Automatic Emotion Recognition from Videos. Does the video file format matter?

Ionut-Daniel Morjan
Department of Computer Science
Babeş-Bolyai University
Cluj-Napoca, Romania
ionut.morjan@stud.ubbcluj.ro

Grigoreta-Sofia Cojocar
Department of Computer Science
Babeş-Bolyai University
Cluj-Napoca, Romania
grigoreta.cojocar@ubbcluj.ro

ABSTRACT

Emotion recognition from video is an important capability for applications like driver monitoring or telemedicine. Although the latest state-of-the-art approaches achieve good accuracy, there is a gap in presenting these results. How do they perform if the video file quality is degraded by variation in resolution or compression rate, or codec? In this work, we compare two state-of-the-art approaches, Poster++ and DAN, by measuring their robustness when the video file quality is degraded. We used a high-quality MP4 sequence as our reference and we created versions that vary in resolution, compression bitrate, and codec. The results obtained showed that under moderate conditions, both approaches maintained consistency close to their high-quality baseline, with most frames yielding identical emotion labels, but when the video quality dropped further, differences emerged. At low resolution with heavy compression, Poster++'s accuracy fell into the mid-eighties for modern codecs, while DAN stayed above the low-nineties for those same inputs. An additional interesting result observed is that both models remained highly confident even when many frames were misclassified, indicating that confidence alone cannot detect poor input quality. Overall, our study demonstrates that while recent emotion recognition approaches can tolerate mild to moderate video degradation, severe downsampling or heavy compression can significantly compromise their performance.

Author Keywords

automatic emotion recognition, video quality,

CCS Concepts

•**Computing methodologies** → **Computer vision problems;**

DOI: 10.37789/icusi.2025.21

1. INTRODUCTION

Emotion recognition from video is a rapidly growing field at the intersection of computer vision and affective computing. It involves automatically identifying human emotional states from visual inputs, typically by analyzing facial expressions in video frames. This capability has numerous practical applications, from improving human-computer interaction and social robotics to enhancing driver monitoring systems and healthcare diagnostics. Facial expression analysis is central to video-based emotion recognition. Psychologists Ekman and Friesen identified six basic emotions: happiness, sadness, fear, anger, surprise, and disgust, which are universally reflected in facial expressions across cultures [4]. These basic emotions, to which sometimes contempt and a neutral state are added, form the typical classification targets for facial expression recognition (FER) approaches. The human face provides rich non-verbal cues: slight movements of the eyes, eyebrows, mouth, and other facial muscles combine to convey complex emotional states. Capturing these subtle facial movements in real-world conditions is a challenging task for automated systems. Early approaches to automatic emotion recognition relied on static images or controlled laboratory settings, but the field has evolved towards dynamic analysis in unconstrained videos. Video-based emotion recognition offers some advantages over still-image analysis by leveraging the temporal dimension; changes in facial expressions over time can provide additional clues, such as the progression of a frown or smile, which can help disambiguate similar expressions and improve robustness. However, working with video also introduces new challenges, including motion blur, varying lighting across frames, and the need to process a large number of frames efficiently. Additionally, in real-world applications, videos may be recorded under suboptimal conditions: cameras might have low-quality, bandwidth constraints may enforce high compression, and formats can vary. All these factors can affect the performance of emotion recognition approaches.

The existing facial emotion recognition approaches have evolved from simple feature-based methods to complex deep-learning based methods. However, one aspect that has not been sufficiently addressed in existing work is the robustness of these new methods under varied video file formats. Most benchmark datasets (RAF-DB [8], AffectNet [11], etc.) consist of relatively high-quality images or frames, but in real contexts video inputs are far from ideal. They could be low-resolution (i.e surveillance footage or video call thumbnails), compressed for streaming or recorded in non-controlled settings with lighting issues. There is no study in the literature about how the most recent emotion recognition approaches perform when the video quality is varied. This gap motivated our research. By comparing two emotion recognition approaches across different video file resolutions, bitrates, and formats, we aim to identify how they perform, whether they maintain robustness and whether they degrade.

The rest of the paper is structured as follows. Section 2 introduces the components of a video file, Section 3 presents the ideas behind facial expression analysis and emotion recognition, and Section 4 the steps emotion recognition from videos and their challenges. Our experiment and the obtained results are presented in Section 5. Threats to validity are given in Section 6 and Section 7 concludes the paper.

2. VIDEO FILE FORMATS

A video file is a type of digital file that can be played on TVs, phones, and computers. It contains both sound and moving images. Besides images and audio data, other optional data like metadata or subtitles are usually included. The following are the main components of a video file.

