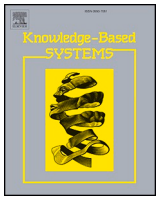
Knowledge-Based Systems 295 (2024) 111856

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Recognizing affective states from the expressive behavior of tennis players using convolutional neural networks

Darko Jekauc [a,*], Diana Burkart [a], Julian Fritsch [a], Marc Hesenius [b], Ole Meyer [b], Saquib Sarfraz [a], Rainer Stiefelhagen [a]

[a] Karlsruhe Institute of Technology: Kaiserstrasse 12 76131 Karlsruhe, Germany
[b] University of Duisburg-Essen: Forsthausweg 2 47057 Duisburg, Germany

ARTICLE INFO

Keywords:
Emotion
Affect recognition
CNN
Tennis
Body language
Automated affect recognition

ABSTRACT

This study describes an AI model by leveraging advanced Convolutional Neural Networks (CNNs) to recognize affective states in real-world sports settings, particularly tennis matches. In contrast to prior studies that primarily utilized data acquired from actors and rudimentary statistical methods, the present research emphasizes the analysis of bodily expressions in real-life contexts, aiming for a more naturalistic representation of human emotions. Our CNN-based models demonstrate an accuracy rate of up to 68.9 %, outperforming or matching human observers in many instances. Intriguingly, both the machine learning models and human observers exhibited a shared propensity to more effectively identify negative affective states, which may be attributed to the more intense and straightforward expression of these states. These results not only advance the state of the art in affective state recognition but also pave the way for broader applications, including in healthcare and automotive safety sectors, thereby constituting a significant advancement in the development of sophisticated and universally applicable emotional recognition systems.

1. Introduction

Affective processes, encompassing emotions and moods, play a pivotal role in various human behaviors and cognitions [1]. These processes pervade numerous aspects of life, influencing decisions in contexts such as consumer behavior, health care, and occupational activities [2]. Recognizing affective states in others is deemed a crucial ability, especially in professions demanding interpersonal interactions [3]. The current research explores the potential for artificial intelligence (AI) to recognize affective states in real-life situations, an ability traditionally considered human.

Affect and emotions are foundational concepts in psychological research, often delineated by their complexity and the processes through which they are experienced. Affect represents a broader spectrum of feelings and states, typically characterized by dimensions such as valence (positive or negative) and arousal (high or low), as outlined in the Circumplex Model of Affect [4]. Emotions, on the other hand, are considered more complex and specific, characterized within categorical models that identify discrete emotional states, such as anger, anxiety, or happiness [5–7]. Recognizing these distinctions, our study draws upon the framework proposed by Baumeister, Vohs, DeWall, and Zhang [8], differentiating between automatic affect—rapid, simple affective reactions captured by dimensional models—and full-fledged emotions, which are more nuanced and align with categorical interpretations. Importantly, the contextual factors surrounding these affective states play a crucial role in their expression and perception, a perspective emphasized by Feldman-Barrett's theory of constructed emotions, which posits that emotions are constructed in the moment, through a complex interplay of the brain, body, and culture [9]. This understanding of affect and emotions as contextually driven processes is particularly relevant to our analysis of tennis players' affective states, grounding our research in a comprehensive understanding of affective processes as they unfold in sports contexts.

Scherer [10] delineates five key components for assessing affective processes: cognitive, neurophysiological, motivational, motor expression, and subjective feeling. While most contemporary methods prioritize the measurement of subjective feelings, significant advancements have occurred in other domains, including cognitive appraisal [11], brain mechanisms [12], physiological response [13,14], and expressive behavior [15]. Such advancements indicate the absence of a 'gold

* Corresponding author. Institute for Sport and Sport Science, Kaiserstrasse 12 76131 Karlsruhe, Germany.
E-mail address: darko.jekauc@kit.edu (D. Jekauc).

https://doi.org/10.1016/j.knosys.2024.111856
Received 20 December 2023; Received in revised form 7 April 2024; Accepted 23 April 2024
Available online 24 April 2024

standard' in affective measurement, highlighting that these metrics capture dimensions of affect rather than discrete emotional states [16]. Traditional approaches also face the limitation of failing to capture the dynamic nature of affective processes in real-world settings [10].

AI offers a promising alternative to traditional measures of affective processes, promising non-invasive and continuous assessment of affective states through motor expression. However, the task of recognizing affective states from expressive behaviors presents inherent complexities and achieves only partial accuracy, even among human evaluators [17,18]. One of the challenges lies in the unpredictable and unsystematic occurrence of emotions and affective states in real-world scenarios. Nonetheless, sports settings offer a unique environment to tackle these challenges, given the rich and varied expressions of affective states frequently observed in competitive situations.

In the pursuit of these research objectives, we have developed and trained an AI model that demonstrates significant proficiency in classifying affective states as either positive or negative in real-world sports scenarios. Utilizing video footage from tennis matches as a testbed for affective expressions in a real-life context, our model achieved accuracies that not only are comparable with human evaluators but also in certain instances surpass them. Notably, our findings reveal a commonality between human and machine assessments: both tend to recognize negative affective states with greater accuracy, which might be attributed to their heightened intensity and more explicit expression. The results of this study offer a promising avenue for further research and potential applications in various domains beyond sports psychology, including healthcare and automotive safety.

In addressing the challenges of affective state recognition in sports, particularly in tennis, our research introduces a novel methodological approach that significantly advances the field. Unlike existing studies that predominantly rely on facial expressions or controlled experimental settings [19], our work harnesses the dynamic and complex nature of real-world tennis matches, utilizing both posture (joint positions) and video frames to analyze athletes' expressive behaviors. This dual-modality approach, combined with advanced CNN architectures, is specifically optimized for the task of affect recognition in athletic performance. By meticulously developing a dataset that captures the authentic expressions of tennis players in competitive scenarios, we bridge the gap between theoretical affective computing models and their practical application in sports. Furthermore, our methodological innovations extend beyond the mere application of AI techniques to include a comprehensive preprocessing pipeline and a novel dataset, setting a new standard for research in the intersection of affective computing, AI, and sports psychology

1.1. Models for affect recognition in real sport scenarios

In our study of affective states in competitive sports, we identify such environments as optimal natural laboratories for observing emotional expressions in naturalistic settings [20]. The inherent structure of sports competitions, with their well-defined rules and goals, creates scenarios that are deeply relevant to an athlete's immediate goals and elicit distinct and pronounced emotional responses. In net and wall games such as tennis, badminton, or volleyball, these emotionally charged scenarios occur with regularity, driven by the dynamics of winning and losing points [21]. This constant ebb and flow in a competition fosters a continuous evolution of the player's emotional state, reflecting the real-time successes and setbacks experienced during the game [22]. We use the term "affective states in real-life" to refer to these emotional responses that occur in direct response to competitive events. The emotional salience of these events, whether they are consistent or inconsistent with the athlete's goal (e.g., winning the game), directly influences the valence of these affective states, resulting in positive or negative emotions based on the outcome's congruence with the athlete's goals [23].

Prior studies, grounded in these theoretical considerations, have endeavored to discern whether external human observers can differentiate between players having won or lost a point in sports like tennis [17, 21] and volleyball [24] by analyzing video sequences of the players' expressive behavior. These investigations imply that situational factors, such as winning or losing, are closely connected with players' affective states, offering robust indicators of athletes' emotional conditions. A significant obstacle within these studies is that recognition by human observers is influenced by multifarious elements, including a) *the player's appraisal of the situation*, b) *the transformation of a player's affective state into discernible expressive behavior*, and c) *the observer's interpretation of the behavior* (see Fig. 1).

