COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS PHYSICS AND INFORMATICS

EMOTION RECOGNITION SYSTEM FOR THE HUMANOID ROBOT NICO

Master thesis

2026 Bc. Šimon Strieška

COMENIUS UNIVERSITY IN BRATISLAVA
FACULTY OF MATHEMATICS PHYSICS AND INFORMATICS

EMOTION RECOGNITION SYSTEM FOR THE HUMANOID ROBOT NICO

Master thesis

Study program:	Applied informatics
Branch of study:	Applied informatics
Department:	Department of Applied Informatics
Supervisor:	RNDr. Kristína Malinovská, PhD.

Bratislava, 2026

Bc. Šimon Strieška

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta:	Bc. Šimon Strieška
Študijný program:	aplikovaná informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor:	informatika
Typ záverečnej práce:	diplomová
Jazyk záverečnej práce:	anglický
Sekundárny jazyk:	slovenský

Názov: Emotion recognition system for the humanoid robot NICO
Systém rozpoznávania emócii pre humanoidného robota NICO

Anotácia: Hlboké neurónové siete sú dnes štandardnom v strojovom učení vrátane spracovania obrazu. Používajú sa aj na rozlíšenie emócií ľudí na obrázkoch či videu [1] a často sa používajú v robotike cielenej na interakciu s ľuďmi. Pri vytváraní systémov pre rozpoznávanie emócii u ľudí sa stretávame s mnohými problémami, či už je to samotná zložitosť úlohy spracovania ľudských tvárí, ale aj klasické problémy strojového učenia ako nedostatočné množstvo dát či nerovnováha zastúpenia klasifikovaných tried v dátovej sade. Mnohé súčasné riešenia a databázy ako napríklad AffectNet [2] trpia práve týmto problémom. Riešení problému zlepšenia presnosti klasifikácie podhodnotených tried je mnoho, či už manipulácia s dátami ale aj augmentácia dát rôznymi metódami. Dnešné moderné metódy na báze adverzariálneho generatívneho učenia ponúkajú napríklad možnosti meniť tzv. osobné atribúty ľudí na fotografiách ako sú napríklad okuliare či vlasy. Ďalším úspešným mechanizmom pre získanie väčšieho množstva kvalitných ľudských dát je generovanie syntentických dát pomocou 3D grafického simulátora ako je napríklad Unreal Engine. Pre úspešné fungovanie rozoznávacieho systému pre robota je potrebné najprv natrénovať systém na dobrej sade dát a otestovať ho a následne ho kalibrovať pre zariadenie na ktorom slúži a otestovať v praxi.

Cieľ: Cieľom práce je preskúmať existujúce modely neurónových sietí pre rozpoznávanie emócii (napr. VGGFace2 a iné) a metódy ako možno tieto modely vylepšiť, či už pomocou vylepšenia samotnej siete ale hlavne vylepšením súčasných dostupných datasetov ako je napr. AffectNet [2] pomocou aj vyššie uvedených netriviálnych augmentácii dát. Pre pochopenie nedostatkov je potrebné analyzovať reprezentácie modelu (feature vectors) napríklad pomocou klasterizačných metód a ďalších techník vysvetliteľnej UI. Pri vývoji siete je súčasne potrebné myslieť na to, aby nebola príliš rozsiahla a bolo možné ju spustiť aj na bežnom hernom laptope aby bol systém prenosný s robotom. Hlavným cieľom práce je vo výsledku hlboká neurónová sieť, ktorú možno ďalej použiť pre výskum interakcie robota a človeka, ktorý vykonávame na katedre s robotom NICO [4].

Literatúra: [1] Li, S. and Deng, W. 2022. "Deep Facial Expression Recognition: A Survey," in IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195-1215. IEEE

[2] Mollahosseini, A., Hasani, B. and Mahoor, M.H., 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1), pp.18-31.
[3] Cao, Q., Shen, L., Xie, W., Parkhi, O.M. and Zisserman, A., 2018, May. Vggface2: A dataset for recognising faces across pose and age. In 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018) (pp. 67-74). IEEE.
[4] Kerzel, M. et al. "NICO - Neuro-inspired companion: A developmental humanoid robot platform for multimodal interaction," 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 2017, pp. 113-120.

Vedúci:	RNDr. Kristína Malinovská, PhD.	
Konzultant:	Ing. Branislav Zigo	
Katedra:	FMFI.KAI - Katedra aplikovanej informatiky	
Vedúci katedry:	doc. RNDr. Tatiana Jajcayová, PhD.	
Dátum zadania:	08.12.2024	
Dátum schválenia:	09.12.2024	prof. RNDr. Roman Ďurikovič, PhD. garant študijného programu

....

študent vedúci práce

I hereby declare that I have written this thesis by myself, only with help of referenced literature, under the careful supervision of my thesis advisor.

. . . .

