Bachelor Thesis

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Cybernetics

Emotion Detection from Video Using Machine Learning

Jan Sadílek

Supervisor: doc. Ing. Daniel Novák, Ph.D.
Study Program: Open Informatics
Specialization: Artificial Inteligence and Computer Science
May 2025

Acknowledgements

I would like to thank doc. Ing. Daniel Novák, Ph.D. for his support, guidance and patience. I would also like to thank Fabián Bodnár for his his tips and practical feedback. Special thanks also go to ChatGPT for invaluable assistance, unwavering support, brainstorming ideas, refining grammar, and countless other improvements. Lastly, I would like to thank my friends and family for their support during the whole process.

Abstract

This thesis addresses facial expression-based emotion detection from video, particularly in the context of online video calls, to enhance human-computer interaction and therapeutic insights. Motivated by a practical application in a platform for transcription and summarization of healthcare conversations primarily, the work developed a lightweight emotion recognition system capable of analyzing longer video sequences of individuals directly facing the camera. The emotions classified are happiness, sadness, anger, fear, disgust, surprise, contempt, and neutrality.

The approach involved fine-tuning the MobileNetV2 convolutional neural network (CNN) on the FER2013Plus dataset, achieving accuracy levels of 73% on FER2013Plus, 60% on CK+, and 47% on RAVDESS datasets. A React-based user interface and Flask backend facilitated video input and analysis, demonstrating the practical applicability of the developed model.

The work noted limitations such as reliance on visual modality only, limited generalization, prediction on a single-frame, and class imbalance.

Keywords: emotion detection, facial expressions, video analysis, MobileNetV2, FER2013Plus, affective computing, emotion classification

Supervisor: doc. Ing. Daniel Novák, Ph.D.

Abstrakt

Tato práce se zabývá detekcí emocí z videa na základě výrazu obličeje, zejména v kontextu online videohovorů, s cílem zlepšit interakci mezi člověkem a počítačem a terapeutické poznatky. Práce je motivována praktickou využitelností v platformě pro přepis a sumarizaci rozhovorů zejména ve zdravotnictví. V rámci práce byl vytvořen nenáročný systém pro rozpoznávání emocí schopný analyzovat delší video sekvence. Klasifikované emoce jsou štěstí, smutek, hněv, strach, znechucení, překvapení, pohrdání a neutralita.

Tento přístup zahrnoval fine-tuning konvoluční neuronové sítě (CNN) MobileNetV2 na datové sadě FER2013Plus, přičemž bylo dosaženo přesnosti 73 % na datové sadě FER2013Plus, 60 % na datové sadě CK+ a 47 % na datové sadě RAVDESS. Uživatelské rozhraní v Reactu a Flask backend server zajistili input a analýzu videa a prokázaly praktickou použitelnost vyvinutého modelu.

Práce popisuje několik omezení tohoto přístupu, zejména: využívání pouze vizuální modality, omezené zobecnění, predikce na základě jednoho snímku a nevyváženost tříd.

Klíčová slova: detekce emocí, výrazy obličeje, analýza videa, MobileNetV2, FER2013Plus, afektivní výpočty, klasifikace emocí

Překlad názvu: Detekce emocí z videa pomocí strojového učení

Contents

Figures

Tables

Chapter 1

Introduction

1.1 Problem Definition

Emotion detection seeks to classify human emotions into discrete categories using computational methods. These methods commonly analyze facial expressions, body language, vocal intonation, physiological signals, and verbal content such as speech transcripts and semantics.

In the context of video, emotion detection involves analyzing sequences of frames while primarily focusing on facial expressions to identify emotional states. Unlike single image analysis, video based detection can capture temporal dynamics, revealing subtle changes in expression over time. This task falls within affective computing, aiming to enable machines to interpret human emotions.

1.2 Motivation

Recognizing emotions from video is essential as it enables more natural human-computer interaction. It can make AI assistants or other robots more empathetic. Facial expressions are our primary source of non-verbal communication, conveying crucial information not present in verbal expression.

Additionally, a direct motivation for our work comes from a practical application in an AI-powered service designed to assist healthcare professionals by transcribing and summarizing online meetings and video calls. One of the main objectives of such a platform is to reduce the volume of paperwork and to help medical professionals focus on patient care. The implementation of emotion detection from video calls could provide a more practical summarization of meeting content, especially if used during therapeutic sessions where such information can be crucial.

1.3 Scope of work

The focus of our work is on facial expression based emotion detection in video. We will focus on the visual modality only any audio or physiological signs will not be taken into account. Emotions will be categorized into 8 categories: the seven basic human emotions happiness, sadness, anger, fear, disgust, surprise, contempt, and neutral. Continuous emotion levels like valence and arousal will not be handled.

