Facial Expression Recognition with Adaptive Frame Rate based on Multiple Testing Correction

Andrey V. Savchenko [1 2 3]

Abstract

In this paper, we consider the problem of the high computational complexity of video-based facial expression recognition. A novel sequential procedure is proposed with an adaptive frame rate selection in a short video fragment to speed up decision-making. We automatically adjust the frame rate and process fewer frames with a low frame rate for more straightforward videos and more frames for complex ones. To determine the frame rate at which an inference is sufficiently reliable, the Benjamini-Hochberg procedure from multiple comparisons theory is employed to control the false discovery rate. The main advantages of our method are an improvement of the trustworthiness of decision-making by maintaining only one hyper-parameter (false acceptance rate) and its applicability with arbitrary neural network models used as facial feature extractors without the need to re-train these models. An experimental study on datasets from ABAW and EmotiW challenges proves the superior performance (1.5-40 times faster) of the proposed approach compared to processing all frames and existing techniques with early exiting and adaptive frame selection.

1. Introduction

Affective behavior analysis and understanding of people's emotions should be essential to a new generation of human-centered interfaces, digital assistants, and personal advertisements (Kollias, 2022). It is known that face is the biometric of choice for this task because of its desirable properties: high accuracy of decisions, low cost of equipment, ease of use, etc. (Jillela & Ross, 2009). Video-based facial expression recognition (FER) is one of the most challenging problems in facial analytics due to the ambiguity of labeling emotions in large datasets, various intensities of the same emotion, and high imbalance of emotions in different situations (Wang et al., 2022). Moreover, since real-world settings entail uncontrolled conditions, FER systems should be robust to various contexts and video recording conditions (Ryumina et al., 2022).

Nevertheless, there exist many methods (Savchenko, 2023; Zhang et al., 2022) that demonstrate reasonable accuracy in several challenges, such as ABAW (Affective Behavior Analysis in-the-Wild) (Kollias, 2022) and EmotiW (Emotion Recognition in-the-Wild) (Dhall, 2019). At first, the majority of recent techniques (Jeong et al., 2022; Li et al., 2019; Savchenko et al., 2022) perform two steps for each video frame: (1) face detection; and (2) facial feature extraction. Even if the further processing is not computationally complex, these steps typically employ slow inference in deep neural networks. Unsatisfactory performance is exceptionally challenging for the second part, in which it is essential to use complex models to reach high accuracy.

It is known that video data is often repetitive: the contents of adjacent frames are usually strongly correlated (Dutson et al., 2022). Hence, a lot of research has been done in processing only a tiny fraction of available video frames by selecting a subset of salient frames (Korbar et al., 2019; Wu et al., 2019b) or conditionally computing using early exiting (Ghodrati et al., 2021; Lim et al., 2022). Unfortunately, most of these methods have been developed for action recognition problems (Yeung et al., 2016) and cannot be directly applied to FER with the same gain in performance due to the following reasons. At first, it is impossible to rapidly analyze the quality of faces on each frame by a lightweight neural net (Lin et al., 2022; Wu et al., 2019b) without running time-consuming face detection and/or tracking for each frame. Secondly, existing emotional datasets are small and dirty due to the high complexity of labeling emotions. The trained models learn too many features specific to a concrete dataset, which is impractical for in-the-wild settings. As a result, it is practically impossible to skip frames reliably using

[1] Sber AI Lab, Moscow, Russia [2] ISP RAS Research Center for Trusted Artificial Intelligence [3] HSE University, Laboratory of Algorithms and Technologies for Network Analysis, Nizhny Novgorod, Russia. Correspondence to: Andrey V. Savchenko <andrey.v.savchenko@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

learned models, such as reinforcement learning (RL)-based policy (Wu et al., 2019b), recurrent neural networks (Yeung et al., 2016) or even MLP (multi-layered perceptrons) in gating models (Ghodrati et al., 2021). Finally, the main difficulty of FER is the potentially rapid changes in an emotional state, so it is necessary to process a video on a fine-grained scale. Indeed, it is enough to make a single decision in video classification or action detection for a clip with a duration of dozens of seconds. Still, it is essential to recognize emotions in near real-time with less than a second delay. The fewer frames available in the input video, the lower the relative efficiency of frame selection algorithms.

