the feature space has been shown to be crucial for improving both model efficiency and interpretability [35], and can thereby simplify the model without losing predictive power. Future work could use methods from the field of explainable AI (XAI), such as Layer-wise Relevance Propagation (LRP) [36] or Spectral Relevance Analysis (SpRAy) [32], to identify the most relevant features and further reduce dimensionality. Additionally, fusing multiple neural network models with Random Forest could be a useful approach to increase overall model accuracy, as has been shown for action recognition by [37].

While our research demonstrates the potential for automated gesture recognition in law enforcement contexts, it is crucial to consider the ethical implications of deploying such technology. There are risks of bias in the training data, potential for misuse or over-reliance on automated systems, and privacy concerns related to surveillance. Future work should involve close collaboration with ethicists, legal experts, and community stakeholders to ensure that any implementation of this technology is responsible, with appropriate safeguards and oversight. Additionally, it is important to view this technology as a tool for training and reflection, rather than as a replacement for human judgment in high-stakes situations.

Although the study presents a promising approach, it is important to acknowledge potential limitations. The use of re-enacted gestures in the dataset creation process, while necessary for ethical and practical reasons, may not fully capture the nuances of real-world conflict situations. Future research should explore ways to validate and refine the model using authentic interaction data while carefully navigating privacy and ethical considerations. Moreover, the cultural specificity of gestures should be considered. The 19 identified gestures may vary in meaning or prevalence across different cultural contexts. Further studies could explore cross-cultural variations in police gestures and their interpretations.

ACKNOWLEDGMENT

We would like to express our deepest gratitude to our associated project partners Bayrische Polizei and Polizei Berlin for their valuable insight and guidance on gestures. Additionally, we are thankful for the participation of all dataset-recording contributors, as their engagement was instrumental in enabling us to successfully conduct this research.

REFERENCES

[1] D. Griffith, "De-escalation training: Learning to back off."
[2] G. Sytschjow, *Was der Körper sagt": nonverbale Kommunikation von Schutzpolizistinnen und Schutzpolizisten im Einsatz*, ser. Schriftenreihe Polizei & Wissenschaft. Verlag für Polizeiwissenschaft, Prof. Dr. Clemens Lorei.
[3] C. Howe, C. Decker, L. Knobloch, H. Can, and A. Bosch, "Bericht zur berliner polizeistudie. eine diskriminierungskritische, qualitative untersuchung ausgewählter dienstbereiche der polizei berlin." p. 141.
[4] U. Füllgrabe, *Psychologie der Eigensicherung: Überleben ist kein Zufall*, 7th ed. Boorberg.
[5] N. Todak and L. James, "A systematic social observation study of police de-escalation tactics." vol. 21, no. 4, pp. 509–543.
[6] D. Yurchenko and U. Pache, "Körpersprache - was ist das überhaupt?"
[7] M. Hermanutz, *Polizeiliches Auftreten - Respekt und Gewalt: eine empirische Untersuchung zum Einfluss verbaler Kommunikation und äußerem Erscheinungsbild von Polizeibeamten auf die Gewaltbereitschaft von Jugendlichen und jungen Erwachsenen*, ser. Polizei & Wissenschaft. Verl. f. Polizeiwissenschaft.
[8] C. Lorei, *Kommunikation statt Gewalt*, 2nd ed., ser. Polizeiwissenschaftliche Analysen. Verlag für Polizeiwissenschaft, no. 32.
[9] R. Voigt, N. P. Camp, V. Prabhakaran, W. L. Hamilton, R. C. Hetey, C. M. Griffiths, D. Jurgens, D. Jurafsky, and J. L. Eberhardt, "Language from police body camera footage shows racial disparities in officer respect," vol. 114, no. 25, pp. 6521–6526.