Due to our usecase, the primary focus is on videos taken during an online video call, a human subject looking directly at the camera. Another important consideration is for the system to be able to process longer input as the video call session can be fairly long.

1.4 Objectives

Our work has multiple objectives. The major one is to develop a working emotion detection machine learning model capable of classifying human emotions from video frames. Specifically, from videos obtained from a camera during an online video call. The second one is to implement a simple user interface capable of ingesting video input, analyzing emotions, and subsequent visualization of results. The third objective is to evaluate the model's accuracy and performance.

The remainder of this thesis is structured as follows: Chapter 2 gives a brief introduction of emotions. Chapter 3 reviews related works in the field of emotion detection, particularly from video. Chapter 4 describes the methodology of our work. Chapter 5 follows the implementation of our work. Chapter 6 evaluates the model used, and Chapter 7 summarizes the results.

Chapter 2

Brief introduction to Emotions

Emotions are physical and mental states brought on by neuro-physiological changes, variously associated with thoughts, feelings, behavioral responses, and a degree of pleasure or displeasure[Wik25] Their physical manifestations are an outward-facing reflection of our inner thoughts.

Systematic scientific interest in emotions dates back to Charles Darwin's book The Expression of the Emotion in Man and Animals (1872). Darwin's work laid the groundwork for later theories which postulate that there are some biologically universally recognizable basic emotions. [Dar72]

There are two main ways to describe emotions. The first one is using categorical (discrete) models which assume a small set of biologically hard-wired emotions that are recognizable by characteristic facial patterns. Paul Ekman proposes six universal emotions anger, disgust, fear, happiness, sadness, and surprise, which he later extended with contempt. [Ekm92] Meanwhile, Robert Plutchik and Henry Kellerman propose a more complex model (Figure 2.1). Eight emotions (joy, trust, fear, surprise, sadness, disgust, anger, anticipation) are arranged as opposing pairs on a wheel, neighboring emotions then blend to form more complex states such as love (joy with trust) or remorse (sadness with disgust). [PK13]

In contrast, dimensional approach treats every emotion as a point in a continuous space, mostly defined by valence (pleasant unpleasant) and arousal (calm activated). James A. Russell proposes a circumplex model (Figure 2.2) where basic emotions occupy characteristic sectors of this plane. Happiness is high-valence with high-arousal while sadness is low-valence with low-arousal. This model allows for nuanced differences and mixed feelings to be represented. [Rus80]

Figure 2.1: Plutchik's emotion wheel [Com25b]

While these are the most established theories, there are many ways that psychologists categorize basic emotions, and one cannot say that there is a consensus on what the correct way to do so is. We highlight only the ones that are relevant to our work.

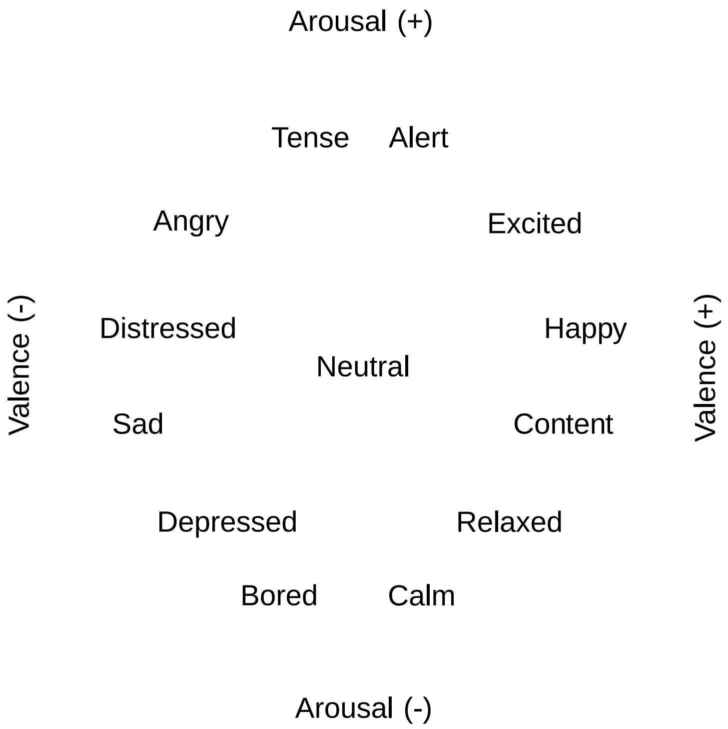

Figure 2.2: Russell's Circumplex model [Com25a]