Hence, this paper examines the possibility of improving the efficiency of FER in videos without significant degradation in accuracy. Our main contribution is the novel framework that can be applied with an arbitrary emotional feature extractor, frame pooling strategy, and video classifier. The proposed method adaptively selects the frame rate in a short video fragment and processes fewer frames with a low frame rate for more straightforward videos and more frames for complex ones. To automatically determine the frame rate whether to stop inference, we use the ideas of the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) from multiple comparisons theory (Hochberg & Tamhane, 2009) to control the false discovery rate by using the confidence of classifiers. The recognition trustworthiness is improved by maintaining only one hyper-parameter, namely, false acceptance rate (FAR). As a result, the running time of FER by our adaptive technique is 2-40 times lower compared to traditional processing of all frames, while the accuracy typically degrades only by 0.1-0.4%. The source code of the proposed method is publicly available[1].

2. Related Literature

2.1. Facial Expression Recognition in Videos

As was stated in Introduction, the typical emotional video datasets are dirty and small. Hence, the vast majority of techniques for FER in the video are based on the frame-level feature extraction with the neural networks trained on face identification and FER in static photos using such large datasets as AffectNet (Mollahosseini et al., 2017). The progress in the video-based FER has been estimated initially on the AFEW (Acted Facial Expression In The Wild) dataset from EmotiW 2013-2019 challenges (Dhall, 2019). For example, the FAN (frame attention network) for features extracted by the lightweight ResNet-18 model reached a validation accuracy of 51% (Meng et al., 2019). In contrast, the accuracy of the same ResNet-18 architecture trained by the noisy student iterative procedure (Kumar et al., 2020) is higher than 55%. One of the best single models is the family of EmotiEffNets, which are EfficientNets pre-trained using a robust optimization on VggFace2 and AffectNet datasets (Savchenko, 2021a; Savchenko et al., 2022). The winners of the last EmotiW 2019 audio-visual challenge are ensembles of several classifiers, namely, bi-modality fusion (Li et al., 2019) and cross-modal feature fusion with factorized bi-linear pooling (FBP) (Zhou et al., 2019).

The AFEW dataset is very small, and it lacks frame-level labels. As a result, the video emotion classifiers trained on this set have very low accuracy in cross-domain settings. This problem has become a focus of many researchers since the appearance of ABAW challenges (Kollias, 2022) that involve different parts of rather large Aff-Wild and Aff-Wild2 databases (Kollias et al., 2019). These challenges were focused on the dynamic nature of the emotional state, so it is necessary to make frame-level predictions. Due to its complexity and high imbalance, the macro-averaged F1-score on its validation and test sets in Expression classification ABAW-3 competition of even the state-of-the-art techniques based on Transformers (Karpov & Makarov, 2022; Zhang et al., 2023) are relatively low (30-40%). For example, the RegNetY-based transformer (Phan et al., 2022) is only 7-8% more accurate than such simple baseline as the VGGFace with new classification head (Kollias, 2022). The fastest algorithm and a single model among the top performers are the EmotiEffNets (Savchenko, 2022). A bit greater F1-score is obtained by an ensemble of the Swin-S, IR152, HRNet, and the RepVGG (Xue et al., 2022). A method based on the ensemble of multi-head cross-attention networks for facial features extracted by ResNet50 was proposed in the paper (Jeong et al., 2022). The winner of the recent challenge (Zhang et al., 2022) introduced a transformer-based fusion module that integrates the static vision features and the dynamic multimodal features from adjacent frames.

2.2. Efficient Video Recognition

Frame-level feature extraction is the most computationally complicated step in a typical video classification pipeline. There are two types of methods used to speed-up the video processing. First, it is possible to improve the efficiency of feature extraction by compressing the original models with structural pruning or quantization (Grachev et al., 2017) and/or adaptive inference computational graphs with early exits (branches) along several hidden layers of a neural net (Teerapittayanon et al., 2016). Though such methods can be successfully applied for very deep networks, they are not appropriate for lightweight architectures (Savchenko, 2021b), such as ResNet-18 (Kumar et al., 2020), MobileNet or EfficientNet (Savchenko, 2021a) mentioned above, that show the state-of-the-art performance for several FER datasets and challenges.