Specifically, the *athlete's appraisal of the situation* is not solely influenced by the immediate outcome of a point, but also incorporates other elements, such as the strategic significance of the point within the competition [23]. Such considerations have been found to augment the intensity of the emotional response to winning or losing a point, particularly during critical phases of a match (e.g., the concluding stages of a closely contested game). For instance, Fritsch et al. [25] empirically demonstrated that an athlete's affective state manifested more expressively in crucial match situations as compared to less significant contexts. Consequently, the perceived importance of a situation has a tangible impact on the intensity of the affective reaction, underscoring the nuanced interplay between context and emotional response in sport.

In the realm of sports, the *translation of affective states into observable expressive behavior* extends beyond mere emotional reactions and also encompasses aspects of social nonverbal communication [26]. In other words, the visible behavior is not solely a manifestation of the individual's affective state but also serves as a communicative signal directed at other participants, such as teammates or opponents [27]. Expressive behavior, in this context, can be both a genuine reflection of the individual's emotional state and a strategic tool, deliberately suppressed or exaggerated to influence others [28]. This dual function of expressive behavior has been empirically substantiated in team sports. For instance, Fritsch et al. [24] revealed that in volleyball, players showed more intense positive affective reactions following a win, whereas negative affective reactions were less conspicuously displayed after losing points. This underscores the complex interplay between expressive behavior, affective states, and social communication in sports, where expression becomes both a reflection of emotion and a calculated response tailored to specific competitive scenarios.

The process of recognizing affective states fundamentally relies on the observer's interpretation of expressive behavior, guided by the principles outlined in Brunswik's lens model, wherein the observer perceives a configuration of various observable cues [29]. Discerning the relevance and meaning of these cues necessitates repeated experience, drawing upon the generalized principles of learning that are applicable across animals [e.g. classical and operant conditioning; 30], humans [31] and AI [e.g. machine learning; 32]. Such experiential learning enables observers to develop an adeptness at interpreting the constellation of cues linked to expressive behavior. Supporting this

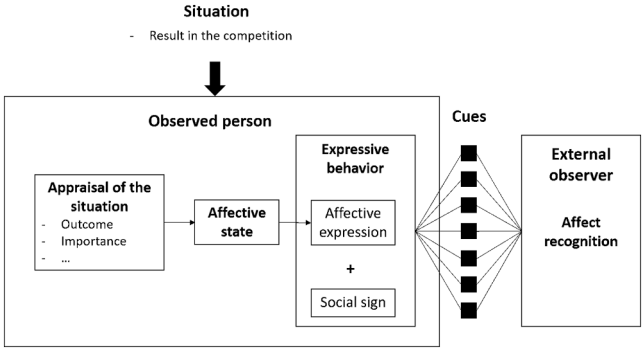


Fig. 1. Model of Affect Recognition.

notion, Fritsch, Preine [17] found that experts in tennis were more proficient at correctly identifying players' affective responses compared to those who were novices in the sport. This underscores the importance of experiential knowledge and nuanced understanding in the accurate assessment of affective states through observed expressive behavior.

The model delineated in Fig. 1 posits specific conditions under which the recognition of affective states from observed expressions can be successfully achieved. First, there must be a clear and discernible link between the situation in which the affect is expressed and the individual's affective state. Second, the intensity of the affective state must surpass a certain threshold to manifest externally. Finally, the observer—whether human or artificial—must accurately interpret the expressive behavior displayed. The successful recognition of the affective state from expressive behavior is contingent upon the concurrent fulfillment of these three conditions. The complex interplay among these elements highlights the intricate nature of affect recognition and underscores the necessity for a nuanced understanding of the underlying mechanisms.

1.2. Automate affect recognition based on AI concerning bodily expressions

Over the past two decades, the convergence of affective computing and machine learning has catalyzed the development of diverse methodologies for automated affect recognition, leveraging data from physiological, auditory, textual, and visual sources [33]. Historically, the focus has largely been on facial expressions for emotion recognition, constituting approximately 90 % of research in the field [33]. Recent shifts towards bodily expressions underscore the expanding interest in non-facial cues [34], echoing broader psychological insights that emphasize the conveyance of affect through body language [35]. Advancements in capturing body postures and movements via video cameras, motion capture, and wearable technologies have broadened the scope of research [36–38]. Benchmark databases like FABO [39], THEATER Corpus [40], GEMEP [41], and EMILYA [42] play a pivotal role in emotion recognition analysis. However, these databases often rely on annotations from acted expressions or external human raters, which may not accurately reflect real-world affective states [43]. This discrepancy highlights the challenges in dataset construction and annotation, as identified by Jemioło et al. [44], emphasizing the need for datasets that more authentically represent the complexity of affective experiences in real-life contexts.

Affective computing is the study and development of systems and devices that can recognize, interpret, process, and simulate human affects, essentially bridging the gap between human emotional experiences and computational technology [45]. While earlier affective computing relied on relatively simple statistical methods, such as decision trees [46], Hidden Naive Bayes [47], analysis of variance [48], or cluster analysis, recent advancements have embraced deep learning techniques like Convolutional Neural Networks (CNN) and Recurrent Neural Network (RNN) [49]. While some studies report high accuracy levels, up to 100 %, it's critical to consider these results in the context of the number of emotion categories evaluated. Such high accuracy rates often depend on the dataset's complexity and the specific scenarios tested [50]. Hence, a comprehensive understanding of these factors is essential before drawing comparisons [44]. Recognizing the importance of context, our study emphasizes the need for affect recognition research to pivot towards naturalistic scenarios that reflect the complexity and spontaneity of human emotions, addressing the challenges of annotation validity and ensuring that the affective states recognized by AI systems closely mirror those experienced in everyday life.

1.3. Objectives of the present study

In light of the prevalent challenges and limitations within the realm of affective recognition, particularly issues concerning the validity of annotations and the essential need for real-world affective assessment, this study sets forth two primary objectives. The first objective of the present study is the development of an AI system specifically designed for recognizing affective states in real-life situations, as opposed to relying on data from acted or posed situations. Recognizing affective states in real-life contexts offers superior utility, as it mitigates the inherent limitations of using actor-based data, which may not capture the nuanced, spontaneous expressions of emotion that occur in everyday life. This focus on real-world applicability aligns with emerging research emphasizing body language as a more informative medium for gauging affective states [39]. The second objective is to rigorously assess the performance of the developed affect recognition system. In striving to achieve these objectives, this study aspires to introduce a more nuanced and applicable methodology to the field of affective recognition, thereby filling existing research gaps and enabling more accurate emotional understanding in real-world situations.

2. Methods

2.1. Data

The data for this study were derived from tennis matches involving 15 players (5 females and 10 males), all competing in official contexts such as tournaments or league matches. These players, aged between 25 and 52 years with an average age of 39.2 (SD = 7.1) years, were amateurs playing at a competitive level, classified within the German Performance Class System (Leistungsklassensystem). The composition of our participant pool was influenced by the availability of athletes who consented to participate and could be filmed during competition. We acknowledge that the gender ratio in our sample does not reflect an equal distribution; however, it mirrors the general participation demographics within the competitive contexts accessible to our research, considering the constraints of athlete availability and willingness to participate. Ethical considerations were paramount in the design and execution of our study. In compliance with the guidelines of the Ethics Committee of the Karlsruhe Institute of Technology, we obtained explicit consent from all participants involved, adhering to strict privacy policies to ensure the integrity and confidentiality of the data collected. Each tennis match was recorded with a focus on capturing the expressive behavior of one player at a time, further reflecting our commitment to conducting our research in an ethically responsible manner.