Chapter 3

Background and Literature Review

Emotion recognition from video is a task of identifying human emotional states by analyzing visual data over time. Historically, researchers have approached this by examining facial cues frame-by-frame or by modeling temporal dynamics across frame sequences. In general, emotion recognition can leverage multiple modalities — facial expressions, voice tone, physiological signals, and verbal content, such as speech transcription and semantic analysis. This section reviews traditional vision-based methods as well as deep-learning approaches, with a focus on still-image-based emotion detection using convolutional neural networks. We discuss datasets like FER2013 and its updated FER2013Plus version, AffectNet, CK+, RAVDESS, and highlight more complex methods — temporal models, 3D facial data, and multimodal fusion, to contrast their advantages and trade-offs. Finally, we justify the choice of a still-image CNN approach for long video streams in a resource-constrained setting.

3.1 Traditional Approaches to Facial Emotion Recognition

Early work on facial emotion recognition mainly used hand-crafted features and classic machine learning classifiers. A widely used description of facial geometry is the Facial Action Coding System (FACS) proposed by Ekman and Friesen, which breaks an expression into elementary 'action units' (AUs). Each AU corresponds to a specific muscle activation, for instance, AU 4 defines the contraction of two muscles resulting in the lowering of the eyebrows. [DTM14] This representation can be used to train classifiers like Support Vector Machines (SVM), Adaboost, and random forests. The conventional approach is divided into three steps: (1) face and facial component detection, (2) feature extraction, and (3) expression classification. [Chu18]

Traditional approaches have the benefit of low computational cost and their usage was feasible on the hardware of the time while providing reasonable accuracy. For instance, a real-time system by Suk and Prabhakaran (2014) used an Active Shape Model (ASM) to track facial landmarks and SVMs to classify expressions, running at approximately 2.4 frames per second on a smartphone (Samsung Galaxy S3) while achieving an accuracy of 72%. [SP14]

While these results are fairly good, the traditional methods present with some drawbacks and limitations. They rely on features manually engineered by experts, and it is possible that they do not capture the full complexity of facial dynamics. Therefore, they might not allow for the best categorization of human emotions. Also, by separating the feature extraction and classification stages, they do not allow for full optimization of the system. In practice, this can result in lower accuracy in scenarios with varied lighting, backgrounds, and subjects. This is the motivation behind the shift towards data-driven deep learning methods that can learn feature extraction automatically from large datasets. [DPM16] [SK09]

3.2 Deep Learning for Facial Emotion Recognition

With deep learning on the rise, especially Convolutional Neural Networks (CNNs), image-based emotion recognition was on the verge of a breakthrough. CNNs can learn directly from pixel data and have proven to outperform traditional feature-based methods on many vision tasks. [Chu18] Specifically, in facial emotion recognition, early successes of CNNs were brought by the FER-2013 challenge. This contest introduced the Facial Expression Recognition 2013 (FER2013) dataset. It contains 35,887 grayscale face images (48×48 pixels) labeled with seven categories (Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutral). The top entries to the competition all used deep CNN models, significantly outperforming approaches based on hand-crafted features. [GEC+13]

3.2.1 CNN Architecture for FER

Modern CNNs for facial emotion recognition often draw from architectures successful in visual object recognition and are fine-tuned on facial emotion datasets. CNNs typically consist of convolutional layers, this helps to learn local feature detectors like edges, textures, and fully connected layers that map the learned features to emotion classes. [PMR+24] [Chu18] CNNs automatically discover which facial features indicate each emotion by being trained on an expansive data volume, without being taught so explicitly.

Overall, CNN-based methods quickly became state-of-the-art for static image emotion recognition, with accuracy surpassing classical methods by a substantial margin. [MHM19]

3.2.2 MobileNetV2 and Efficient CNNs

While very deep CNNs like ResNet-50 can achieve very high accuracy on the facial emotion recognition task, they come with a high computational cost which can be problematic for real-time or mobile deployment. This has spurred the development of lightweight CNN architectures that strike a better trade-off between accuracy and efficiency. Such network is represented by MobileNetV2 by Sandler et al. (2018), which was designed specifically for resource-constrained environments. MobileNetV2 achieves high accuracy on ImageNet with only ~3.4 million parameters (while ResNet-50 has ~25.6 million parameters). The efficiency of MobileNetV2 aligns with our goal of deploying facial emotion recognition for long video calls which is the reason behind our focus on fine-tuning this architecture in our work. [SHZ+19]

3.3 Datasets for Facial Emotion Recognition

Over the years, a variety of datasets were created. They differ in scale, environment (if they were made in a controlled environment or were taken in-the-wild), modalities, and labels for emotions. We summarize key datasets below, as they provide context for the chosen approach.