[1] https://github.com/HSE-asavchenko/face-emotion-recognition/

Hence, this paper focuses on the second group of methods that decreases the number of frames processed. The simplest solution is to reduce the frame rate and process only every k-th frame. Unfortunately, it is practically impossible to reliably choose the hyper-parameter k to balance between accuracy and complexity for an arbitrary input video. Hence, most existing works reduce the computational cost by solving the frame selection problem (Lin et al., 2022). It is the same technique discussed in the previous paragraph but adaptively selects frames rather than layers/units in neural networks for fast inference. Its application for face identification is studied in (Jillela & Ross, 2009), where frames can be automatically disregarded based on inconsistencies with optical flow.

Most modern algorithms in this direction have been developed based on RL for action recognition tasks. One of the first methods, FrameGlipses (Yeung et al., 2016), formulated the model as a recurrent neural network-based agent that observes video frames and decides where to look next and when to emit a prediction. A similar end-to-end deep reinforcement approach was proposed in (Fan et al., 2018), which enables an agent to classify videos by watching a tiny portion of frames. Another technique that uses complex RL, AdaFrame (Wu et al., 2019b), contains a Long Short-Term Memory (LSTM) network trained with a policy gradient method to generate a prediction, determine which frame to observe next and compute the expected future reward of seeing more frames at each time step. The LiteEval (Wu et al., 2019a) contains coarse and fine LSTMs and exploits features derived at a coarse scale with a lightweight model. The AdaFocus (Wang et al., 2021) tries to localize the most informative region in each frame (small image patch) by using a light-weighted ConvNet to quickly process the entire video sequence, whose features are operated by a recurrent policy network to localize the most task-relevant regions. The Dynamic-STE (StudentTeacher Ensemble) employed two networks of different capabilities: the lighter network processes more frames while the heavier one only processes a few (Kim et al., 2021). The D-STEP (Dynamic Spatio-Temporal Pruning) (Raviv et al., 2022) is a cascade of lightweight policy networks to dynamically filter out channels and regions that do not provide information. The recent technique that models temporal sampling as a decision-making process with RL is the OCSampler (Lin et al., 2022). It processes a whole sequence of frames at once rather than picking up frames sequentially by deriving the policies from a light-weighted skim network. Unfortunately, all such methods contain lightweight networks to process each frame efficiently, so they can hardly improve the FER's speed because of the need for face detection and efficient and accurate emotional feature extractors, e.g., EmotiEffNets (Savchenko et al., 2022).

An exciting idea is studied in (Gao et al., 2020) that uses audio as a preview mechanism to eliminate short-term and long-term visual redundancies. It cannot be used if the audio modality is unavailable or the training set with synchronous audio and video tracks is as small as a typical FER dataset (Khokhlova & Savchenko, 2014). The SMART (Gowda et al., 2021) considers frames jointly using an attention mechanism instead of selecting one at a time to look for suitable frames more effectively distributed over the video. The AR-Net (Meng et al., 2020) sets the optimal resolution for each frame in long untrimmed videos and processes the frames with different resolutions based on their relative importance. Similarly, VideoIQ (Video Instance-aware Quantization) (Sun et al., 2021) trains a very lightweight network in parallel with the recognition network to produce a dynamic policy indicating which precision to be used per frame in recognizing videos. As the resolution of typical facial models is much smaller than for the action detection problems, one can hardly expect a significant speed-up in using a small resolution for FER and face-related parts.

An adaptive frame selection network (AFSNet) (Tao & Duan, 2023) selects the most valuable frames in the image sequence by stacking some adaptive frame selection convolutions. The FrameExit (Ghodrati et al., 2021) employs a deterministic frame sampling strategy and a cascade of gating MLP modules to automatically determine the earliest point in processing where a decision is reliable. The SCSampler (Korbar et al., 2019) is a lightweight clip-sampling model that aggregates temporal information from long videos so that it may be inefficient for short video fragments and rapid changes in facial expressions. A similar model, NSNet (Xia et al., 2022), generates pseudo labels that can distinguish between salient and non-salient frames to guide the frame saliency learning.

Thus, to the end of our knowledge, there are no efficient video classification techniques that take into account the main features of the FER task, namely, (1) the potentially rapid evolution of emotions and the need to process relatively short videos; (2) presence of face detection/tracking step that limits the widely-used preprocessing of all frames via small neural nets; (3) small training sets with dirty and ambiguous labeling that limits the potential of deep models, especially in cross-domain scenarios, and forces the usage of lightweight models, such as ResNet-18 or EfficientNet. This paper tries to fill this gap and overcome the above-mentioned drawbacks of the known methods by using sequential analysis and the theory of multiple comparisons (Hochberg & Tamhane, 2009).