- **Video data**. A series of pictures that make up a moving picture is called video data. These pictures are called video frames. The quantity of pixels that comprise a video frame is known as *video resolution* and it determines the degree of clarity and detail. It can be represented by abbreviations like 720p, 1080p, or 4K, or using the width by height in pixels, such as 1920x1080. While lower resolutions contain fewer pixels and may appear blurrier, higher resolutions have more pixels, producing pictures that are more detailed.
- **Audio data**. The sound that comes with the video is referred to as audio data.
- **Metadata**. The title, creation date, resolution, and other details of the video are considered metadata.
- **Container**. The video and audio data are wrapped in a container, such as MP4, AVI, or MOV. The container is responsible for arranging and saving these elements, together with any additional data.
- **Codecs**. The audio and video data inside the container is compressed and decompressed using codecs. They decide how efficiently the data is stored and how it is played back. They are also important for managing large video files for playing, transmission, and storage. Some of the most known video codecs are AV1, VP9, H.264, and H.265 (HEVC).
- **Video bitrate**. It is the quantity of data required to depict a video per unit of time and it is usually measured in bits per second (bps), kilobits per second (kbps), or megabits per second (Mbps). More data is used to represent the video at a higher bitrate, which often produces greater visual quality, but also results in larger file sizes and possibly slower streaming speeds.

The most common video file formats are MOV (an Apple format, frequently linked to higher quality), AVI (an older format, still in use, but typically creates larger files), MKV (a flexible format, frequently used for high-quality video), and MP4 (a very popular format known for its versatility and compatibility with various devices and platforms).

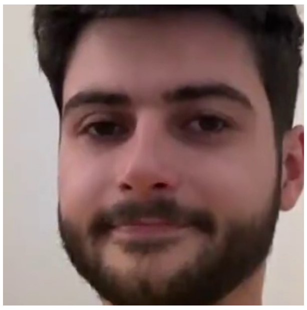

Figure 1: Example frame from a MP4 video at 720p and 1000kbps.

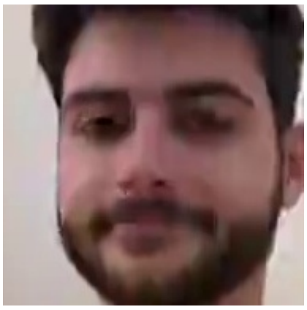

Figure 2: Example frame from a MP4 video at 360p and 128kbps.

To better understand the differences that the quality of a video file may bring, Figure 1 shows a video frame extracted from a MP4 video at 720p and 1000kbps, while Figure 2 shows the same video frame extracted from a MP4 video at 360p and 128kbps.

3. FACIAL EXPRESSION ANALYSIS AND EMOTION RECOGNITION

Facial expression analysis systems attempt to automatically analyze and recognize facial motions and facial feature changes from visual information. A key assumption in facial expression analysis and emotion recognition is that facial muscle movements correspond to displayed emotions [4]. For example, a smile (raising the lip corners) typically indicates happiness, while a furrowed brow might indicate anger or confusion. Psychologists tried to catalogue these facial muscle movements and their emotional interpretations [4]. Software systems also attempt to replicate this interpretative process by extracting visual features that correlate with facial muscle movements and mapping them to emotion categories. Early computer vision methods used hand-crafted features—such as distances between facial landmarks (eyes, mouth corners, etc.), or texture descriptors like local binary patterns on regions of the face to characterize expressions. These features were then fed into a classifier (i.e., SVMs or neural networks[9, 17]) trained to recognize different emotions. Such approaches worked well under constrained conditions but struggled with variations in lighting, pose, and individual differences in appearance. Modern approaches use deep learning, which can learn features automatically from large datasets of face images. Convolutional Neural Networks (CNNs) [7] have shown great success in static facial expression recognition (on single images) by learning hierarchical feature representations. In video (dynamic) emotion recognition, deep learning models often extend this by capturing temporal information. Recurrent Neural Networks or temporal convolution can be used to model how expressions evolve over time [6]. An alternative is to perform frame-by-frame emotion prediction using a CNN and then aggregate the predictions over time to decide the overall emotion. Both strategies benefit from the temporal continuity that video provides, potentially increasing accuracy over single-image methods. For example, a brief smirk might be missed in one frame but caught in another, or the context of successive frames can help differentiate a genuine smile from a fleeting expression.

Despite these advances, video-based emotion recognition remains challenging in unconstrained settings. Facial expressions can be subtle and brief. People in videos may move their heads or have partial occlusions (e.g., a hand on the face), and the illumination can change from one frame to the next. Another complexity is person-specific expression variability. Not everyone shows emotion in the same way, and models must generalize across different faces and demographics.