For each point contested within a match, detailed records were maintained, including both the score and the outcome of the point. Videographic data was captured using a Sony HDR-CX240E Handycam (Full HD 1920×1080), equipped with a back-illuminated Exmor R CMOS sensor for superior low-light performance. This camera, known for its clear image zoom up to 54x and SteadyShot image stabilization, was placed near the net to capture one player from a frontal perspective. To ensure a consistent frontal view, the camera's orientation was adjusted to the opposite side of the court whenever players switched sides (see Fig. 2).

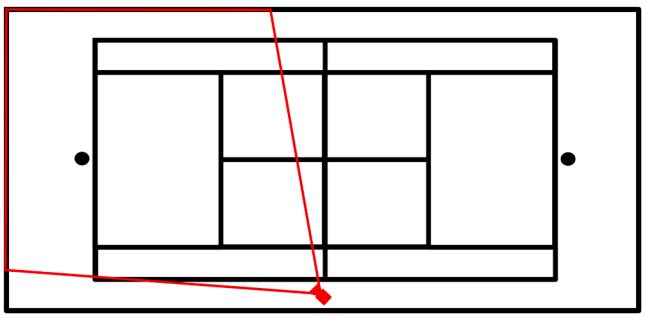


Fig. 2. The placement and orientation of the camera on the court.

Subsequently, the video footage was meticulously edited to isolate the segments displaying players' expressive behavior in the immediate aftermath of a rally until the commencement of the subsequent rally. The selection criteria for individual points included within the study stipulated that the ball's trajectory must not allow inferences regarding the point's outcome, and that the player's expressive behavior must be discernible in the video recording. These criteria were strictly adhered to, enabling us to compile 622 analyzable match scenes from a larger pool of video material, where the proportion of points won and lost was nearly identical. To avoid introducing any asymmetries in the dataset and ensure equitable learning conditions for the AI, we employed a stratified sampling method to deliberately select an equal number of scenes from won and lost points. This approach allowed us to achieve a balanced distribution, resulting in an even allocation of 311 points won and 311 points lost. The stratified sampling ensured that the selection was not random but systematically aligned with our objective to maintain parity between the outcomes represented in our dataset.

The collected video footage of the match scenes underwent a subsequent editing process to generate three distinct versions, each differing in length. The first version captured the initial three seconds following the conclusion of the rally, while the second version encompassed the first ten seconds post-rally. The third version included the entire duration of the pause between rallies; however, if a video scene's length was less than ten seconds in this version, it was allocated to the second version category. Empirical findings from studies involving human observers have demonstrated that the recognition rate, when utilizing video excerpts from the first version (i.e., the initial three seconds after the point's end), was comparable to those obtained from the second or third versions of the footage [17]. As a result, it was concluded that the first three seconds of video would suffice for a human, and therefore also for a neural network, to accurately recognize a player's affective state. Owing to specific computational constraints, the analysis actually utilized a slightly extended timeframe—specifically, the first 160 frames were extracted at a 50 FPS rate. This particular frame count was chosen as it is divisible by 32, conforming to the input requirements of the pre-trained pose network deployed in part of this study.

As delineated in Table 1, the complete dataset comprising 622 match scenes was randomly partitioned into two distinct subsets to evaluate the model's performance. Specifically, 500 videos were allocated to the training set for the purpose of training the model, while the remaining 122 videos were designated as a test set. To ensure representativeness, an equal distribution of points won and lost was maintained within both the training and test sets.

A neural network was trained to predict the expressive behavior corresponding to positive or negative affect as exhibited in the video recordings. To facilitate this analysis, joint positions were extracted for the players in the videos, allowing the expressive behavior to be assessed using either pose data alone, image data alone, or a combination of both, utilizing a pre-existing and pre-trained neural network from MMPose [MMPose Contributors, 51]. This particular neural network accepts an image as input and returns the positional coordinates of every joint for each person depicted in the image. The output format is aligned with the COCO (Common Objects in Context) dataset and adheres to a specific sequence of joint positions.

Subsequently, this format was mapped to the OpenPose joint position order, where the neck position—absent in the COCO dataset—was computed as the midpoint between both shoulders. The transformation of the joint position format was necessitated to enable joint position tracking within the videos and to facilitate the creation of skeleton-images, following the procedure detailed by Schneider, Sarfraz [52]. Their work, involving activity recognition, utilized the OpenPose joint position format. In accordance with the methodology employed in our study, joint positions were extracted for each image independently, thus lacking continuity in the mapping of joint position groups to an individual person (see Fig. 3). To establish this continuity, a separate program was implemented by Schneider, Sarfraz [52] to analyze the distances between joint positions from one frame to another. Through a comparative analysis of these distances, different tracks of joint positions were identified, continuously spanning multiple images. The length of these tracks is indicative of the duration a person remained visible in the video.

In the analysis of each video, multiple tracks were identified corresponding to various individuals present within the frame. This presented a challenge, as only the individual playing in the foreground was pertinent to the study, while others, such as spectators or individuals participating in adjacent games, were not relevant. To differentiate between the subject of interest and irrelevant individuals, a tailored algorithm was employed to define bounding boxes around each person in the video sequence. The individual corresponding to the tallest bounding box for the majority of the video's duration was identified as the primary subject, a criterion selected on the basis that the person of interest, being closest to the camera, appeared taller than others in the background. This specific track was consequently designated as the primary subject's track for the analysis.

For the specific condition involving pose, the relevant pose data were extracted from the video sequences, focusing on the identified tracks of joint positions over time for the main subject. Utilizing skeleton-image encoding, a methodology described by Caetano, Sena [53], the trajectories of the joint positions were processed. Both magnitude and orientation were encoded, with each stored in a distinct file for every video and encompassing the entire track. Within the resulting matrix, the rows were structured to encode the spatial configuration of the joints, while the depth layers of the image were determined as one for the magnitude and three for the orientation matrices. By maintaining the magnitude and orientation as separate files, the data were prepared in a format amenable for direct utilization as input to a neural network, facilitating subsequent analysis.

In the case of image data, it was imperative that the neural network's focus be directed solely toward the relevant player within the video. To achieve this, the portion of the image encapsulating the player was isolated and cropped. Utilizing the player's joint positions, which were previously extracted, a bounding box containing all of the player's joints was defined. This bounding box, complete with additional padding, was

Table 1
Training and test set split.

	N	N win	N loose
Train	500	250	250
Test	122	61	61

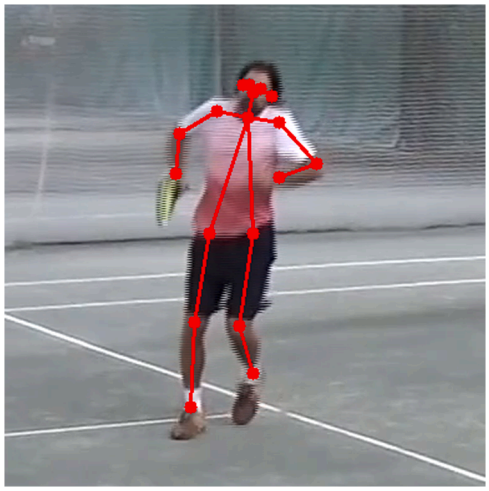


Fig. 3. Sample cropped image with visualized joint positons.

then cut out to create a fixed square size with the player centrally positioned. The inclusion of padding ensured that the player remained fully within the frame, compensating for potential inaccuracies in joint position detection. In instances where players momentarily exited the frame, a bounding box of fixed size was maintained from the point at the frame's edge where the player departed, holding this position until the player's return. This occurrence was rare, and each instance was meticulously reviewed to ensure consistency in the approach.