Figure 3.1: FER2013 Emotion histogram

3.3.1 FER2013

The Facial Expression Recognition 2013 dataset was created by Pierre Luc Carrier and Aaron Courville. 35 887 images harvested via web search and labeled them with one of the seven basic emotions happiness, sadness, anger, fear, disgust, surprise, contempt, and neutral. FER2013 is considered a challenging in-the-wild set due to deeply varying image quality and labeling noise (the human accuracy is ~65%). This dataset has become a standard benchmark for CNN-based facial emotion recognition. The dataset is not balanced (Figure 3.1) [GEC+13]

3.3.2 FER2013Plus

To address the noisy labels of the original FER2013, Barsoum et al. (Microsoft Research, 2016) re-annotated the dataset via crowdsourcing. The new relabeled dataset was labeled by 10 independent humans. The authors defined 8 emotions, adding contempt to the 7 original. FER+ provides a higher-quality training set. It carries over the strengths of the original dataset while addressing its main issue of inaccurate labels. The resulting FER2013+ provides, for each of the 35 887 images, a probability distribution over the eight emotions rather than a single (hard) label. This enables training with soft targets which better reflects annotation uncertainty and can yield higher accuracy than majority-vote labels. [BZCFZ16] [MLS24]

If we refer to the figure 3.2 displaying the class distribution of the dataset,

Figure 3.2: FER2013Plus Emotion histogram

we see it is still heavily skewed towards some emotions.

3.3.3 AffectNet

AffectNet is one of the largest facial emotion recognition dataset. It contains over 1 million faces collected from the web. Part of the dataset, about 440,000 pictures, were labeled manually with 7 discrete emotions, the same as the original FER2013 dataset. The dataset also features continuous labels for valence and arousal values. It is an in-the-wild dataset with great diversity in subjects, poses, and lighting. Due to its source, the dataset is heavily imbalanced (Figure 3.3). While networks trained on AffectNet offer great in-the-wild performance, training on this dataset is computationally expensive.[MHM19]

3.3.4 CK+

The Cohn-Kanade Extended dataset is a video dataset taken in a lab-controlled environment. The dataset contains 593 video sequences from 123 subjects where each sequence depicts a facial expression evolving from neutral to peak emotion. It is useful for validating algorithms under ideal conditions. The dataset's small size and controlled environment limit its use to testing and validating purposes only, as it is insufficient for training. [LCK+10]

Figure 3.3: AffectNet Emotion histogram

Figure 3.4: CK+ Emotion histogram

There is a variety of adaptations of this dataset to a picture form. They sample multiple frames from each video where most of the frames are neutral (Figure 3.4). This is due to the transition from neutral to peak emotion in each video.

3.3.5 RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song is a multimodal, studio-recorded resource aimed at joint facial and vocal emotion analysis. It contains 7,356 files captured at 720p/30 fps video with 48 kHz audio. These are evenly split across three modality formats audio-only, video-only, and full audio-video and across speech (1,440 audio + 2,880

Figure 3.5: RAVDESS Emotion histogram

video clips) and song (1,012 audio + 2,024 video clips) performances from 24 professional actors (12 male, 12 female). The dataset is labeled with 8 discrete emotions (Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised) with two intensity levels (normal, strong) for all emotions except Neutral. In our work, only the visual speech portion of the dataset is used (distribution is in figure 3.5). [LR18]

3.4 Temporal Dynamics in Video

When recognizing emotions from video, leveraging temporal information changes in expression over time can improve accuracy.[HM17] There are two main strategies to tackle this. The first one is to combine CNN (for its spatial feature extraction on individual frames) with a Recurrent Neural Network (RNN), typically Long Short-Term Memory (LSTM), to capture temporal changes across frames.[Chu18] An alternative to the CNN + RNN combination is to use convolutional networks operating on spatiotemporal volumes of video, using time as an additional dimension. While these model architectures provide improved accuracy in the facial emotion recognition task over purely static single-frame-based recognition, these gains come at the expense of increased resource usage and complexity. [PMR+24]

3.5 Impact of environment on accuracy

Varying conditions significantly influence facial expression recognition accuracy. Poor lighting conditions can cause a substantial decline in model performance, reducing accuracy by up to approximately 30% compared to brightly lit scenarios [LSZ$^+$23].

Similarly, changes in camera angles also have a notable impact. When the subject's face is rotated to a three-quarter view rather than fully frontal, recognition accuracy typically decreases by approximately 10 25% across most emotional categories [MF25].