4. EMOTION RECOGNITION FROM VIDEOS

Emotion recognition from videos is a video analysis problem that consists of several processing steps, as illustrated in Figure 3, and described in the following:

- **Frame Extraction.** The first step is to extract the frames at a certain rate. For a given video, one might extract every frame or every *n*-th frame, depending on the needs. Higher frame rates preserve more temporal detail but also increase computational load. The extracted frames are typically stored as images in memory or on disk for further processing. It is important to ensure the frames are extracted in the correct order and timing so that the dynamics of expressions are preserved. Additionally, frames can be pre-processed (resized, converted to grayscale, or equalized for lighting) to normalize the input for the emotion recognition model. A tool like OpenCV [1] is often used for frame extraction.
- **Face Detection.** After the frames are obtained, detecting the face in each frame is a critical step. Face detection localizes the region of the image that contains a face, allowing the emotion recognition algorithm to ignore the background and focus on facial cues. A commonly used method for face detection is the Haar cascade classifier [15]. This method can process images in real time and is one of the first algorithms to enable face detection in live video. The use of Haar cascades in an emotion recognition from video settings means each video frame is scanned to find a face, and if found, the face region is extracted for emotion analysis.
- **Face Alignment and Normalization.** After detecting a face, some systems apply alignment, i.e., rotating and scaling the face so that the eyes or other landmarks are at predefined locations. This can compensate for head tilt or distance. Alignment often uses facial landmark detection (identifying points like the corners of eyes, tip of nose, etc.) and then warps the image to a standardized pose. Algorithms like the Active Shape Model [2] or more recent deep learning-based [7] landmark finders can be used for face alignment.
- **Feature Extraction.** Before classification, features are extracted from the face region. This might involve computing descriptors like Histogram of Oriented Gradients (HOG) [3] or Local Binary Patterns (LBP) [12] on the face image to encode textures and shapes of facial components. In deep learning methods, the raw pixel data of the face (or a resized version) is fed into a neural network, which produces a feature representation internally.
- **Emotion Recognition.** The previous steps laid the groundwork for emotion recognition by isolating faces and converting the raw video into a stream of face images or face features. The identified features are then fed into an emotion classification model that predicts the emotional state (happiness, sadness, etc.). In recent years, many deep-learning based approaches, such as Poster++[10] and DAN[16], were proposed for automatic emotion recognition from images.

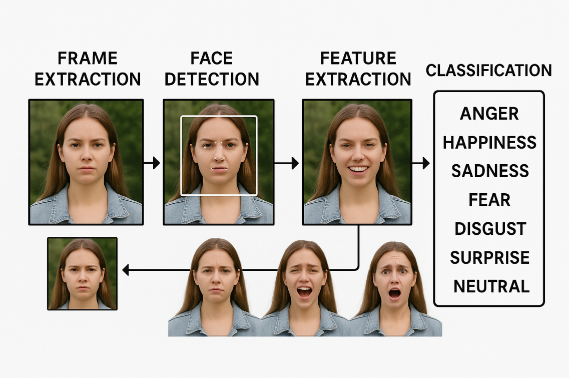

Figure 3: Video-based facial expression recognition pipeline.

Poster++

The POSTER++ network, proposed by Mao et al. [10], uses a transformer-based framework where one branch processes the face image and another branch processes facial landmark coordinates or a representation of facial geometry. The network then uses an attention mechanism to fuse these two—effectively, the model "knows" where key facial points are (such as eyes, eyebrows, mouth corners, etc.) and can pay special attention to those regions in the image. The strengths of POSTER++ include its high accuracy and efficiency. By fusing landmark and appearance data, it can be robust to variations such as head pose or slight occlusions—

even if part of the face is not clearly visible, the landmarks might still anchor the attention. It also tends to handle subtle expressions well, since landmarks capture small movements and the cross-attention can amplify those signals in the image features. One limitation of this approach is that it relies on having accurate facial landmark detection as an input. If the landmarks are inaccurate (which can occur if the video quality is poor or the face is at an extreme angle), it could mislead the model.

DAN

Another recent deep-learning approach used in emotion recognition is the DAN architecture, which stands for Distract-Your-Attention Network, introduced by Wen et al. [16]. It uses multiple attention heads to focus on different parts of the face and then distracts or encourages them not to all look at the same region, ensuring comprehensive coverage of facial features [16]. Wen et al. observed two key challenges for facial expression recognition:

- *Different emotions can look quite similar.* For example, anger and disgust share some facial characteristics (furrowed brow, tense mouth) and can be easily confused by a model because the underlying facial appearances overlap significantly. This means the feature space of expressions has clusters that are close together, making classification hard.
- *An emotion manifests in multiple facial regions simultaneously.* If a model only concentrates on one region (such as the mouth), it might miss critical information elsewhere (like the eyes). A holistic understanding requires integrating cues from various regions of the face.