2.2. Model structure

The structure of the model is delineated into two primary networks: the pose network and the image network. The pose network operates on pose data, specifically utilizing skeleton images as fabricated in the previous section, whereas the image network directly engages with the image data. These two networks are subsequently integrated into a composite network, amalgamating information from both the pose and image networks. Generally, the input to these networks comprises information from 32 frames, though an extended sequence length of 64 frames was also investigated to discern the impact of sequence duration. The unifying output across all networks is a one-dimensional binary discrete prediction, discerning whether the preceding point in a given situation was lost or won.

The architecture of the pose network is conceived as a CNN. Predicated on a framework proposed and pre-trained by Schneider, Sarfraz [52] for an action recognition classification task, the architecture undergoes adaptation by replacing the final two fully connected layers with new ones to facilitate retraining for the affect recognition task (see upper part of Fig. 4). Conversely, the image network is constructed on the foundations of the S3D network, as proposed by Min and Corso [54], and pretrained on the Kinetics-400 dataset for an analogous action recognition classification task [55]. Manifested as a 3D-CNN network, it augments the traditional CNN architecture by introducing time as a third dimension. To allow retraining for affect recognition, the terminal convolutional layer is replaced by two new convolutional layers (see lower part of Fig. 3).

The outputs of both networks are flattened and concatenated following the final layers, culminating in a unified feature vector. This vector is subsequently subjected to a linear layer, culminating in the final classification result (see Fig. 3). The employment of the S3D network architecture prescribes a fixed timeframe for analysis, an attribute not necessary with an RNN architecture. Within this particular application, the emphasis is placed on the examination of discrete temporal segments, seeking cues within the player's body language. To facilitate the identification of such cues, the input was methodically confined to a fixed size, encompassing either 32 or 64 frames, as dictated by the specifications of the network.

To ensure a comprehensive and robust evaluation of the model's performance, we implemented four distinct test modes, as detailed in Table 2. Test Mode 1 entailed the analysis of a 32-frame segment, chosen randomly from the preliminary 160 frames of the video. Test Mode 2 examined the video in its entirety, utilizing all 160 frames. Test Modes 3 and 4 both utilized the full span of 160 frames but divided the frames into five equidistant segments, each comprising 32 frames. Consequently, both these modes generated five distinct neural network predictions. For Test Mode 3, these outcomes were averaged to yield a collective prediction. Test Mode 4, in contrast, relied on a majority voting mechanism wherein individual predictions were rounded based on their proximity to a median value. This approach mainly manifested divergent results when the predictions hovered around the midpoint; otherwise, the outcomes between the two methods were largely

Table 2
Settings for different test modes.

Test Mode	# frames	# samples	selection method	combination method
1	32	1	Random	–
2	160	1	All	–
3	32	5	Sequential	Average
4	32	5	Sequential	Majority voting

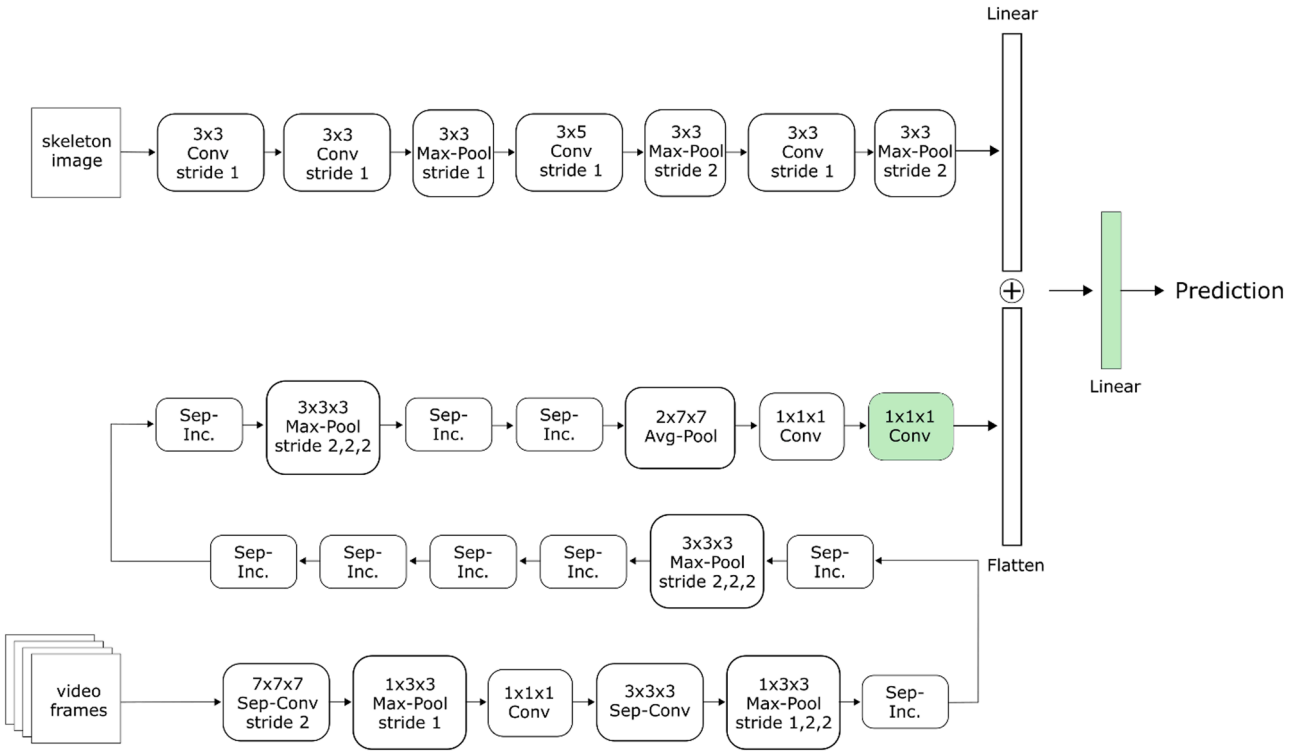


Fig. 4. Architecture of the combined network with pose network at the top and the image network at the bottom. Note: The green parts represent the adjusted layers of the original networks.

congruent.

In the computational setup for training our models, we utilized servers equipped with Nvidia GeForce graphics cards of the GTX series, which feature up to 12GB of graphics memory. In the process of model training, three distinct network types were employed: the pose network, the image network, and a combined network variant. These networks were subject to training using various hyperparameters, the objective being to ascertain the network configuration that yielded the most efficacious performance metrics. The hyperparameters selected for the experimental comparison of these models encompassed the optimizer, the learning rate, and the decision as to whether the entire pre-trained model would be fine-tuned to the newly acquired data. Within the realm of optimizers, both the Adaptive Moment Estimation (Adam) optimizer and the Stochastic Gradient Descent (SGD) optimizer were subjected to comparative analysis. The learning rates that were evaluated spanned a range, varying in powers of ten, from 0.1000 down to 0.0001. The specific hyperparameters corresponding to the models that demonstrated the most favorable results are documented in Table 3.