Chapter 4

Methodology

This chapter provides an overview of the methodology used in our work. First, the overall project handling of video is described. Next, we proceed with the selection of the dataset and data preprocessing. Lastly, the model architecture and subsequent training is described.

4.1 System overview

The system is made up of two parts: the user application, which handles interaction with the user, and a detection program, which handles the detection itself. The user application is a simple web app made with a React frontend and a Python backend. It is only a demo application allowing users to easily use the emotion detection system and showcasing its capability. The user can upload or record a video which is then saved and passed to the detection module. The detection module is written in Python, video is taken as input, individual frames are extracted, faces detected, and cropped to subsequently classify the facial expression of those faces into an emotion label using a pre-trained CNN model. The output of the detection program is then turned into a chart visualizing emotions present in the processed video.

4.2 Dataset choice and data preprocessing

For training and evaluating the model, the FER2013 image dataset is used with updated and more precise emotion labels of FER2013Plus. FER2013 is used due to its large size, as it contains 35,887 48×48 grayscale images including very diverse subjects. The official data split is used for training, validation, and testing purposes. Additionally, the CK+ and RAVDESS datasets are employed to help evaluate the model. They both contain images from a controlled environment.

Due to the selection of image datasets for training, there is a need to process the video to the same format. One in a couple of frames is extracted, this helps the system not to waste resources on frames that are almost identical. This proves crucial, especially when processing long video files. We need to be mindful as to where our processing time is allocated, while still capturing valuable information in the video.

Next, face detection using Haar Cascades is run, which allows cropping the frame to the desired 224×224 containing only the subject's face. Haar Cascade for face detection is used because it is lightweight and sufficient for frontal faces present within our domain. Any undetected or misdetected faces will affect results, but no individual frame that is not detected should disrupt the overall emotion detection during the processing of a long video. Also, due to our usecase, only one face present in the video.

Random augmentation on training images is performed. This allows the model to better generalize to images taken in the wild images taken in varying conditions. Techniques such as flipping, slight rotation (±15°), and brightness adjustment are used to provide variability. None of these adjustments disturb the dataset's labels, a vertically flipped face still conveys the same expression as the original not-flipped face. Data augmentation is a well established technique which is used to minimize overfitting when data is insufficient. [WWL+25]

Due to the fact that the FER2013Plus is a dataset with uneven class distribution (Figure 3.2), class imbalance handling is employed. Class specific weights (which are derived from the training dataset class distribution) are computed and applied during training. This forces the model to pay more attention to minority classes by increasing their impact on the loss function. This way, bias toward the majority class is reduced and it leads to a more balanced performance across classes.

4.3 Model architecture and training

In our work, we analyze emotions by classifying individual video frames independently, thus explicitly disregarding temporal dynamics. While temporal information can enhance emotion recognition accuracy by including some valuable information in between frames, our frame-by-frame approach is justified by the requirement for computational efficiency and simplicity. This approach also allows for easier model training and faster inference, which aligns with our constraints.

The MobileNetV2 convolutional neural network architecture pre-trained on ImageNet as a base model is utilized before fine-tuning on the chosen dataset. This architecture was chosen for its accuracy and computational efficiency. The pre-trained MobileNetV2 network was originally designed for image classification to 1,000 classes. This is adapted to our 8-class classification by removing the final classification layer and adding our own custom classification head tailored to our 8-class dataset.

Two-phase transfer learning strategy was adopted. In the first phase, the MobileNetV2 weights are frozen and the new classification head is being trained exclusively. In the subsequent phase, the top 20 layers of the network are unfrozen and finetuned with a decreased learning rate to better adapt learned representations specifically to facial emotion recognition.

The model is implemented in Python using the TensorFlow/Keras library, providing the pre-trained architectures and training utilities. OpenCV is used for face detection and image preprocessing.

Chapter 5

Demo application

This chapter details the practical implementation of the emotion detection demo application, building upon the methodological framework introduced in Chapter 4 Methodology. The focus here is on the technical realization of the system, made up of a backend processing engine and a user-facing frontend interface. This application aims to showcase the theoretical capabilities of the developed model and demonstrate its usage.

5.1 System Architecture

The demo application comprises a React client and a Flask server. A video is captured or uploaded by the user, the server extracts frames, detects faces, and classifies expressions, then returns the emotion statistics. The results are subsequently visualized to the user. The data flow of the system is visualized in the diagram (Figure 5.1).

5.2 Backend Implementation

The backend portion of the system consists of two parts. The first one is a Flask server where the communication with the frontend is handled, videos