One strength of DAN lies in its attention mechanism design. By covering multiple facial regions, DAN is less likely to miss important cues. If one region doesn't clearly indicate the emotion, another might. In terms of input requirements, DAN doesn't explicitly use landmarks or other modalities. It relies solely on image features and learned attention, which means it needs high-quality visual data to attend properly. If the video quality is poor (blurry or low resolution), the attention heads might struggle to find meaningful regions to focus on, or they might all focus on the least noisy part of the face.

Challenges in Video-Based Emotion Recognition

Real-world video data comes with many variations that pose challenges to emotion recognition systems:

- *Video resolution and quality:* The resolution of a video determines how much detail is present in each frame. High-resolution videos provide clear details of facial features, making it easier for algorithms to pick up subtle cues like wrinkles around the eyes or slight lip curves. However, low-resolution videos may render these fine details indistinguishable, a smile and a neutral expression might look similar if the mouth region only spans a few pixels. When resolution drops, the face detector might even struggle to find the face, and any recognition model has less information to work with. These quality degradations can significantly reduce emotion recognition accuracy, and the system may misclassify emotions or become less confident in its predictions when key facial features are obscured.
- *Video compression and bitrate:* Most digital videos are compressed using codecs to reduce file size. Compression removes redundant information and can introduce artifacts, especially at low bitrates. Typical artifacts include blockiness, blurriness, and motion artifacts where moving parts of the face may lag or ghost. These artifacts alter the true appearance of the face in the image. For instance, heavy compression might smooth out the wrinkles of a frown or create blocky noise around the eyes, confusing the feature extraction. Research has shown that compression-induced distortions can lead to loss of facial detail and a drop in recognition accuracy [14]. In highly compressed video, an algorithm might even detect a different expression than intended, or fail to recognize any clear expression. Pitrey and Hlavács [14] specifically noted that when video quality is degraded by compression or downscaling, facial expression features become distorted, which can either yield incorrect emotion predictions or none at all. They also observed that compression artifacts can distort facial features to such an extent that viewers may experience strange feelings. While this response is psychological, it demonstrates how drastically facial appearance can be altered by compression.
- *Video file formats and codecs:* The format of a video can refer to both the container (e.g., MP4, AVI, MKV) and the codec used to encode the video stream (H.264, VP9, etc.). Generally, the choice of container format (MP4 vs AVI) does not inherently affect the visual content, as it is mostly a wrapper. What matters more is the codec and its settings (bitrate, keyframe interval, etc.). Different codecs, however, have different compression efficiencies and artifact characteristics. For example, some codecs might introduce more blurring, while others create more blocking. If an emotion recognition approach was developed and tested mostly on videos from a certain source or codec, a drastically different compression scheme could affect its performance.
- *Lighting, pose, and occlusion:* Other worth mentioning, well-known challenges correlate with real-world video quality. Lighting changes (e.g., glare, shadows) can occur across frames and severely impact face analysis. A person turning their head (pose variation) might momentarily give a non-frontal view, making it difficult for a detector or requiring the recognition algorithm to handle side views. Occlusions, like the person touching their face or wearing glasses or masks, can hide facial features. These factors can be exacerbated by low resolution and compression. For example, a slight turn of the head combined with low resolution might render the face barely detectable. Robust models attempt to mitigate these issues by data augmentation (training on varied conditions) or by using temporal information (if a face is occluded in one frame, maybe the next frame is clear).

In our study, we are particularly interested in how the video file quality, compression, and format variations affect the performance of emotion recognition approaches. Controlling the other factors (lighting, pose, and occlusion) is quite complex and was not in the scope of our study.

5. STUDY

Study's objective

Facial emotion recognition has evolved from simple feature-based methods to sophisticated deep networks leveraging attention and multi-stream information. POSTER++ [10] and DAN [16] stand out as state-of-the-art approaches that have obtained the best accuracy on standard datasets. However, one aspect that has not been thoroughly addressed in existing work is the robustness of these advanced approaches under varied video quality conditions. Most benchmark datasets (RAF-DB, AffectNet, etc.) consist of relatively high-quality images or frames. The assumptions are that faces are clearly visible and reasonably large. However, in real-world applications, video inputs might be far from ideal. They could be low resolution (i.e., surveillance footage or video call thumbnails), or compressed for streaming (introducing artifacts), or simply recorded in non-controlled settings with lighting issues. We did not find in the literature a comparative analysis of how top emotion recognition models like POSTER++ or DAN perform when the video quality is systematically varied. There are a few studies (such as Pitrey and Hlavács [14]) hinting at the drop in performance with compression, but those were done with older techniques.

The focus of our experiment is to understand how the video quality factors impact the models (Poster++ and DAN) performance in recognizing the seven basic emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral), as we have previously observed that for Figure 1 Poster++ predicted neutral emotion, but for Figure 2 predicted disgust.