2.3. Statistical analysis

In order to assess the performance of the models presented in Table 3 this study employed metrics such as the accuracies, precision, recall, F1 score, and ROC-AUC score. The accuracy metric reflects the proportion of all predictions that the AI got correct, be it for won or lost points. An accuracy of 0.50 corresponds to the guess probability. A high accuracy indicates a significant alignment between the AI's estimations and the actual point outcomes. Precision denotes the proportion of points the AI correctly identified as 'won' out of all its 'won' predictions. Recall provides insight into the AI's ability to detect truly won points. It measures the fraction of actual 'won' points that the AI managed to correctly identify. F1 Score is a harmonic mean of precision and recall, it balances the trade-off between precision and recall. It is especially valuable when class distributions are imbalanced. The Negative Predictive Value (NPV) is defined as the proportion of accurately identified lost points relative to the total number of lost points. Conversely, the Positive Predictive Value (PPV) quantifies the likelihood of correctly identifying won points among the overall set of won points. Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) metric is instrumental in gauging the AI's capacity to distinguish between won and lost points. The ROC-AUC of 0.50 corresponds to the guess probability. A high ROC-AUC score reveals that the model can effectively differentiate outcomes based on body language, independent of the specific classification threshold chosen. ROC-AUC is equivalent to accuracy when the outputs can only take discrete binary values.

3. Results

Table 4 elucidates the performance characteristics of the top seven models, as further outlined in Table 3. The accuracy ranged from 63.9 % to 68.9 %, with the image network basing on sequence length of 64 frames in test mode 1 (Image 64 t1) achieving the highest accuracy of 68.9 %. In contrast, both Pose t1 and Image t2 demonstrated the lowest accuracy of the seven best models used in this study, standing at 63.9 %. Precision values varied considerably, from 62.7 % (Image t2) to 80.6 % (Image 64 t3). Likewise, recall values exhibited a range between 41.0 % (Image 64 t3) and 80.3 % (Combined t1), demonstrating the trade-off between the two metrics. The F1 score, which provides a balanced view of precision and recall, had a maximum value of 71.0 % for Combined t1. Image 64 t1 and Image t3 followed closely, registering F1 scores of 64.2 % and 66.7 % respectively. NPV ranged from 60.4 % (Image 64 t3) to 73.3 % (Combined t1), while PPV spanned between 62.7 % (Image t2) and 80.6 % (Image 64 t3). It should be noted that the ROC-AUC values mirror the accuracy values because our output only took discrete binary values (point won vs. point lost) without rounding.

4. Discussion

The primary aim of this research was to develop a machine learning model designed to identify affective states by analyzing bodily expressions, using video footage from tennis matches as a data source. Our approach distinguishes itself primarily by its focus on real-life sports situations for model training, in contrast to other methodologies that often rely on data obtained in controlled laboratory settings with actors. We contend that by training our model on more naturalistic contexts, we are better positioned to make reliable predictions in real-world scenarios. While this work strives to be a valuable contribution to the emerging field of applying AI algorithms for emotional state detection in real-world athletic settings, it is important to acknowledge that the field is rapidly evolving, and other parallel efforts may also be in progress.

In the current investigation, the most robust accuracy of 68.9 % was achieved by utilizing an image-based neural network with a sequence length of 64 frames in Test Mode 1. This performance aligns well with existing literature in the domain of gesture-based emotion recognition. Comparing our results with state-of-the-art methods highlights the contributions of our study to the field of affective computing. Recent studies, such as those by Ly, Lee, Kim, and Yang [56], have reported accuracies in the vicinity of 67.5 % utilizing a combination of CNN and ConvLSTM networks, drawing from databases like FABO. Similarly, Avola et al.'s research [57], employing an LSTM-MLP hybrid model on the UCLIC Affective Body Posture and Motion database, achieved accuracies between 35.0 % and 78.9 % for different affective postures. In the sports context, however, we are not aware of any study that has examined the accuracy of emotion recognition by AI. The advantage of our method is underscored by its application in real-life contexts, providing a more accurate reflection of genuine affective states as opposed to the controlled or acted expressions often used in training datasets. This aspect underlines the significance of our work within the broader trajectory of affective computing research, marking a step forward in the development of AI systems capable of emotion recognition in complex, real-life situations.

For transparency, comparisons with human observers have been made, and our trained model shows comparable, if not superior, performance. The studies that were based on the same sample of video footage revealed an average human observer accuracy between 55.9 % and 63.0 %, with individual rates reaching as high as 75.0 % [17,21],

Table 3
Overview of the optimal hyperparameter configurations for the models.

Model	Model type	Test Mode	Optimizer	Learning rate	Fine-tuning	Seq length
1. Pose t1	Pose	1	SGD	0.0001	False	32
2. Image t1	Image	1	SGD	0.0010	True	32
3. Image t2	Image	2	Adam	0.0010	True	32
4. Image t3	Image	3	Adam	0.1000	False	32
5. Image 64 t1	Image	4	Adam	0.0010	False	64
6. Image 64 t3	Image	1	SGD	0.0010	False	64
7. Combined t1	Combined	1	SGD	0.1000	True	32

Note: pose = model type pose; image = model type image; combined = model type combined; t1 = test mode 1; t2 = test mode 2; t3 = test mode 3; t4 = test mode 4; 64 = sequence length of 64 frames.

Table 4
Overview of the performance indicators.

Model	True Positives		True Negatives		False Positives		False Negatives		ACC	F1	Precision	Recall	NPV	PPV	ROC-AUC score
	n	%	n	%	n	%	n	%							
Pose t1	33	27.0	45	36.9	16	13.1	28	23.0	63.9	60.0	67.4	54.1	61.6	67.3	0.639
Image t1	34	27.9	45	36.9	16	13.1	27	22.1	64.8	61.3	68.0	55.7	62.5	68.0	0.648
Image t2	42	34.4	36	29.5	25	20.5	19	15.6	63.9	65.6	62.7	68.9	65.5	62.7	0.639
Image t3	39	32.0	44	36.1	17	13.9	22	18.0	68.0	66.7	69.6	63.9	66.7	69.6	0.680
Image 64 t1	34	27.9	50	41.0	11	9.0	27	22.1	68.9	64.2	75.6	55.7	64.9	75.6	0.689
Image 64 t3	25	20.5	55	45.1	6	4.9	36	29.5	65.6	54.3	80.6	41.0	60.4	80.6	0.656
Combined t1	49	40.2	33	27.0	28	23.0	12	9.8	67.2	71.0	63.6	80.3	73.3	63.6	0.672

Note: pose = model type pose; image = model type image; combined = model type combined; t1 = test mode 1; t2 = test mode 2; t3 = test mode 3; t4 = test mode 4; 64 = sequence length of 64 frames; ACC = accuracy; NPV = Negative Predictive Value; PPV = Positive Predictive Value; ROC-AUC score = Receiver Operating Characteristic - Area Under the Curve.

and are thus comparable to the accuracies achieved by our CNN models. This suggests that the machine learning algorithms applied here offer equal or greater proficiency in identifying affective states from expressive behaviors of tennis players compared to human observers. However, it's worth noting that both human cognition and AI present unique challenges in accurately recognizing affective states. Human affect recognition may be influenced by subjective biases [58], emotional state [59], or cultural background [60], which can impede objective assessment. On the other hand, AI models like CNNs may struggle with issues related to the learning with imbalanced data, the validity of training data, generalization to diverse situations, or interpreting nuanced, context-dependent expressions [61]. These issues highlight the inherent biases in AI methodologies, emphasizing the need for careful consideration in the design and application of machine learning algorithms for affect recognition.