3. Proposed Approach

The task of this paper is formulated as follows. Given the input facial video $X = \{X(t), t = 1, 2, ..., T\}$ with T frames,

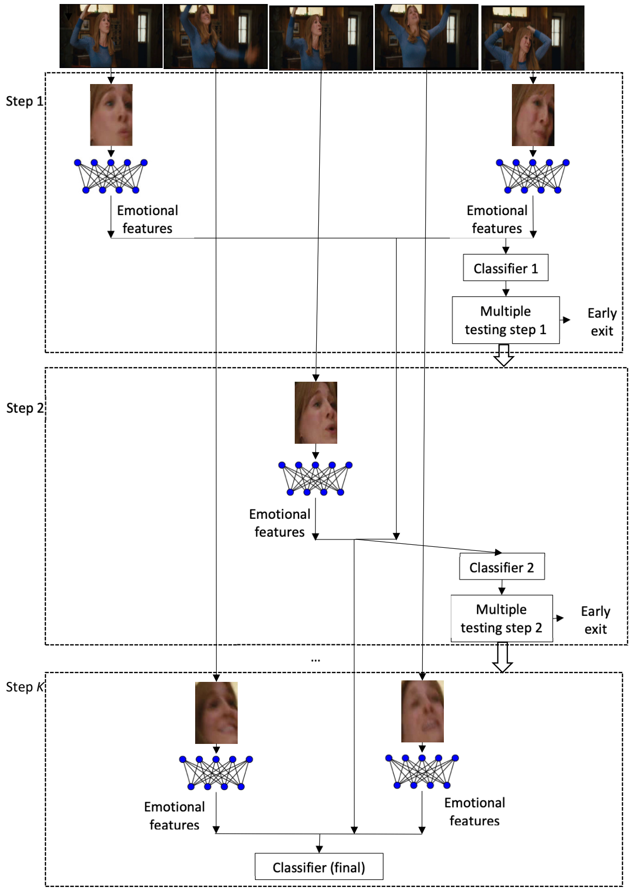

Figure 1. The overview of our approach with adaptive frame rate.

it is necessary to associate it with one of $C > 1$ emotional classes. The classes are specified by the training set of $N > 1$ facial videos $X_n = \{X_n(t), t = 1, 2, ..., T_n\}, n = 1, 2, ..., N$ with known class label $y_n \in \{1, 2, ..., C\}$ of the n-th training example, where T_n is its number of frames. For simplicity, we assume that only one subject is presented in every frame of both input and training videos.

The proposed pipeline is shown in Fig. 1. Here we consider the processing of the entire video with different frame rates inspired by sequential statistical analysis (Wald, 2013). Regarding efficient video classification techniques (Wu et al., 2019b; Ghodrati et al., 2021), the deterministic frame sampling policy function with one hyper-parameter $k > 1$ is used. Let us denote $\lceil \log_k T \rceil$ as L, where $\lceil x \rceil$ is the ceiling function. Our sequential processing will have at most L steps. The frame rate factor (concerning the original frame rate of the input video) at the l-th stage ($l = 1, 2, ..., L$) is computed as follows: $FR^{(l)} = k^{L-l}$. Thus, every $FR^{(l)}$-th frame is considered in the input video, i.e., frames with numbers from a set $T^{(l)} = \{1, FR^{(l)}, 2FR^{(l)}, ..., T\}$. For example, only two frames $\{X(1), X(T)\}$ are analyzed at the first (coarsest) step, but all frames $\{X(1), X(2), ..., X(T)\}$ are processed at the last (finest) step. Such frame rate adaptation will let us reuse the processing results from the previous step as every sequence of frames at l-th stage is a subset of frames processed at the $(l + 1)$-th stage.

The following procedure is repeated for every l-th step in our pipeline. At first, an appropriate face detector and/or tracker is used to obtain the facial image from every new frame $t \in (T^{(l)} - T^{(l-1)})$, where $T^{(0)} = \emptyset$ is an empty set. Thus, $\frac{T}{k^{L-l}} - \frac{T}{k^{L-(l-1)}} = \frac{T(k-1)}{k^{L-l+1}}$ new frames should be processed.