Methodology

To determine if the video file format affects automatic emotion recognition from videos, a video of 9 minutes and 24 seconds, recorded at 30 frames per second, was used, which filmed a person speaking, and, in each frame, the face of the person was visible. In the absence of manually annotated emotion labels for the videos, the model's predictions on the highest-quality video (MP4 format, 720p, 1000 kbps) are assumed to be the "true" or baseline reference. This approach provides a consistent baseline to evaluate how much the model's predictions change when the same video content is degraded. While this is an approximation, as the model's baseline output may itself contain errors, it is a reasonable proxy for ground truth in a comparative study. This methodology is similar in spirit to practices in video quality assessment and has been implicitly used in prior work where high-quality data are treated as an oracle to evaluate degraded inputs [5]. The underlying assumption is that the model's output on a high-quality video is the most reliable; any deviations caused by lower-quality inputs indicate performance degradation.

Considering all possible combinations for a video file: container (MP4, AVI, WebM, MOV, MKV, OGV, 3GP, WMV), resolution (720p, 480p, or 360p), and bitrate (1000kbps (high), 500kbps(moderate), or 128 kbps(very low)), we obtained 65 possible variants, as 3GP is an older format container, and the only accepted resolution is 360p, and MP4 with resolution 720p and bitrate 1000kbps was considered the baseline.

For each variant video, we performed frame-by-frame emotion recognition. We first identified the face in each frame, and then used the pre-trained POSTER++ and DAN models (the same model and parameters used for the baseline, with no fine-tuning) on the face's region. Afterwards, we compared the predicted emotion for the frame of that variant to the baseline's frame's prediction. In the end, we obtained 130 results: 65 results from using Poster++ and 65 results from using DAN.

Video Conversion Process

All video variants were generated from the original source video using the FFmpeg tool [13]. FFmpeg is a widely-used, open-source command-line utility for video/audio processing and encoding. It allows precise control over output format, codec, resolution, and bitrate, making it suitable for our systematic experiment. The conversion was automated with scripted FFmpeg commands to ensure consistency across all variants. Each variant had a specified container format (e.g., MP4, AVI, WebM, MOV, MKV, OGV, 3GP, WMV), a target resolution (360p, 480p, or 720p), and a video bitrate (128kbps, 500kbps, or 1000kbps). We selected codecs appropriate to each container format in order to produce playable, standard-conforming files. We used the H.264 video codec for MP4 and MKV outputs, since MP4 typically contains H.264 video with AAC audio. WebM is traditionally associated with the VP8/VP9 video codecs, but in our experiment, we also used H.264 for WebM to isolate the effect of the container itself. OGV files (Ogg Video) were encoded with the Theora codec for video and Vorbis for audio. 3GP (a format for older mobile devices) was encoded with the H.263 video codec and AAC audio, consistent with the 3GP specification for broad compatibility. MOV (QuickTime) was created using an older MPEG-2 video codec (mpeg2video) and MP3 audio, reflecting the legacy usage of MOV with MPEG codecs. AVI, being an older container format, was encoded with a legacy MPEG-4 Part 2 codec to ensure it would play back in default players without modern codec support. Finally, WMV (Windows Media Video) was encoded using a Windows Media codec (specifically the WMV2 codec for video) and WMA for audio. All conversions maintained the same frame rate as the original video (30 frames per second) and used a constant frame size corresponding to the target resolution (no letterboxing or cropping was introduced beyond scaling the frames).

Evaluation Metrics

For the baseline video, we computed the **BaseNoFrames**, the number of frames in which the face was detected, and the **AvgConfBaseline**, the confidence score of the model (Poster++ or DAN).

For each variant video, we computed the following:

- **VNoFrames** The number of frames in which the face was detected using the Haar algorithm.
- **VMatchedEmotion** The number of frames for which the identified emotion was the same as the emotion identified for the baseline.
- **VMismatchedEmotion** The number of frames for which the identified emotion was not the same as the emotion identified for the baseline.
- **VAccuracy** The percentage of frames for which the identified emotion was the same as the emotion identified for the baseline video, if the face was detected in both frames ($VAccuracy = MatchedEmotion/VNoFrames * 100$).
- **VAccuracyALL** The percentage of frames for which the identified emotion was the same as the emotion identified for the baseline ($VAccuracyALL = MatchedEmotion/BaseNoFrames * 100$).
- **AvgConfVariant** The confidence score of the model (Poster++ or DAN) on the variant video, to determine whether there are any differences with the **AvgConfBaseline**, as in general, the models tend to be slightly less confident on lower-quality inputs.

Results

For the baseline video, we obtained 13216 frames which contained faces ($BaseNoFrames = 13216$), and the **AvgConfBaseline** is 0.8610 for Poster++ and 0.8646 for DAN, respectively.

As the results obtained for the selected video exceed 100 lines, in Table 1 we present the best 10 results and in Table 2 the worst 10 results. The **Variant** column represents the combination used for the video format, having the following format *container_resolution_bitrate*.