Bratislava, 2026

Bc. Šimon Strieška

Acknowledgement

Abstract

Facial Expression Recognition (FER) is a technology for automatically distinguishing human emotions, one of the most common ways to implement FER is through deep neural networks. A deep neural network is a type of artificial neural network composed of multiple hidden layers of neurons, enabling the network to learn complex data properties. Deep neural network models focused on FER achieve the best results in research areas such as emotion recognition or Human-Robot Interaction (HRI). However, the existence of these models depends on extensive and diverse datasets, which often suffer from class imbalance. This imbalance leads to the creation of models that achieve suboptimal results when working with data from underrepresented classes. In our work, we experiment with existing methods to address the problem of class imbalance, which we apply to the ImageNetV1 model to achieve higher performance. The outcome of the work is an improved model that surpasses the original in both accuracy and robustness. The model can be utilized in future HRI research with the NICO robot.

Keywords: facial expression recognition, artificial neural networks, deep neural networks,

Abstrakt

Rozpoznávanie výrazov tváre (facial expression recognition, FER) je technológia pre automatické rozlišovanie ľudských emócii, jedným z najčastejších spôsobov implementácie FER je prostredníctvom hlbokej neurónovej siete. Hlboká neurónová sieť je druhom umelej neurónovej siete, ktorá je tvorená niekoľkými skrytými vrstvami neurónov, vďaka ktorým je sieť schopná naučiť sa komplexné vlastnosti dát. Modely hlbokých neurónových sietí zamerané na FER dosahujú najlepšie výsledky v oblastiach výskumu ako rozpoznávanie emócii alebo interakcia robota a človeka (human robot interaction, HRI). Avšak, existencia týchto modelov závisí na rozsiahlych a rozmanitých datasetoch, ktoré často trpia nevyváženosťou tried (class imbalance). Táto nevyváženosť vedie k tvorbe modelov, ktoré dosahujú neoptimálne výsledky pri práci s dátami z nedostatočne zastúpených tried. V našej práci experimentujeme s existujúcimi metódami pre riešenie problému nevyváženosti tried, ktoré aplikujeme na model ImageNetv1 s cieľom dosiahnúť vyššiu úspešnosť. Výsledkom práce je vylepšený model, ktorý prekonáva ten pôvodný v presnoti aj robustnosti. Model možno v budúcnosti využiť pri výskume HRI s robotom NICO.

Kľúčové slová: rozpoznávanie výrazov tváre, umelé neurónové siete, hlboké neurónové siete,

Contents

List of Figures

List of Tables

Terminology

Terms

Abbreviations

Motivation

Chapter 1

Introduction

Chapter 2

State of the Art

2.1 Introduction

Facial Emotion Recognition (FER) plays a central role in affective computing and in technologies that aim to interpret and react to human emotion. In human-robot interaction (HRI), FER enables humanoid platforms such as NICO to understand users more intuitively and to produce behavior that appears responsive [22]. The transition from hand-crafted feature extraction techniques to deep learning has significantly advanced the field, allowing modern systems to cope with challenges such as variation in illumination, occlusion, head poses and more. Architectures built on convolutional neural networks (CNNs) [24, 8] and transformer-based mechanisms [29] provide representational capacity needed for robust FER in real world environments.

Deploying FER on robotic hardware, however, demands far more than achieving good accuracy on benchmarks. Robots operate with limited computation, varied lighting conditions, moving subjects, and varying camera viewpoints. They must process facial information in real time, maintain reliability across diverse users, and offer interpretable behaviors that align with safety standards for HRI. FER datasets often suffer from class imbalance, noisy annotations, and inconsistent labeling [18]. These issues have to be mitigated using synthetic data generation and advanced augmentation strategies [3, 20, 16].

The following sections outline the theoretical foundations, classical approaches, modern deep-learning methodologies and HRI constraints that shape current FER systems.

2.2 Foundations of Emotion Recognition

2.2.1 Categorical Models

Many FER datasets and models rely on the categorical emotion framework introduced by Ekman, which proposes a set of basic and universal emotions such as happiness, sadness, anger, fear, disgust, surprise and contempt [5]. These categories serve as discrete labels that can be easily integrated into machine-learning classification tasks. The framework is widely used in large-scale FER datasets, including AffectNet [18], RAF-DB [12], and FER-2013 [7], and has remained dominant because its simplicity works well with the structure of deep classification architectures.

2.2.2 Dimensional Models

In contrast to categorical labels, dimensional models describe affect on continuous axes of valence and arousal. This representation focuses on emotional intensity and the transfer between different emotional states. Dimensional approaches are particularly valuable when a system must interpret subtle or mixed expressions rather than select from a small set of discrete categories. Multi-task learning architectures often use valence and arousal regression alongside categorical prediction, benefiting from the more nuanced signal provided by this continuous emotional space [18].

2.2.3 Facial Action Coding System (FACS)

The Facial Action Coding System (FACS) provides a physiological perspective on emotion by decomposing expressions into Action Units (AUs), each corresponding to a specific facial muscle movement [15, 17]. Because combinations of AUs can describe a wide spectrum of expressions, FACS offers a more detailed representation than categorical labels and enables increased interpretability. Datasets annotated with AUs are valuable for training FER systems that operate in sensitive settings where transparency or fine grained behavioral interpretation is required, both of which are important in HRI.

2.3 Classical Approaches

2.3.1 Hand-Crafted Feature Extractors

Before deep learning systems became widespread, FER pipelines relied on explicitly engineered image descriptors. Approaches using Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), Gabor filter banks and geometric relationships between facial landmarks were used in early FER literature [23]. Although these methods offered reasonable performance on controlled datasets, they often failed under real-world conditions. Their sensitivity to variation in lighting, occlusions and head pose made them unsuitable for use in dynamic environments.