A closer examination of performance metrics reveals intriguing patterns, such as the Image 64 t3 model's 90.2 % accuracy rate in identifying negative affective states for points lost. This high accuracy underscores the model's sensitivity to subtle cues indicative of negative emotions, a finding that aligns with psychological theories suggesting that humans are more attuned to recognizing negative emotional expressions due to their evolutionary significance [62]. The trend observed in our models, where accuracy for negative states generally exceeds that for positive states, is consistent with findings from human-based studies [17,24], where human observers showed a higher recognition rate for negative affective states (ranging from 60.1 % to 68.3 %) compared to positive affective states (ranging from 51.1 % to 56.6 %). This context helps to understand the significance of the 90.2 % figure, which highlights the advanced capabilities of CNN models in emotion recognition, especially for negative states, compared to traditional human observer-based assessments.

A closer examination of performance metrics reveals some intriguing patterns. For instance, the Combined T1 model displayed an accuracy rate of 80.3 % for identifying positive affective states during points won, whereas the Image 64 t3 model achieved a 90.2 % accuracy rate for detecting negative affective states during points lost (as mentioned in the previous paragraph). In five out of the seven models under investigation, the accuracy for identifying negative affective states outstripped that for positive states.

The propensity for both CNNs and human observers to more accurately identify negative affective states than positive ones may be multifaceted, encompassing both evolutionary underpinnings and psychosocial factors. From an evolutionary perspective, the ability to rapidly recognize and respond to negative emotional states is advantageous, as these states often signal imminent threats or unfavorable circumstances requiring swift action [63]. Additionally, negative emotions tend to manifest through more salient and universally recognizable facial expressions, body language, and physiological changes, making them easier to detect. In social contexts, the accurate identification of negative emotions is crucial for maintaining social cohesion, as it allows for timely intervention and support [64]. Moreover, empirical studies suggest that individuals often experience negative affective states with greater intensity than positive ones, further amplifying the physiological and expressive cues associated with them [62]. Consequently, this heightened intensity might make negative affective states more readily recognizable. Collectively, these factors suggest that both machine learning algorithms and human cognitive processes are more finely tuned to the cues and expressions associated with negative affective states, thereby accounting for the higher rates of accurate identification.

The study employs an innovative methodology for the accurate assessment of affective states in real-world settings, specifically utilizing objective situation descriptors like "points won" or "points lost" in tennis matches. This methodology is strengthened by the standardized rules of the sport, which generally yield consistent appraisal mechanisms across players, thereby enhancing the accuracy of interpretive analyses. However, the accuracy of this method for recognizing affective states is conditional upon three primary factors. First, it assumes a uniformity in how players evaluate situations, a premise that, while supported to some extent by existing studies [25], still necessitates further empirical validation for accuracy. Second, the accuracy of affective state recognition can be affected by the level of affective expressivity displayed by the observed individual, and may be compromised if true affective states are masked for social or strategic reasons. Lastly, the accurate recognition of affective states hinges on the precise interpretation of expressive cues, a challenge that presents inherent difficulties for both human evaluators and AI systems. Consequently, achieving complete accuracy in affect recognition in real-world settings remains an aspirational goal, given the three conditions that influence the detection and interpretation of expressive cues have yet to be fully satisfied by either human observers or AI systems.

Furthermore, we acknowledge the critical importance of fairness in AI and affective computing research. While our study focused on the technical aspects of emotion recognition, the aspect of fairness—ensuring equitable and bias-free models—remains a paramount concern for future work [65]. As AI technologies continue to evolve, dedicating efforts to understand and mitigate potential biases becomes essential to developing tools that are just and equitable for all users.

4.1. Strength and limitations

This study has a number of strengths and limitations. A significant strength of this study lies in the annotation method employed. By referring to objective indicators in situational contexts, the study advances a novel form of affective state assessment. Furthermore, the use of real-life video footage from actual tennis matches enhances the ecological validity of the research. This provides an understanding of how affective states are naturally experienced and expressed.

One clear limitation is the restricted size of the video dataset. A limited dataset size constrains the model's learning capacity, potentially affecting the accuracy and generalizability of the neural network. Additionally, the small test set of 122 videos may result in inflated differences between model performances. Additionally, all video footage

was confined to a tennis hall environment, subjecting the data to specific lighting conditions that occasionally impaired visibility. Further, the participant demographic was limited to competitive tennis players of young to middle adult age from the European region. A more diverse participant pool across skill levels, ages, and ethnic backgrounds could enrich the dataset and improve the model's applicability. In addition, the study could benefit from higher-resolution video capture to glean more nuanced details of expressive behavior. The involvement of a professional camera operator could further refine the quality of the collected data. Finally, creating high-quality videographic data in real-life scenarios presents its own set of challenges, including varying environmental conditions and the unpredictability of live events, which can complicate data collection and affect the fidelity of the captured footage.

4.2. Implications for research and practice

To mitigate these limitations, subsequent research should focus on expanding the dataset both in terms of volume and diversity. High-resolution video equipment and professional camera operation could also enhance data quality. If these limitations are addressed and a sufficiently accurate neural network is developed, the system's application could potentially extend to various other domains such as healthcare, automotive safety, and workplace environments. Such advancements would not only improve the understanding of human affective expressive behavior but also facilitate the development of non-invasive, reliable systems for measuring affective states across diverse contexts.

In addressing the critical need for comparative analysis between AI and human observers in recognizing affective states, future research should systematically explore the distinctions and similarities in their performance. Such studies could offer invaluable insights into the precision, biases, and efficiency of both AI systems and human judgment in interpreting expressive behavior. This endeavor not only holds the promise of advancing our understanding of affective state recognition technologies but also of enhancing the applicability of these tools in real-world scenarios where nuanced interpretation of human emotions is crucial.

The findings from this study underscore the importance of exploring machine learning methodologies for affective state recognition in sports contexts. One clear avenue for future research is to investigate the observed divergence between machine and human performance in recognizing positive versus negative affective states. Understanding the biases or strengths in the employed algorithms could offer insights into the mechanisms driving these variances. Further studies could also delve into the effect of sequence length and test modes on model performance, potentially contributing to the refinement of future affective recognition systems. Moreover, the observed limitations in affective state recognition, both in human observers and AI systems, call for multi-disciplinary research involving psychology, data science, and sports science. Such collaborations could facilitate the development of more nuanced and accurate models that better mimic human appraisal mechanisms.

For professionals in the sports sector, the study's outcomes could be particularly impactful. Coaches, athletes, and sport psychologists could benefit from a system capable of recognizing affective states, allowing for more targeted emotional regulation strategies. Automated real-time emotional analysis could provide a data-driven basis for tactical or motivational adjustments during a match or training. Furthermore, sports organizations could potentially adopt these machine-learning models for fan engagement activities, offering real-time emotional narratives of ongoing matches. However, caution must be exercised in the ethical application of such technologies, especially in regard to player consent and data privacy.