An initial observation from these results is that under various formats, resolutions and bitrate conditions, the Haar algorithm is not able to properly identify the face. As the video format is seriously degraded (using the lowest resolution and the lower bitrate) in almost 27% of the frames the face is not recognized.

Impact of Video Resolution

Consistent with expectations, we found that higher resolution video leads to higher emotion prediction consistency with the baseline. When the video was kept at full 720p resolution (but with no other changes like format or bitrate), the models typically showed almost no drop in performance. At 480p, which is a moderate reduction in resolution (44% fewer pixels per dimension than 720p), there was typically a slight drop in performance, but still generally high. The 360p resolution (a significant downsampling from the baseline, with only 50% of the width/height of 720p, i.e. a quarter of the pixel count) presented a bigger challenge. At 360p, we observed noticeable drops in consistency, especially for the lowest-quality streams. When video quality was otherwise decent (e.g. using a high bitrate and an efficient codec), the drop due to resolution alone was moderate. The real impact of resolution became clearer when interacting with other factors: at 360 p combined with low bitrate, performance suffered markedly. But isolating resolution, one interesting observation is that DAN sometimes showed greater resilience to resolution loss than Poster++ in specific contexts. For example, in one scenario with extremely low resolution but also low quality (360 p at 128 kbps in MP4), DAN achieved 91.41% consistency, slightly higher than Poster++ at 83.08% for the same scenario. In summary, resolution reduction alone (from 720 p down to 480 p or 360 p) causes a gradual decline in emotion recognition consistency, but moderate resolutions (480 p) are largely sufficient for these models. Only at fairly low resolutions (360 p or below) do we begin to see substantial divergence, reinforcing the idea that ensuring a minimum face resolution (on the order of fifty pixels height or more) is important for reliable emotion recognition.

Impact of Video Compression Bitrate

For most formats and resolutions, moving from 1000 kbps to 500 kbps did not dramatically reduce performance. In fact, in several cases the consistency remained virtually the same, or dropped by only a few tenths of a percent, when cutting the bitrate in half. For example, Poster++ at 480 p matched about 83.92% of frames at 1000 kbps and about 83.90% at 500 kbps in MP4, a negligible difference. DAN likewise showed almost no change between 1000 kbps (92.11% consistency) and 500 kbps (92.04%) at 480 p, indicating that when the video is of reasonably good visual quality ("DVD quality" or better), the models are near their peak performance and minor compression differences have minimal impact. In other words, as long as the visual quality is "good to excellent," emotion recognition loss is negligible. This trend holds especially at lower resolutions: for a 360 p video, 1000 kbps is more than enough to preserve quality (since the frame is small), and even 500 kbps is generally sufficient to keep artifacts low. The real effects of compression emerge at the very low bitrate of 128 kbps. Here we start to see significant drops in performance, especially at higher resolutions. A striking finding is that 128 kbps at 720 p resolution was particularly damaging to consistency. For instance, Poster++ at 720 p/128 kbps (MP4) only achieved about 84.79% consistency (compared to around 90.38% at 500 kbps, and nearly 100% at the original quality). DAN managed about 91.86% at 720 p/128 kbps in the same scenario—higher than Poster++—perhaps indicating DAN could cope with the blurring and noise at high-resolution, low-bitrate inputs slightly better. But in other formats, both models struggled: for example, with a less efficient codec at 720 p/128 kbps, consistency could drop into the 70–80% range. In general, 720 p at 128 kbps introduced enough artifacts that the models frequently disagreed with their baseline predictions—roughly 1 in 7 frames (about 15%) changed emotion label for Poster++ in MP4, and almost 8% changed for DAN—highlighting the severity of extreme compression.

Impact of Video Encoding Format

The choice of video codec significantly influences emotion recognition accuracy. Container changes alone (MP4, MOV, MKV) have no impact when they preserve the identical H.264 stream. VP9 (WebM) consistently outperforms H.264 at constrained bitrates; for instance, at 720p and 128kbps, WebM achieves 89.01% consistency, well above

Variant	VNoFrames	VMatchedEmotion	VMismatchedEmotion	VAccuracy	VAccuracyALL	AvgConfVariant	Model
mkv_720p_1000kbps	13216	13216	0	100	100	0.8610	POSTER++
mov_720p_1000kbps	13216	13216	0	100	100	0.8649	DAN
mkv_720p_1000kbps	13216	13216	0	100	100	0.8649	DAN
mov_720p_1000kbps	13216	13214	2	99.9849	99.9849	0.8609	POSTER++
ogv_720p_1000kbps	12830	12175	655	94.8947	92.1232	0.8674	DAN
webm_720p_1000kbps	12790	12089	701	94.5192	91.4725	0.8638	DAN
webm_720p_500kbps	12768	12073	695	94.5567	91.3514	0.8654	DAN
wmv_720p_1000kbps	12752	12031	721	94.3460	91.0336	0.8661	DAN
mp4_720p_500kbps	12770	12029	741	94.1973	91.0185	0.8677	DAN
mov_720p_500kbps	12770	12029	741	94.1973	91.0185	0.8677	DAN

Table 1: The top 10 results obtained under various formats, resolutions, and bitrate conditions.