2.3.2 Machine Learning Classifiers

Hand-crafted features were typically paired with machine learning classifiers that attempted to map these descriptors onto emotion categories. Popular approaches included support vector machines, random forest ensembles, k-nearest neighbor classifiers, and when working with video-based FER, temporal models such as Hidden Markov Models. While these classifiers helped establish early baselines in FER research, they were limited by the representational power of the handcrafted features. As a result, they struggled to model the complex, high-dimensional variations of facial expressions encountered in real world conditions.

2.4 Deep Learning

2.4.1 Convolutional Neural Networks (CNNs)

The emergence of CNNs revolutionized FER by making it possible to learn hierarchical features directly from data. Architectures such as VGG [24], ResNet [8] and Inception [25] demonstrated that deep models could extract both low-level and high-level features necessary for emotion recognition. Pretraining on large face recognition datasets such as VGGFace2 [2] further enhanced the transferability of learned features to FER.

VGGFace2 is particularly valuable because it contains large identity diversity, pose variability, and age differences that enables models to generalize better across subjects. ResNet-based architectures introduced residual connections, which allowed networks to grow deeper without suffering from vanishing gradients. This stability made them great candidates for FER backbones. Inception architectures contributed by incorporating multiscale convolutions, which lead to better performance on datasets with varying resolutions and facial proportions.

2.4.2 Receptive Fields and Hierarchical Abstraction

A crucial concept in CNNs is the receptive field, which defines the spatial extent of the input that influences a neuron's activation. As images pass through successive convolutional layers, receptive fields expand, enabling deeper layers to capture long-range relationships.

However, increasing depth introduces computational challenges. Deep architectures require more parameters and are susceptible to vanishing gradients during training.

Architectural innovations, such as residual connections in ResNet, address these limitations by enabling efficient propagation of gradients through identity mappings [8]. Residual design is used extensively in FER models because it supports deeper backbones without sacrificing trainability.

2.4.3 Pooling and Spatial Downsampling

Pooling layers reduce the spatial resolution of feature maps, summarizing local regions with operations such as max pooling or average pooling. This downsampling introduces robustness to small translations and deformations. Modest pooling is beneficial because expressions involve small but semantically meaningful movements that should remain detectable even when the face undergoes slight positional shifts.

Nevertheless, excessive pooling can lead to loss of fine-grained information, which is detrimental for distinguishing subtle emotional cues. As a result, modern FER architectures often use reduced pooling, dilated convolutions, or strided convolutions to preserve more spatial detail. Transformer-based FER models go further by eliminating pooling altogether and relying instead on patch embeddings.

2.4.4 Normalization and Activation Functions

Techniques such as batch normalization stabilize training by normalizing intermediate activations across mini-batches. This reduces internal covariate shift and allows the use of higher learning rates, improving convergence speed. Normalization is especially important due to the variability of illumination and facial appearance across datasets. It helps maintain stable feature distributions even when input conditions vary significantly.

Activation functions introduce nonlinearity into the network, enabling the representation of complex patterns. The rectified linear unit (ReLU) remains the most widely used activation in CNN-based FER models because of its simplicity and empirical performance. More recent alternatives, such as LeakyReLU and GELU, appear in deeper backbones and transformer-based systems to improve gradient flow and representational smoothness.

2.4.5 Feature Maps and Channel Semantics

A CNN processes facial images by constructing multiple feature maps, each corresponding to a filter that responds to specific facial characteristics. Channel-wise interpretability becomes increasingly important when deploying FER in human-robot interaction because developers must understand which cues the model prioritizes. For example, if an FER network responds disproportionately to background texture rather than meaningful facial regions, explainability tools such as Grad-CAM can reveal these dependencies and guide architectural adjustments.

Different channels often encode complementary emotional signals. Some channels may specialize in detecting eye-region intensity changes, while others track lip curvature or jaw tension. This diversity allows FER models to represent compound and subtle expressions, which are common in natural human interaction.

2.4.6 Residual and Dense Connectivity

As FER datasets grew and models increased in depth, architectural innovations emerged to improve information flow. ResNet introduced skip connections to mitigate training degradation, allowing networks with hundreds of layers to be trained effectively [8]. DenseNet further extended this idea by connecting each layer to all subsequent layers, strengthening feature reuse and reducing the number of parameters. Both concepts are relevant to FER because deeper models often capture more nuanced emotional cues but require stable gradient propagation.

Residual and dense connectivity also improve robustness to occlusions by allowing the model to aggregate information from multiple scales and feature depths. When parts of the face are occluded, earlier layers may still provide useful cues that propagate through skip connections.

2.4.7 Attention Within Convolutional Networks

Even before the rise of fully transformer-based vision models, researchers integrated attention mechanisms directly into CNNs to enhance their ability to focus on emotionally relevant features. Channel attention modules, spatial attention blocks, and hybrid attention units improved FER performance by emphasizing informative facial regions while suppressing irrelevant background or identity-specific cues. These mechanisms are particularly beneficial in robotic environments, where lighting changes, camera movement, and user pose variation can introduce distracting information.