The potential ramifications of this research extend far beyond the sports arena, offering transformative possibilities across multiple sectors. In healthcare, the algorithmic recognition of affective states could significantly enhance patient monitoring and personalized care, potentially identifying early signs of emotional or psychological distress that may otherwise go unnoticed. In automotive safety, an understanding of the driver's emotional state could trigger appropriate preventive measures, such as alerting systems or automated driving controls, to minimize the risk of accidents caused by emotional impairment. Similarly, in sectors like education, customer service, and mental health, the nuanced understanding and recognition of human emotions could pave the way for more empathetic and effective interactions. Also, in areas such as robotics or Human-AI Interaction, successful detection of the user's affective state can influence how machines act and behave towards humans and thus yield a better user experience. Future research could focus on adapting the algorithms for these specific applications, including considerations for ethical implications and data privacy, to fully harness the potential of affective state recognition in creating more adaptive and responsive systems across these diverse fields.

5. Conclusion

In summary, this study serves as a milestone in the utilization of machine learning for affective state recognition in real-world sports contexts. Our findings indicate that machine learning models can outperform human observers in affective state recognition. Notably, both machine learning models and human observers displayed a common tendency to more accurately recognize negative affective states, potentially due to their heightened intensity and more explicit expression. The implications of this research are far-reaching, impacting sectors such as healthcare and automotive safety where understanding and recognizing human emotions can be transformative. By laying the foundation for substantial improvements through larger and more diverse datasets and higher-quality video capture, this study sets the stage for the development of non-invasive, highly accurate systems for detecting human affective states across a range of applications.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used DeepL and ChatGTP in order to increase readability of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Darko Jekauc: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Formal analysis, Conceptualization. **Diana Burkart:** Visualization, Validation, Software, Investigation, Formal analysis, Data curation. **Julian Fritsch:** Writing – review & editing, Resources, Methodology. **Marc Hesenius:** Writing – review & editing, Supervision. **Ole Meyer:** Writing – review & editing, Supervision. **Saquib Sarfraz:** Writing – review & editing, Supervision, Software, Resources. **Rainer Stiefelhagen:** Writing – review & editing, Supervision, Resources, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Funding

This research did not receive any specific grant from funding

agencies in the public, commercial, or not-for-profit sectors.