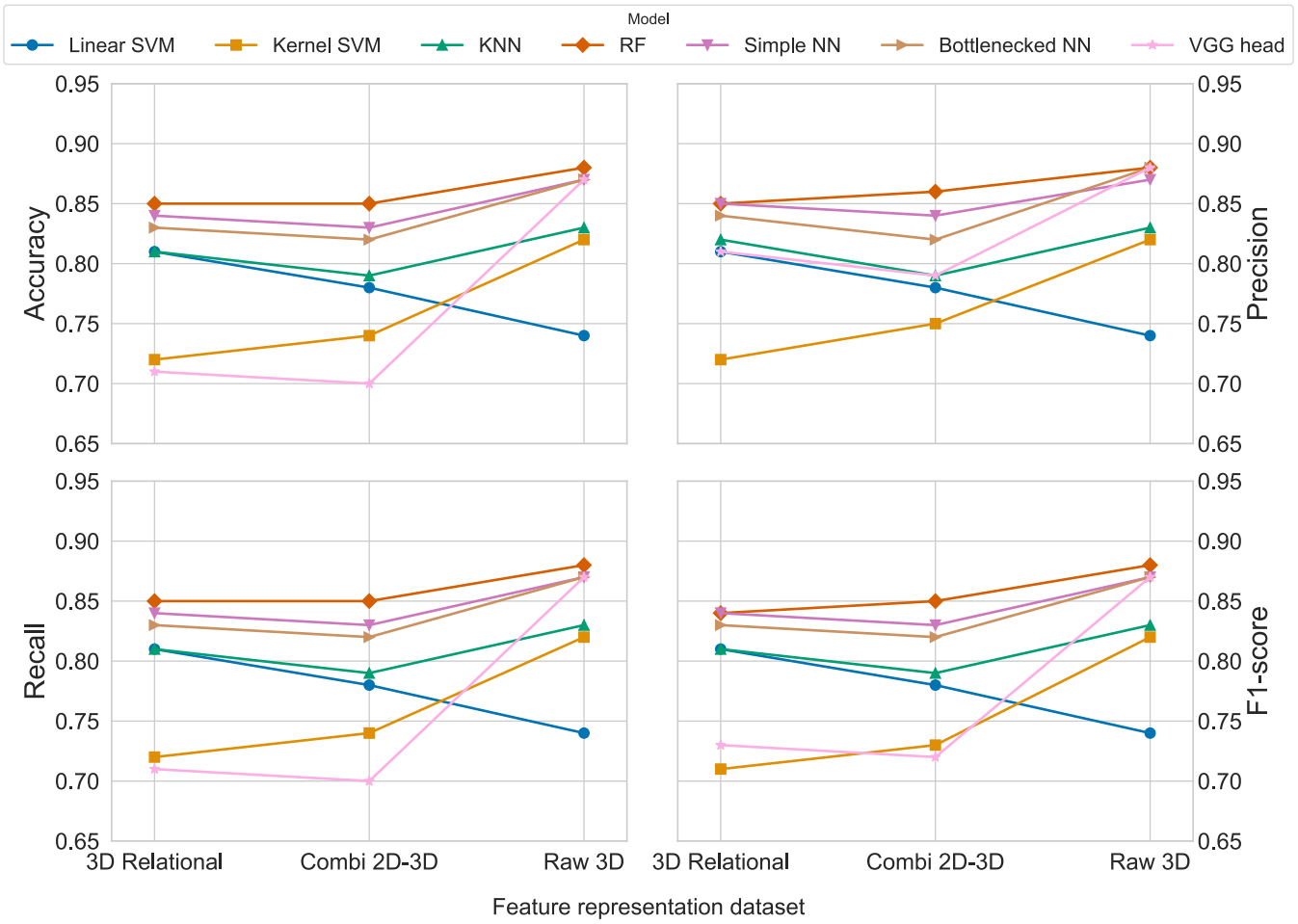

Fig. 7. Overall models' average performance of 5-fold cross-validation in terms of accuracy, precision, recall, and F1-score. The evaluation was conducted on three different datasets: 3D relational geometric representations, a combination of 2D and 3D relational data, and raw 3D coordinates. All four metrics indicate that Random Forest (RF) performed the best, followed by the simple neural network and the bottlenecked NN.