Architectures such as the Dual-Direction Attention Mixed Feature Network (DDAMFN) [28] illustrate the effectiveness of combining convolutional features with specialized attention pathways. By modulating the flow of information through channel-wise and spatial attention, such networks learn more robust and discriminative representations tailored to emotion recognition.

2.4.8 Specialized FER Architectures

As FER advanced, several models were introduces with intention to address specific challenges. Some architectures attempted to mitigate identity bias by separating features related to identity from those related to expression, while others focused on leveraging attention mechanisms to isolate emotionally relevant regions of the face. For example, systems that integrate multilevel supervision [12], architectures such as DAN that guide attention toward informative facial areas [27], networks such as DDAMFN that combine dual direction attention with mixed kernel convolutions to capture both local and long range dependencies [28]. Approaches using graph convolutional networks incorporate facial landmark structures into the learning process, which provides another way to represent geometric relationships that underline expressions [11].

2.4.9 Comparison of FER

Table 2.1 summarizes several FER architectures and shows how the field has moved toward attention based and transformer based designs. The performance gap between classical CNNs and state-of-the-art transformer models shows that modeling global relationships between facial regions leads to high accuracy on large datasets.

Table 2.1: Performance of deep FER architectures on AffectNet.

Model	**Backbone**	**Parameters**	**AffectNet Accuracy**
VGGFace2 fine-tuned	VGG-16	138M	55–60% [2]
ResNet-50	ResNet-50	25M	58–63% [8]
DAN	ResNet-50	26M	63–66% [27]
POSTER++	Transformer	40M	67–70% [30]
Emotion-GCN	GCN + CNN	5M	60–65% [11]

2.5 Attention Mechanisms and Transformer-Based FER

2.5.1 Attention in FER

Facial expressions can be defined by subtle and localized cues like eyebrow tension, eye aperture or small variations in lip curvature. Traditional convolutional neural networks require increasingly deep architectures to capture interactions between distant facial regions, since their receptive fields expand gradually with depth. Attention mechanisms address this by enabling models to selectively emphasize features that are relevant for emotion recognition while suppressing those that are not. This allows FER systems to focus on critical regions such as the eyes during expressions of fear or surprise, or on the mouth when identifying happiness, rather than the entire face. These mechanisms also improve robustness to factors like background noise or illumination variations.

2.5.2 Self-Attention and Vision Transformers

The introduction of Vision Transformers (ViT) marked a major shift in computer vision by replacing convolutions with self-attention operations applied to image patches. In FER, this approach has become increasingly influential because self-attention can model long-range dependencies across the face more effectively than convolutions alone. Transformers evaluate pairwise relationships between all spatial positions simultaneously, offering a global perspective on the facial configuration that is particularly beneficial in expressions where interactions between distant features carry emotional meaning. Capacity for global reasoning allows transformer-based FER systems to better handle non-frontal faces and varied lighting. Thanks to this, transformer-based models such as POSTER and POSTER++ have set new performance benchmarks for FER [29, 30].

2.5.3 Hybrid Transformer-CNN Models

Pure transformer models often require large datasets to train effectively. To address this, hybrid architectures combine the capabilities of both CNNs and transformers. In these systems, convolutional layers typically extract low level edges and textures, while transformer layers reason about global relationships between high level facial features. Cross-attention mechanisms allow the model to integrate global contextual information with localized details,this leads to more stable representations of facial expressions, it is especially effective when FER systems must generalize across datasets.

2.5.4 Regional Attention Networks

In addition to global attention techniques, researchers have explored architectures that emphasize emotion relevant regions through guided attention. DAN, for example, constructs multiple attention maps focusing on different parts of the face and integrates them into a unified representation [27]. By systematically highlighting or suppressing particular regions, such networks achieve better discrimination among similar emotions. DDAMFN, which combines dual direction attention with mixed kernel convolutions, further improves the modeling of complementary local and global features [28].

2.5.5 Graph-Based Attention Models

Graph based architectures represent another important branch of FER. Instead of viewing the face as a pixel grid, these models see it as a structured graph composed of landmarks with spatial relationships. Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs) treat facial expression recognition as a problem of learning patterns in these landmark relationships. Emotion-GCN combines CNN-extracted features with graph reasoning to focus on deformable facial geometry [11]. This approach is great when dealing with occlusions or pose variations that obscure entire facial regions but leave the geometric configuration of visible landmarks intact.

2.6 Explainable AI in Facial Emotion Recognition

Explainability has become an essential for modern emotion recognition systems. Since these models are deployed in settings that involve direct interaction with humans, the importance of interpretability is both in technical transparency and safety. Developers must understand not only the outputs of an FER system but also the internal rationale behind its predictions. The survey *Explainable AI: A Review of Machine Learning Interpretability Methods* [14] provides a systematic overview of the interpretability landscape and identifies key distinctions such as intrinsic versus post-hoc interpretability and global versus local explanations, distinctions that are directly relevant for the design of reliable FER systems in human-robot interaction.