References

[1] D. Kahneman, Thinking, fast and slow, Pinguin Group, New York, NY, 2011.
[2] R.P. Bagozzi, U.M. Dholakia, S. Basuroy, How effortful decisions get enacted: the motivating role of decision processes, desires, and anticipated emotions, J. Behav. Decis. Mak. 16 (4) (2003) 273–295.
[3] M. Zeidner, G. Matthews, R.D. Roberts, Emotional intelligence in the workplace: a critical review, Appl. Psych. 53 (3) (2004) 371–399.
[4] J.A. Russell, A circumplex model of affect, J. Pers. Soc. Psychol. 39 (6) (1980) 1161–1178.
[5] P. Ekman, An argument for basic emotions, Cogn. Emot. 6 (3–4) (1992) 169–200.
[6] R.S. Lazarus, Relational meaning and discrete emotions, editors, in: KR Scherer, A Schorr, T Johnston (Eds.), Appraisal Processes in Emotion, Oxford University Press, New York, NY, 2001.
[7] A. Ortony, G.L. Clore, A. Collins, The cognitive structure of emotions, Cambridge University Press, New York, NY, 1990.
[8] R.F. Baumeister, K.D. Vohs, N.C. DeWall, L. Zhang, How emotion shapes behavior: feedback, anticipation, and reflection, rather than direct causation, Personal. Social Psychol. Rev. 11 (2) (2007) 167–203.
[9] L.F. Barrett, C.D. Wilson-Mendenhall, L.W. Barsalou, The conceptual act theory: a roadmap, editors, in: L Feldman Barrett, JA Russell (Eds.), The psychological construction of emotion, The Guilford Press, New York, NY, US, 2015, pp. 83–110.
[10] K.R. Scherer, What are emotions? And how can they be measured? Social Sci. Informat. 44 (4) (2005) 695–729.
[11] H.L. Meiselman, Emotion measurement, Woodhead Publishing, 2016.
[12] P.A. Kragel, K.S. LaBar, Decoding the nature of emotion in the brain, Trends Cogn. Sci. (Regul. Ed.) 20 (6) (2016) 444–455.
[13] M. Egger, M. Ley, S. Hanke, Emotion recognition from physiological signal analysis: a review, Electron. Notes. Theor. Comput. Sci. 343 (2019) 35–55.
[14] L.A.R. Sacrey, S. Raza, V. Armstrong, J.A. Brian, A. Kushki, I.M. Smith, et al., Physiological measurement of emotion from infancy to preschool: a systematic review and meta-analysis, Brain Behav. 11 (2) (2021) e01989.
[15] D. Keltner, J. Tracy, D.A. Sauter, D.C. Cordaro, G. McNeil, Expression, in: LF Barrett, M Lewis, JM Haviland-Jones (Eds.), Handbook of Emotion, 42016, 2016, pp. 467–482.
[16] I.B. Mauss, M.D. Robinson, Measures of emotion: a review, Cognit. Emot. 23 (2) (2009) 209–237.
[17] J. Fritsch, L. Preine, D. Jekauc, The examination of factors influencing the recognition of affective states associated with tennis players' non-verbal behaviour, Psychol. Sport Exerc. 61 (2022) 102206.
[18] H. Aviezer, Y. Trope, A. Todorov, Holistic person processing: faces with bodies tell the whole story, J. Pers. Soc. Psychol. 103 (1) (2012) 20–37.
[19] A. Kumar, A. Kaur, M. Kumar, Face detection techniques: a review, Artif. Intell. Rev. 52 (2) (2019) 927–948.
[20] Jekauc D.Emotionen im Sport, J. Appl. Sport Exerci. Psychol. 25 (2) (2018) 51–52.
[21] J. Fritsch, K. Seiler, M. Wagner, C. Englert, D. Jekauc, Can you tell who scores? An assessment of the recognition of affective states based on the nonverbal behavior of amateur tennis players in competitive matches, J. Sport Exerc. Psychol. 45 (3) (2023) 138–147.
[22] D. Jekauc, L. Mülberger, S. Weyland, Achtsamkeitstraining Im Sport: Das Übungsprogramm Zur Förderung der Sportlichen Leistungsfähigkeit, Springer, 2022.
[23] D. Jekauc, J. Fritsch, A.T. Latinjak, Toward a theory of emotions in competitive sports, Front. Psychol. 12 (6046) (2021).
[24] J. Fritsch, S. Ebert, D. Jekauc, The recognition of affective states associated with players' non-verbal behavior in volleyball, Psychol. Sport Exerc. 64 (2023) 102329.
[25] J. Fritsch, E. Finne, D. Jekauc, D. Zerdila, A.M. Elbe, A. Hatzigeorgiadis, Antecedents and consequences of outward emotional reactions in table tennis, Front. Psychol. 11 (2020) 578159.
[26] P. Furley, G. Schweizer, Body language in sport, editors, in: Gershon Tenenbaum, RC Eklund (Eds.), Handbook of sport psychology, John Wiley & Sons, 2020, pp. 1201–1219.
[27] P. Furley, G. Schweizer, G. Tenenbaum, R.C. Eklund, Body language in sport, editors. Handbook of sport psychology, John Wiley & Sons, Hoboken, NJ, 2020, pp. 1201–1219.
[28] J. Fritsch, D. Redlich, A. Latinjak, A. Hatzigeorgiadis, The behavioural component of emotions: exploring outward emotional reactions in table tennis, Int. J. Sport Exerc. Psychol. 20 (2) (2022) 397–415.
[29] E. Brunswik, The conceptual framework of psychology, University Chicago Press, Oxford, 1952.
[30] E.L Thorndike, Animal intelligence: an experimental study of the associative processes in animals, Psycholog. Rev. 2 (4) (1898) 1–109.
[31] G.R. Lefrancois, Theories of human learning - Mrs. Gribbin's Cat, Cambridge University Press, Cambridge, UK, 2019.
[32] Z.H. Zhou, Machine Learning (2021).
[33] S. Baloch, S.A. Rahman Abu-Bakar, M. Mohd Mokji, S Waseem, Affective computing and anger expression through bodily movements: a review, Social Sci. Res. Network [Internet] (2022). Available from: https://ssrn.com/abstract=4264031 [access date: 15.10.2023].
[34] B. de Gelder, A.W. de Borst, R. Watson, The perception of emotion in body expressions, WIREs Cognitive Sci. 6 (2) (2015) 149–158.
[35] H. Aviezer, Y. Trope, A. Todorov, Body cues, not facial expressions, discriminate between intense positive and negative emotions, Science (1979) 338 (6111) (2012) 1225–1229.
[36] M. Karg, A.A. Samadani, R. Gorbet, K. Kühnlenz, J. Hoey, D. Kulić, Body movements for affective expression: a survey of automatic recognition and generation, IEEe Trans. Affect. Comput. 4 (4) (2013) 341–359.
[37] F. Noroozi, C.A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, G. Anbarjafari, Survey on emotional body gesture recognition, IEEe Trans. Affect. Comput. 12 (2) (2018) 505–523.
[38] T. Sapiński, D. Kamińska, A. Pelikant, G. Anbarjafari, Emotion recognition from skeletal movements, Entropy 21 (7) (2019) 646.
[39] H. Gunes, C. Shan, S. Chen, Y. Tian, Bodily expression for automatic affect recognition, editors, in: A Konar, A Chakraborty (Eds.), Bodily Expression for Automatic Affect Recognition, Wiley, Hoboken, NJ, 2015, pp. 343–377.
[40] M. Kipp, J.C. Martin, Gesture and emotion: can basic gestural form features discriminate emotions?, in: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops IEEE, 2009.
[41] Bänziger T., Scherer K.R. Introducing the geneva multimodal emotion portrayal (gemep) corpus. In: Scherer KR, Bänziger T, Roesch E, editors. Blueprint For Affective computing: A sourcebook 2010. p. 271–94.
[42] N. Fourati, C. Pelachaud, Emilya: emotional body expression in daily actions database, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA), 2014.
[43] B.W. Schuller, Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends, Commun ACM 61 (5) (2018) 90–99.
[44] P. Jemioło, D. Storman, M. Mamica, M. Szymkowski, W. Żabicka, M. Wojtaszek-Główka, et al., Datasets for automated affect and emotion recognition from cardiovascular signals using artificial intelligence— a systematic review, Sensors 22 (7) (2022) 2538.
[45] J. Tao, T. Tan, Affective computing: a review. Affective computing and intelligent interaction, editors, in: J Tao, T Tan, RW Picard (Eds.), ACII, Springer, Beijing, 2005, pp. 981–995.
[46] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, G. Volpe, Multimodal analysis of expressive gesture in music and dance performances, in: Gesture-Based Communication in Human-Computer Interaction: 5th International Gesture Workshop, GW 2003, Genova, Italy, Springer, 2004. April 15-17, 2003, Selected Revised Papers 5.
[47] G. Castellano, S.D. Villalba, A. Camurri, editors, in: Recognising human emotions from body movement and gesture dynamics. International conference on affective computing and intelligent interaction, Springer, 2007.
[48] D. Glowinski, A. Camurri, G. Volpe, N. Dael, K. Scherer, Technique for automatic emotion recognition by body gesture analysis, in: 2008 IEEE Computer society conference on computer vision and pattern recognition workshops, IEEE, 2008.
[49] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, et al., A systematic review on affective computing: emotion models, databases, and recent advances, Informat. Fusion 83-84 (2022) 19–52.
[50] S.K. Khare, V. Blanes-Vidal, E.S. Nadimi, U.R. Acharya, Emotion recognition and artificial intelligence: a systematic review (2014–2023) and research recommendations, Informat. Fusion 102 (2024) 102019.
[51] Contributors M. Openmmlab pose estimation toolbox and benchmark. URL: https://github.com/open-mmlab/mmpose; 2020.
[52] D. Schneider, S. Sarfraz, A. Roitberg, R. Stiefelhagen, Pose-based contrastive learning for domain agnostic activity representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.
[53] Skelemotion: a new representation of skeleton joint sequences based on motion information for 3d action recognition, in: C Caetano, J Sena, F Brémond, JA Dos Santos, WR Schwartz (Eds.), 2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS), IEEE, 2019.
[54] K. Min, J.J. Corso, Tased-net: temporally-aggregating spatial encoder-decoder network for video saliency detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.
[55] Kay W., Carreira J., Simonyan K., Zhang B., Hillier C., Vijayanarasimhan S., et al. The kinetics human action video dataset. Computer Vision and Pattern Recognition. 2017;arXiv:1705.06950.
[56] Emotion recognition via body gesture: deep learning model coupled with keyframe selection, in: ST Ly, G-S Lee, S-H Kim, H-J Yang (Eds.), Proceedings of the 2018 international conference on machine learning and machine intelligence, 2018.
[57] D. Avola, L. Cinque, A. Fagioli, G.L. Foresti, C. Massaroni, Deep temporal analysis for non-acted body affect recognition, IEEe Trans. Affect. Comput. 13 (3) (2020) 1366–1377.
[58] V. Colonnello, P.M. Russo, K. Mattarozzi, First impression misleads emotion recognition, Front. Psychol. 10 (527) (2019).
[59] V. Ovsyannikova, Influence of emotional states on emotion recognition, Psychol. J. Higher School Econ. 11 (1) (2014) 86–101.
[60] U. Schimmack, Cultural Influences on the Recognition of Emotion by Facial Expressions:individualistic or Caucasian Cultures? J. Cross. Cult. Psychol. 27 (1) (1996) 37–50.
[61] D. Dablain, K.N. Jacobson, C. Bellinger, M. Roberts, N.V. Chawla, Understanding CNN fragility when learning with imbalanced data, Mach. Learn. (2023) 1–26.
[62] R.F. Baumeister, E. Bratslavsky, C. Finkenauer, K.D. Vohs, Bad is stronger than good, Rev. General Psychol. 5 (4) (2001) 323–370.

[63] M. Mendl, O.H.P. Burman, R.M.A. Parker, E.S. Paul, Cognitive bias as an indicator of animal emotion and welfare: emerging evidence and underlying mechanisms, Appl. Anim. Behav. Sci. 118 (3) (2009) 161–181.

[64] J.P. Forgas, When sad is better than happy: negative affect can improve the quality and effectiveness of persuasive messages and social influence strategies, J. Exp. Soc. Psychol. 43 (4) (2007) 513–528.

[65] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A Galstyan, A survey on bias and fairness in machine learning, ACM. Comput. Surv. 54 (6) (2021). Article 115.