[10] M. B. Rosenberg, *Gewaltfreie Kommunikation: eine Sprache des Lebens*, 12th ed., ser. Reihe Kommunikation. Junfermann Verlag.

[11] R. Behr, "polizeigewalt hat es nicht gegeben" – cop culture als disposition für dominanz, Überlegenheit und grenzüberschreitung im polizeilichen alltagshandeln," pp. 217–238.

[12] T. Feltes and M. Alex, "Polizeilicher umgang mit psychisch gestörten personen," in *Polizeiarbeit zwischen Praxishandeln und Rechtsordnung*, D. Hunold and A. Ruch, Eds. Springer Fachmedien Wiesbaden, pp. 279–299, series Title: Edition Forschung und Entwicklung in der Strafrechtspflege.

[13] L. Kleygrewe, R. I. V. Hutter, M. Koedijk, and R. R. D. Oudejans, "Virtual reality training for police officers: a comparison of training responses in VR and real-life training," vol. 25, no. 1, pp. 18–37. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/15614263.2023.2176307

[14] S. Kishore, B. Spanlang, G. Iruretagoyena, S. Halan, D. Szostak, and M. Slater, "A virtual reality embodiment technique to enhance helping behavior of police toward a victim of police racial aggression," vol. 28, pp. 5–27.

[15] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, May 2007.

[16] F. Noroozi, C. A. Corneanu, D. Kaminska, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on Emotional Body Gesture Recognition," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 505–523, Apr. 2021.

[17] V. Sharma, H. Kolivand, S. Asadianfam, D. Al-Jumeily, and M. Jayabalan, "Gesture Recognition Techniques," in *2023 15th International Conference on Developments in eSystems Engineering (DeSE)*. Baghdad & Anbar, Iraq: IEEE, Jan. 2023, pp. 244–249.

[18] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2684–2701, 2020.

[19] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1653–1660.

[20] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. L. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," *ArXiv*, vol. abs/2006.10204, 2020.

[21] I. Grishchenko, V. Bazarevsky, A. Zanfir, E. G. Bazavan, M. Zanfir, R. Yee, K. Raveendran, M. Zhdanovich, M. Grundmann, and C. Sminchisescu, "Blazepose ghum holistic: Real-time 3d human landmarks and pose estimation," 2022.

[22] H. Xu, E. G. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, "Ghum ghuml: Generative 3d human shape and articulated pose models," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6183–6192.

[23] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, p. 273–297, 1995.

[24] C. M. Bishop, *Pattern recognition and machine learning*, ser. Information science and statistics. New York: Springer, 2006.

[25] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, p. 21–27, 1967.

[26] E. Fix and J. Hodges, *Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties*. USAF School of Aviation Medicine, 1951.
[27] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1. Montreal, Que., Canada: IEEE Comput. Soc. Press, 1995, p. 278–282.
[28] ——, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, p. 832–844, Aug. 1998.
[29] *Neural Networks: Tricks of the Trade: Second Edition*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, vol. 7700.
[30] *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, ser. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2022, vol. 13200.
[31] O. Pfungst and C. L. Rahn, *Clever Hans (the horse of Mr. Von Osten) a contribution to experimental animal and human psychology,*. New York,: H. Holt and company, 1911.
[32] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, Mar. 2019.
[33] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, ser. Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press, 2016.
[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.
[35] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, no. null, p. 1157–1182, mar 2003.
[36] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, pp. 1–46, 07 2015.
[37] Y. Lavinia, H. H. Vo, and A. Verma, "Fusion based deep cnn for improved large-scale image action recognition," in *2016 IEEE International Symposium on Multimedia (ISM)*, 2016, pp. 609–614.