2.6.1 Categories of Interpretability Methods

Interpretability methods can be categorized along several dimensions. Intrinsic methods are those that are inherently interpretable by design, relying on transparent model structures such as decision trees or sparse linear models. These approaches provide global interpretability because the entire decision-making process can be inspected directly. However, these models generally lack the representational power needed for FER, where high-dimensional image data and subtle facial cues demand more expressive architectures.

In contrast, post-hoc interpretability methods operate on trained models and generate explanations without altering the underlying architecture. These approaches are far more applicable to FER, where deep neural networks dominate state-of-the-art performance. According to the review [14], post-hoc methods can be further divided into global and local explanations. Global explanations attempt to describe the overall behavior of a model, whereas local explanations focus on individual predictions. FER systems typically benefit from the latter because the emotional content of a single image or sequence may require targeted diagnostic analysis to ensure correct robot behavior in real time.

Table 2.2: Overview of interpretability method categories relevant to FER and HRI.

Category	**Scope**	**Relevance to FER**
Intrinsic models	Global	Limited, as FER requires high-capacity models
Post-hoc global	Global	Useful for dataset auditing and bias inspection
Post-hoc local	Local	Essential for inspecting individual FER decisions
Model-agnostic	Local/Global	Flexible; applicable to any FER architecture
Model-specific	Local/Global	Provides deeper insights into CNN/Transformer features

2.6.2 Visual Attribution for FER in Robotics

Among post-hoc interpretability approaches, visual attribution has become one of the most widely used tools for assessing the behavior of FER models. Methods such as Grad-CAM [21] generate heatmaps that highlight the regions of the input image most responsible for a model's decision. When applied to FER in robotics, these heatmaps help determine whether a robot relies on socially meaningful cues such as the tension around the eyes during expressions of fear or whether it is influenced by irrelevant contextual features such as background patterns, clothing, or hair. These methods can also serve as mechanisms for identifying dataset biases or model oversights.

Layer-Wise Relevance Propagation [1] provides a more fine-grained interpretability method by decomposing a model's output into pixel-level relevance scores. These explanations can reveal how a network interprets minute visual cues such as subtle eyebrow contractions or faint lip depressions.

2.6.3 Interpreting Feature Embeddings

The structure of the feature space itself can offer valuable insights into the behavior of FER models. Dimensionality reduction techniques such as t-SNE or UMAP project high dimensional feature embeddings into interpretable two dimensional spaces, revealing how the model clusters different emotional categories. The review [14] notes that such global interpretability tools are useful for identifying overlapping categories, dataset imbalance effects or even systematic misclassifications.

2.6.4 Explainability as a Requirement in HRI

Explainability plays a unique functional role in human-robot interaction. In realtime deployments, robots continuously interpret facial expressions and use these interpretations to modulate speech, gesture, or gaze behavior. When a robot misinterprets an emotion, users are likely to notice immediately. Explainability tools allow developers not only to understand why a misclassification occurred but also to evaluate whether the underlying cause poses a safety or ethical risk. For example, if a robot systematically misclassifies expressions from certain demographic groups, explainability methods can reveal whether the model has internalized unintended biases. The review [14] highlights the importance of such techniques for ethical auditing, fairness assessment, and transparent decision-making.

Explainability also supports the integration of FER with other subsystems in the robot. For example, if an FER module triggers a behavioral controller to adopt a comforting posture in response to sadness, developers must verify that this action is grounded in reliable perceptual cues.

2.7 Dataset Challenges in FER

2.7.1 Imbalanced Class Distributions

Most FER datasets exhibit strong class imbalance, with expressions such as happiness and neutrality appearing far more frequently than emotions like disgust or fear. AffectNet is one of them, as it contains hundreds of thousands of happy or neutral faces but far fewer samples of rare emotions [18]. This imbalance skews model predictions toward majority classes and diminishes performance on minority classes. Mitigation strategies include adjusting loss functions, oversampling minority classes, balancing through data augmentation and generating synthetic samples.

Table 2.3: Distribution of categorical labels in the AffectNet dataset [18].

Emotion	**Number of samples**
Neutral	75374
Happy	134915
Sad	25959
Surprise	14590
Fear	6878
Disgust	4303
Anger	25380
Contempt	4250

2.7.2 Annotation Noise and Label Subjectivity

Because emotion interpretation is subjective, large FER datasets often contain mislabeled or ambiguous samples. Crowdsourced annotation protocols, like those used for RAF-DB [12], may introduce inconsistent labeling due to individual annotator biases and uncertainty in interpreting subtle expressions. Deep learning models trained on noisy labels may inherit these inconsistencies, leading to skewed predictions. Robust loss functions and noise modeling techniques have been proposed to cope with such challenges, careful curation and dataset refinement are also very effective.

2.7.3 Domain Shift and Real-World Deployment

A major challenge in FER research arises from the discrepancy between training datasets and real world deployment environments. Many FER datasets consist of images scraped from the web, whereas robots such as NICO operate with onboard cameras that produce images of different resolution, lighting and distance. This gap called domain shift can reduce performance when a model trained on standard datasets is deployed. Domain adaptation approaches [19], fine-tuning on robot-captured data, and synthetic augmentation are important steps in building effective robotic FER systems.