Secondly, this image is fed into a neural network to extract the D-dimensional feature vector (embeddings) $\mathbf{x}(t)$. In video-based FER, it is typical to pre-train the neural feature extractor on face identification and/or emotion classification (Savchenko, 2021a) on large datasets of photos, such as AffectNet (Mollahosseini et al., 2017).

Thirdly, the features of all frames from $T^{(l)}$ (including embeddings computed at previous steps) are aggregated into a single representation of the entire video

$$\mathbf{x}^{(l)} = Pool(\{\mathbf{x}(t) | t \in T^{(l)}\}). \qquad (1)$$

One can employ any simple strategies, such as statistical functions: mean (AvgPool), max (MaxPool), min, standard deviation or their concatenation (Savchenko, 2021a) or more complicated techniques, e.g., LSTMs or FAN (Meng et al., 2019).

Fourthly, descriptor $\mathbf{x}^{(l)}$ is classified. During the training procedure, faces are detected in all frames of every example X_n, and their embeddings $\mathbf{x}_n(t)$ are obtained with the same neural network-based feature extractor. These embeddings are aggregated in the same way as described above into $\mathbf{x}_n^{(l)}$ descriptors for chosen frame rate factors $FR^{(l)}, l = 1, 2, ..., L$, and an arbitrary classifier $\mathcal{C}^{(l)}$, such as MLP, random forest, gradient boosting or SVM (support vector machine), is trained using the set of pairs $\{(\mathbf{x}_n^{(l)}, y_n)\}$.

By feeding $\mathbf{x}^{(l)}$ into classifier $\mathcal{C}^{(l)}$, we obtain predicted class label $\hat{y}^{(l)}$ and its indicator of reliability (confidence score). Without a lack of generality, let us assume that each classifier represents a decision function $\mathbf{s}^{(l)}(\mathbf{x}^{(l)}) = [s_1^{(l)}(\mathbf{x}^{(l)}), ..., s_C^{(l)}(\mathbf{x}^{(l)})]$, where $s_y^{(l)}(\mathbf{x}^{(l)}) \geq 0$ is a confidence score of the classifier, such as the estimate of class posterior probability at the output of MLP or the signed distance of $\mathbf{x}$ to the separating hyperplane for SVM. The decision is made in favor of the class with the maximal confidence

$$\hat{y}^{(l)} = \underset{y \in \{1, ..., C\}}{\text{argmax}} \, s_y^{(l)}(\mathbf{x}^{(l)}). \qquad (2)$$

Fifthly, the inference will be terminated, and the classifier output $\hat{y}^{(l)}$ will be returned if l is equal to L or decision (2) is rather reliable:

$$s^{(l)}_{\hat{y}^{(l)}}(\mathbf{x}^{(l)}) > s^{(l)}_{\hat{y}^{(l)}}. \quad (3)$$

Otherwise, the frame rate is adjusted ($FR^{(l+1)} = FR^{(l)}/k$), and all five above-mentioned processing parts are repeated for the $(l+1)$-th step.

The best estimate of threshold $s^{(l)}_y$ for y-th classifier output at level l strongly depends on the chosen type of classifier. For example, if the features are matched with the Kullback-Leibler (KL) divergence, which has a chi-squared distribution with D degrees of freedom in asymptotic (Kullback, 1997), then $s^{(l)}_y$ is proportional to the α_l-quantile of the non-central chi-squared distribution (Savchenko, 2020), where $\alpha^{(l)}$ is the fixed FAR.

Suppose there is no theoretical knowledge about the distribution of the decision function of a classifier. We propose to compute these thresholds using a random subset of the training set for every class y to train a classifier. Next, the confidence scores are obtained for the remaining M examples $\{\mathbf{x}^{(l)}_{n_1}, ..., \mathbf{x}^{(l)}_{n_M} | y_{n_m} = y\}$. Threshold $s^{(l)}_y$ is chosen as the α_l-quantile of the maximal scores of other classes $\left\{ \max\limits_{c \neq y} s^{(l)}_c(\mathbf{x}^{(l)}_{n_m}) \middle| m \in \{1, ..., M\} \right\}$. The training set is split into two equal parts in all experiments of this paper. Moreover, we use the same threshold $s^{(l)}_y = s^{(l)}$ for each class, i.e., M is equal to $N/2$. As a result, the threshold can be estimated more accurately due to the larger size of the available maximal scores $\left\{ \max\limits_{c \neq y_{n_m}} s^{(l)}_c(\mathbf{x}^{(l)}_{n_m}) \middle| m \in \{1, ..., M\} \right\}$.