Variant	VNoFrames	VMatchedEmotion	VMismatchedEmotion	VAccuracy	VAccuracyALL	AvgConfVariant	Model
ogv_360p_128kbps	11938	8828	3110	73.9487	66.7978	0.8382	POSTER++
ogv_720p_128kbps	11340	8411	2929	74.1711	63.6426	0.8139	DAN
ogv_480p_128kbps	11782	8010	3772	67.9851	60.6084	0.8331	POSTER++
3gp_360p_128kbps	9651	6763	2888	70.0756	51.1728	0.7233	DAN
ogv_720p_128kbps	11340	6616	4724	58.3422	50.0605	0.8381	POSTER++
3gp_360p_1000kbps	9685	6259	3426	64.6257	47.3593	0.7105	DAN
3gp_360p_500kbps	9650	6245	3405	64.7150	47.2533	0.7145	DAN
3gp_360p_1000kbps	9685	4618	5067	47.6820	34.9425	0.7797	POSTER++
3gp_360p_500kbps	9650	4601	5049	47.6788	34.8139	0.7785	POSTER++
3gp_360p_128kbps	9651	4426	5225	45.8605	33.4897	0.7726	POSTER++

Table 2: The worst 10 results obtained under various formats, resolutions, and bitrate conditions.

MP4's 84.79%. AVI (MPEG-4 ASP) yields lower performance at limited bitrates and medium resolutions—77.65% at 480p and 128kbps versus MP4's 82.48%. OGV (Theora) and 3GP (H.263) severely degrade performance unless bitrates are very high—OGV at 720p and 128kbps matches only 58.34%, and 3GP at 360p and 1000kbps matches only 47.68%. WMV (WMV9) performs on par with MP4 at high bitrates—90.39% at 720p and 1000kbps—but falls behind at low bitrates—73.59% at 360p and 128kbps. These empirical findings demonstrate that modern, efficient codecs such as H.264 and VP9 preserve facial features much better for POSTER++, whereas older or less-efficient codecs introduce artifacts that lead to frequent misclassification.

Confidence

Under most degraded conditions, the confidence of the models on those same frames dropped by a small amount (usually a few percentage points). This indicates that the classifiers are aware of some ambiguity introduced by lower quality, and they are not as completely sure of their predictions as they were on the clear video. However, the confidence reduction is often modest, and in some cases, the degraded video even yielded higher confidence than the baseline for a model. The latter counter-intuitive phenomenon can happen if compression artifacts or smoothing cause the model to latch onto a particular expression more strongly (even if that may be a misclassification). Overall, an important conclusion is that a high confidence from the model does not always mean the input was high quality. The models can be confidently wrong when video quality distorts the facial cues.

6. THREATS TO VALIDITY

For our study, we have identified the following threats:

- *Number and selection of video.* In our study we used only one video, and we selected it considering the format used for recording, its content, and its quality. We wanted to use a video that has only one face in each frame and the face appears in as many frames as possible. An increased number of faces in a frame may influence the results of face extraction and emotion recognition for the variant videos.
- *Frame rates.* In our experiments, we used only one frame rate, namely 30 frames per second, for the baseline and the variant videos. Different frame rates (e.g. 25 or 50 frames per second, etc) may yield different results.

7. CONCLUSIONS

Bringing all the results together, we can draw several important conclusions about the impact of video quality on facial emotion recognition and the comparative behavior of the two evaluated models:

- Both POSTER++ and DAN handle moderate video quality reductions with minimal impact on their predictions. In practical terms, this means that if an emotion recognition system processes standard definition video or a moderately compressed stream, it will perform almost as well as on a high definition, high bitrate feed. Occasional frames might be misclassified, but the overall emotion trend or dominant expressions will generally be captured correctly. When video quality drops severely, such as very low resolution (360 p) combined with very low bitrate (128 kbps), the models begin to disagree with their high-quality predictions more frequently. A substantial fraction of frames can be assigned a different emotion than they would at high quality. In a real-world deployment, this is critical; if the application cannot guarantee minimally decent video quality, the reliability of emotion inference becomes highly questionable. At the extreme, more than one in three frames may be misclassified under archaic codecs, or one in five under extremely low-bitrate H.264.
- Not all video formats preserve facial expression information equally. Modern, efficient codecs such as H.264 in

MP4, MOV, or MKV, and VP9 in WebM, maintain high recognition consistency even at lower bitrates.