2.8 FER Datasets

AffectNet stands as one of the most comprehensive datasets available for FER, containing almost half million annotated images distributed across eight primary emotional categories [18]. Its scale and annotation depth make it particularly valuable for training high capacity deep learning models, but its strong class imbalance creates challenges that must be addressed in order to achieve good results. RAF-DB offers crowdsourced annotations of both basic and compound emotions [12], making it a valuable resource for studying the complexity of blended affective states. FER-2013, although substantially smaller and consisting of low-resolution grayscale images, remains widely used as a baseline dataset due to its simplicity and the availability of standardized benchmark protocols [7].

Datasets such as BP4D [15] and DISFA [17] provide detailed FACS-based annotations that allow models to learn fine grained Action Unit combinations. These datasets are sometimes used for research directions that require interpretability or physiologically grounded representations. VGGFace2, though not an FER dataset, is widely used for pretraining because its wide pose variation and identity variety help build models that generalize well to FER tasks [2].

2.9 Synthetic Data Generation and Advanced Augmentation Methods

2.9.1 Synthetic Data in FER

The performance of FER systems depends heavily on the diversity and balance of their training datasets. Real world facial expression data, particularly for rare emotions such as disgust, fear, or contempt, is difficult to collect at scale. Even large datasets like AffectNet remain highly imbalanced and contain inconsistent labels, which complicates model training and can lead to biased predictions [18]. Synthetic data generation therefore plays a crucial role in addressing these limitations. By creating balanced datasets or improving the balance of existing ones, we can mitigate class imbalance, increase diversity in lighting and pose variation, and improve the robustness of FER models to real world scenarios. Synthetic data further enables robot specific calibration by allowing the generation of faces captured under the same camera geometry and lighting conditions as in real conditions.

2.9.2 GAN-Based Facial Expression Synthesis

Generative Adversarial Networks (GANs) have introduced a powerful paradigm for producing high quality synthetic facial images. Models such as StyleGAN [10] can generate photorealistic faces with a high degree of control over attributes. StarGAN [3] extends this functionality by enabling multi domain translation, allowing a single generator to convert images from one emotional category to another without requiring multiple models. This capability is particularly advantageous for FER, since it enables the creation of balanced datasets by producing additional samples for underrepresented emotion classes.

GANimation [20] takes a more physiologically grounded approach by manipulating Action Units (AUs). Because AUs correspond directly to facial muscle activations, GANimation can generate realistic emotional expressions with varying degrees of intensity. This degree of control allows the creation of subtle, compound, or ambiguous expressions—types that are rarely well represented in natural datasets but commonly appear in real humans.

2.9.3 Augmentation Techniques

Traditional forms of data augmentation such as rotation, scaling, color jitter, noise injection, or random cropping help improve FER robustness. These transformations allow models to tolerate variations in lighting conditions, camera distance, or small misalignments, all of which are common during human robot interaction. While such basic augmentations improve generalization, they often do not provide sufficient variation to compensate for severe dataset imbalance or challenging environmental conditions.

To enhance robustness further, modern augmentation strategies increasingly rely on adversarial perturbations or style-transfer methods. Adversarial augmentation, for instance, forces models to remain stable even when inputs contain deliberately challenging distortions. Style-transfer techniques can simulate variations in lighting, texture, or camera domain, making them useful for adapting models to the visual characteristics of robotic sensors.

2.10 3D Simulation and Synthetic Rendering

2.10.1 3D Synthetic Humans for FER

Three dimensional simulation environments, like those built using engines such as Unreal Engine or Unity, offer means of generating controlled and highly customizable datasets for FER research. Unlike purely generative models, 3D simulations allow precise manipulation of facial expressions, head pose, lighting and background. These environments can be used to create synthetic datasets in which each emotion category is represented equally, counteracting the imbalance. They also allow us to simulate images using the exact camera geometry that a robot employs, which enables domain-specific calibration and decreases the gap between training and deployment conditions [16].

2.10.2 Blendshape-Based Expression Modeling

Most 3D engines provide sophisticated facial rigs that rely on blendshapes, skeletal animation, or procedural deformation. Blendshape systems are particularly relevant for FER because they correspond closely to FACS Action Units. By controlling parameters such as inner brow raising or eyelid tightening, it becomes possible to generate expressions that match recognized emotional categories with a high degree of realism. Because blendshapes allow continuous adjustment of AU intensity, they provide a mechanism for producing subtle emotional expressions that mirror the complexity of real human affect.

2.10.3 Domain Randomization

Domain randomization introduces intentional variability in simulated environments by altering lighting, background textures, camera angles, or facial appearance in extreme and random ways. The goal is to train models that are not overly sensitive to any specific environmental condition. By exposing a network to a wide range of synthetic variations, we encourage it to learn domain invariant features that transfer more effectively to real world settings. For robots like NICO, which frequently operate under inconsistent lighting or with users who move unpredictably, domain randomization provides a practical way to enhance robustness and reduce the need for extensive real world data collection.

2.10.4 Synthetic-to-Real Transfer

Despite the advantages of synthetic data, a performance drop typically occurs when models trained primarily on synthetic images are deployed on real world images. This phenomenon, known as the “sim2real gap,” arises from differences in texture realism. Domain adaptation methods address this challenge by aligning the feature distributions between synthetic and real datasets. Approaches vary from adversarial domain adaptation and Maximum Mean Discrepancy minimization to style transfer techniques that transform synthetic images to resemble real ones [19]. Fine tuning on small quantities of real robot-captured data further reduces this gap.

2.11 Domain Adaptation and Generalization

2.11.1 Supervised and Unsupervised Adaptation

Domain adaptation can occur under supervised or unsupervised conditions. In supervised adaptation, a small amount of labeled data from the target domain, such as images acquired directly from the robot’s camera, is used to refine the model. In unsupervised adaptation, no labels are available in the target domain, algorithms align distributions using statistical or adversarial constraints instead. FER research increasingly relies on unsupervised methods because acquiring emotion labels for robot specific images is labor intensive and unreliable.

2.11.2 Cross-Dataset Generalization

Cross-dataset generalization remains a significant challenge in FER because datasets vary widely in style, demographic composition, annotation quality and environmental conditions. A model trained on AffectNet may not perform well on RAF-DB or FER-2013, and the other way around as well. Robots that operate in varied environments must deal with this generalization problem on an even broader scale. Approaches integrating diverse datasets, transformer architectures and domain adaptation techniques tend to generalize more effectively across contexts. These trends indicate that FER research is shifting toward models that explicitly prioritize robustness rather than optimizing for accuracy on a singular dataset.

2.12 Integration of FER into Human-Robot Interaction Systems

The incorporation of FER into human-robot interaction typically relies on a multi stage processing pipeline that is made up of face detection face alignment, feature extraction and emotion classification. Although this pipeline may appear straightforward, the demands of realtime interaction and the unpredictability of human behavior place significant constraints on each component. Facial regions must be located accurately even as subjects move, turn their heads or enter and exit the robot's field of view. The alignment process must standardize orientation and scale with minimal computational overhead, and the feature extractor must produce stable embeddings under changing conditions. The classifier must map these embeddings to discrete or continuous emotional states consistently enough for the robot to behave in a meaningful way. In the absence of such consistency, the downstream behavior module may trigger inappropriate actions that disrupt interaction and human experience.

Face detection and tracking constitute important stages of the pipeline. FER accuracy declines drastically when the input region is misaligned or incomplete, and the robot must therefore maintain a stable view of the user's face even under motion. Modern detectors such as MTCNN, RetinaFace and some mobile variants of YOLO have become common choices in robotic systems because they provide a balance of accuracy and speed suitable for realtime deployment. When integrated with lightweight tracking methods, these detectors help stabilize temporal fluctuations in the detected face region, improving the reliability of subsequent expression classification.

Many FER models operate on individual images, but robots function in real time can benefit greatly from temporal modeling. Recurrent neural networks, temporal convolutional architectures and transformer-based sequence models all offer mechanisms for capturing the temporal evolution of expressions. Such temporal smoothing reduces frame-by-frame noise and yields more stable predictions, allowing robots to respond to users more naturally. Even simple filters like moving-average can reduce fluctuations in predicted emotion labels, helping humanoid robots avoid abrupt behavioral shifts.

2.13 HRI Challenges in FER

Robots must interpret expressions that are not posed for a camera but produced spontaneously as a part of an interaction. Spontaneous expressions tend to be more subtle and more contextually dependent than acted ones. These differences introduce complexity that many FER datasets fail to capture, since they frequently contain exaggerated or stylized expressions from web images or controlled laboratory recordings. A robot relying on a model trained exclusively on such datasets may misclassify subtle emotions or fail to detect them altogether. This is why approaches that incorporate synthetic augmentation, unsupervised adaptation and realworld calibration are useful for correctly classifying authentic human behavior.

Cultural and demographic variability further complicates FER. Emotional expressions differ subtly across cultures, and datasets may under or over represent particular demographic groups, leading to model biases. These biases can manifest as systematic misclassification of certain populations, raising ethical concerns when deploying FER in socially interactive robots. Ensuring that FER systems behave fairly requires attention to dataset composition, augmentation strategies that diversify appearance, and explainability tools that reveal unwanted dependencies on identity-related cues.

Realistic deployments also expose robots to occlusions arising from glasses, masks, hats, or hair, all of which interfere with the visibility of key facial regions. Traditional CNNs often degrade significantly under such conditions. Given the increased prevalence of face coverings in recent times, robustness to occlusions may hold higher significance than before.

Camera distance and viewpoint variation introduce additional challenges, as robots frequently interact with users who move around a physical space. Interactions may lead to vastly different perspectives and thus absence of certain facial features. These variations degrade the performance of FER models trained predominantly on frontal images. Synthetic augmentation and domain randomization provide partial solutions by generating data that mimic these conditions, still it remains a concern when it comes to FER.

Realtime responsiveness is another dimension of HRI-specific difficulty. Robots must classify expressions quickly in order to react in expected time frames. High latency can cause delays in emotional responses, making the robot appear inattentive or disengaged. This restricts the feasible size and complexity of FER models. Techniques such as model pruning and quantization are therefore essential for embedding FER within computationally constrained robotic platforms. Alternatively, the use of lightweight backbones can also improve response times, but at a cost of performance.