- Although both models are generally robust, they exhibit different failure modes. POSTER++ tends to falter more under extreme compression noise, while DAN shows greater vulnerability to very low resolution under archaic codecs. These complementary strengths suggest a hybrid approach: one might use POSTER++ when faces are small but relatively clear, and DAN when faces are larger but potentially heavily compressed.

- Our results reinforce prior research showing that both downsampling and heavy compression negatively affect facial expression recognition. Training data should include low-quality variants so that the model learns to be invariant to those distortions. The differences between POSTER++ and DAN also highlight how architectural choices—such as leveraging temporal information or extracting specific convolutional features—affect robustness.

In conclusion, while state-of-the-art emotion recognition models are fairly robust to mild reductions in video quality, they are not immune to severe degradation. Video resolution and bitrate each have a significant effect on performance, and using efficient, modern codecs is crucial to maintain accuracy. POSTER++ and DAN handle these challenges in slightly different ways, underscoring that robustness is an essential dimension in model evaluation. Future work could combine POSTER++'s spatial resilience with DAN's compression resilience to achieve more consistent performance across a wider range of video qualities. Our findings could also help practitioners set realistic expectations and understand limitations when deploying emotion recognition on real video streams.

REFERENCES

[1] Bradski, G. 2000. The OpenCV Library. https://www.drdobbs.com/open-source/the-opencv-library/184404319. (2000). Accessed: 2025-05-28.

[2] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. 1995. Active Shape Models—Their Training and Application. https://doi.org/10.1006/cviu.1995.1004, *Computer Vision and Image Understanding* 61, 1 (1995), 38–59.

[3] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. IEEE, 886–893.

[4] Ekman, P. and Friesen, W.V. 1971. Constants across Cultures in the Face and Emotion. *Journal of Personality and Social Psychology* 17, 2 (1971), 124–129.

[5] Experts Illinois. 2017. Robust emotion recognition from low quality and low bit rate video: A deep learning approach. http://bit.ly/43XQWBB. (2017). Accessed: 2025-06-03.

[6] Alex Graves. 2013. Generating Sequences With Recurrent Neural Networks. https://arxiv.org/abs/1308.0850. (2013). Accessed: 2025-06-12.

[7] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep Learning. https://www.nature.com/articles/nature14539. (2015). Accessed: 2025-06-12.

[8] Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2584–2593. DOI: http://dx.doi.org/10.1109/CVPR.2017.277

[9] Heng Liu, Ying Sun, Miao Zhang, Qi Wang, and Lei Zhang. 2022. Deep Learning Approaches for Emotion Recognition: A Survey of Convolutional Neural Networks and Recurrent Neural Networks. https://arxiv.org/abs/2203.09213v1. (2022). Accessed: 2025-06-12.

[10] Mao, J., Xu, R., Yin, X., Chang, Y., Nie, B., and Huang, A. 2023. POSTER++: A simpler and stronger facial expression recognition network. https://doi.org/10.48550/arXiv.2301.12149. (2023). Accessed: 2025-06-06.

[11] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. 2019. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing* 10, 1 (2019), 18–31. DOI:http://dx.doi.org/10.1109/TAFFC.2017.2740923

[12] Mark Ojala, Matti Pietikäinen, and Timo Mäenpää. 2002. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 7 (2002), 971–987. DOI: http://dx.doi.org/10.1109/TPAMI.2002.1017623

[13] Parentnode.dk. 2022. Internet Video Optimisation Using FFmpeg. https://parentnode.dk/blog/internet-video-optimisation-using-ffmpeg. (2022). Accessed: 2025-05-25.

[14] Pitrey, R. and Hlavács, H. 2013. Impact of Video Quality on Automatic Facial Expression Recognition. https://www.isca-archive.org/pqs_2013/pitrey13_pqs.pdf. In *Proceedings of the International Workshop on Perception and Quality of Sensor Signals (PQS)*.

[15] Viola, P. and Jones, M. 2004. Robust Real-Time Face Detection. https://doi.org/10.1023/B:VISI.0000013087.49260.fb, *International Journal of Computer Vision* 57, 2 (2004), 137–154.

[16] Wen, Z., Lin, W., Wang, T., and Xu, G. 2023. Distract Your Attention: Multi-Head Cross Attention Network for Facial Expression Recognition. https://doi.org/10.3390/biomimetics8020199, *Biomimetics* 8 (2023), 199.

[17] Jie Yang, Xian Zhang, Li Liu, Kai Zhao, and Weiming Zhang. 2023. Facial Emotion Recognition Using Support Vector Machines and Feature Fusion. https://arxiv.org/abs/2308.10755v1. (2023). Accessed: 2025-06-12.