2.14 Humanoid Robot NICO

The NICO humanoid robot is designed as a research platform for cognitive development and social interactions [6, 22]. Its sensing capabilities include stereo cameras and audio input, and computation is expected to be performed on a laptop connected to the robot. While this setup provides greater flexibility than embedded microcontrollers, it still imposes practical limits on model size and inference speed. FER system designed for NICO must therefore balance accuracy with efficiency, ensuring that emotion classification occurs smoothly.

Previous work with NICO has explored domains such as hand gesture recognition, gaze estimation and imitation learning. Emotion recognition, however, introduces distinct challenges because it demands both perceptual sensitivity and reasoning. Facial expressions captured by NICO's cameras are subject to the robot's specific field-of-view constraints, its head movement patterns, and the dynamic behavior of human users. These conditions amplify the need for domain adaptation and calibration.

The constraints inherent to NICO's hardware make explainability particularly interesting. If the robot misinterprets an expression and reacts inappropriately, users must be able to understand why the misclassification occurred. Saliency-based tools such as Grad-CAM [21] or relevance propagation [1] enable diagnosis of these errors and adjust the model or data pipeline accordingly.

Chapter 3

Thesis Objectives

Chapter 4

Methodology

Chapter 5

System Design

Chapter 6

Implementation

Chapter 7

Experiments and Results

In this section, we evaluate and compare techniques for handling an imbalanced dataset using the validation split of the AffectNet dataset. We first analyze the performance of the baseline model under different learning rate settings and subsequently examine methods such as oversampling, class weighting, and data augmentation. Each technique is assessed based on its impact on the average F1 score.

7.1 Baseline Model

We began by creating a baseline model without applying any techniques specifically designed for imbalanced data. The objective of this experiment was to establish a reference point for subsequent evaluations and to compare the results of our baseline model with those reported in the original project.

Our model was trained on the same training and validation splits as the original implementation. We introduced several modifications to key hyperparameters, most notably increasing the batch size from 16 to 64 in order to improve training efficiency and stability. We also observed that the learning rate used in the original project was likely too high, which may have hindered model convergence. To address this, we experimented with three different learning rates to identify the value that would yield optimal performance. We hypothesized that these adjustments would lead to improved results relative to the original model's mean F1 score of 52.57

During fine-tuning, we found experimentally that a learning rate of 0.000005 provided the best performance, as shown in Table 7.1. Models trained with this learning rate achieved a significantly improved average F1 score of 53.51

Learning rate	Loss	Average F1 score
0.0005	0.6789	0.308
0.00005	0.6164	0.460
0.000005	0.5720	0.529

Table 7.1: Results of different learning parameters

Chapter 8

Discussion

Conclusion

Chapter 9

User Manual

Bibliography

[1] Sebastian Bach et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140, 2015.

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[3] Yunjey Choi et al. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797, 2018.

[4] Stefano d'Apolito, Danda Pani Paudel, Zhiwu Huang, Andres Romero, and Luc Van Gool. Ganmut: Learning interpretable conditional space for gamut of emotions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 568–577, 2021.

[5] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98(45):16–22, 1999.

[6] Elisabeth andothers Friedrich. Nico: Neuro-inspired companion for human-robot interaction. In *HRI*, 2019.

[7] Ian Goodfellow, Dumitru Erhan, Pierre Carrier, et al. Challenges in representation learning: A report on facial expression recognition challenge. In *ICML Workshops*, 2013.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[9] Martin Heusel et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, pages 6626–6637, 2017.

[10] Tero Karras et al. A style-based generator architecture for generative adversarial networks. *CVPR*, pages 4401–4410, 2019.

[11] Jiabei Li et al. Facial expression recognition with graph convolutional networks. In *ACM MM*, pages 1438–1447, 2021.

[12] Shan Li and Weihong Deng. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, pages 2852–2861, 2017.

[13] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215, 2020.

[14] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[15] Patrick Lucey et al. The extended cohn–kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. *CVPR Workshops*, pages 94–101, 2010.

[16] Meysam Madadi et al. Synthesizing training data for facial expression recognition using unreal engine. *IEEE Transactions on Games*, 2019.

[17] Mohammad Mavadati et al. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[18] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[19] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[20] Albert Pumarola et al. Ganimation: Anatomically-aware face animation from a single image. In *ECCV*, pages 818–833, 2018.

[21] Ramprasaath R. Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[22] Rohan Sharma et al. The nico humanoid robot platform for research in cognitive interaction. *International Journal of Social Robotics*, 2020.

[23] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[25] Christian Szegedy et al. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[26] Kai Wang, Xiangyu Peng, Jian Yang, and Yu Qiao. Region attention networks for pose- and occlusion-robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020.

[27] Yanan Wen et al. Distract your attention: Multi-head cross attention network for facial expression recognition. In *ICCV*, pages 153–162, 2019.

[28] Jie Zeng et al. Deep dual-branch attention mixed feature networks for facial expression recognition. *IEEE Access*, 8:153509–153522, 2020.

[29] Jun Zhao et al. Poster: A pyramid cross-fusion transformer for facial expression recognition. In *CVPR*, pages 2407–2416, 2021.

[30] Jun Zhao et al. Poster++: A transformer-based facial expression recognition framework. *IEEE Transactions on Affective Computing*, 2022.

[31] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.