An essential question in this paper is how to choose concrete values $\alpha_1, ..., \alpha_L$ if a confidence level α for the whole image recognition procedure is specified? Let us consider the task regarding statistical hypothesis testing (Belova & Savchenko, 2015). At each step l, there is a null hypothesis H_l that the decision $\hat{y}^{(l)}$ (2) is correct, and a decision boundary for this hypothesis is specified by inequality (3).

It is an example of multiple hypothesis tests (Hochberg & Tamhane, 2009). In this theory, it is typical to control the false discovery rate, i.e., the expected ratio of false positives to the total number of positive classifications. Such correction usually requires sorting the p-values of all L hypotheses. In this paper, it is assumed that the reliability increases with the availability of additional information about input video, i.e., with an increase in the frame rate. Thus, a typical solution would be the Benjamini-Hochberg test (Benjamini & Hochberg, 1995; Savchenko, 2021b):

$$\alpha_l = \frac{\alpha \cdot l}{L}. \quad (4)$$

The proposed training procedure for our inference pipeline (Fig. 1) is summarized in Algorithm 1. It is important to emphasize that here, in contrast to existing works (Ghodrati et al., 2021; Wu et al., 2019b; Yeung et al., 2016), the same classifier $\mathcal{C}$ can be used for all different frame rate factors $\{FR^{(l)}\}$, especially if simple statistical functions are applied to estimate the video descriptor.

Algorithm 1 Proposed Training Procedure

```
for each training example $n \in \{1, ..., N\}$ do
  for each frame $t \in \{1, ..., T_n\}$ do
    Extract facial region in $X_n(t)$ using an arbitrary face detector
    Feed the facial image into a neural network feature extractor and compute the embeddings $\mathbf{x}_n(t)$
  end for
  Compute video descriptor $\mathbf{x}_n = Pool(\{\mathbf{x}_n(t) | t \in \{1, 2, ..., T_n\}\})$
  for each step of adjusted frame rate $l \in \{1, ..., L-1\}$ do
    Compute $\mathbf{x}^{(l)}_n = Pool(\{\mathbf{x}_n(t) | t \in T^{(l)}\})$ (1)
  end for
end for
for each step of adjusted frame rate $l \in \{1, ..., L-1\}$ do
  Split $N$ instances in a stratified fashion to get indices $\{n_1, ..., n_M\}$ of validation set
  Train the $l$-th classifier $\mathcal{C}$ using remaining training examples
  Initialize a list $S = []$
  for each validation instance $m \in \{1, ..., M\}$ do
    Append the maximal inter-class confidence score $\max\limits_{y \neq y(n)} s^{(l)}_y(\mathbf{x}^{(l)}_{n_m})$ to $S$
  end for
  Assign the $\lfloor \alpha l / L \rfloor$-th largest element from $S$ to the threshold $s^{(l)}$ using the Benjamini-Hochberg correction (4)
end for
Train an arbitrary classifier $\mathcal{C}$ using set of pairs $\{(\mathbf{x}_n, y_n)\}$.
return classifier $\mathcal{C}$ and thresholds $s^{(l)}, l = 1, 2, ..., L$
```

If classifier $\mathcal{C}$ and feature pooling $Pool$ (1) are computationally cheap, the number of processed frames mainly defines the run-time complexity of the Algorithm (Fig. 1). If the decision is made after the first step for the frame rate factor $FR^{(1)}$, one will need to perform inference only twice (for the first and the last frames). In the worst case, all T frames should be analyzed. If we assume that every step has the same exit probability of $1/L$, the average complexity will be estimated as follows:

$$\frac{1}{L}\sum_{l=1}^{L}(1 + k^{l-1}) \approx \frac{T}{L}$$

Table 1. Mean F1-score and average relative inference time per one frame $\bar{t}$ (ms), top participants of the ABAW-3 challenge.

Team	Model	Val F1-score	Test F1-score	Time $\bar{t}$
Netease Fuxi Virtual	InceptionResNet	0.295	0.2846	196.85 ±0.35
Human (Zhang et al., 2022)	Ensemble	0.394	0.359	983.20 ±0.54
IXLAB	DAN (ResNet50)	0.317	0.3064	89.79 ±0.19
(Jeong et al., 2022)	Ensemble	0.346	0.3377	264.26 ±0.27
AlphaAff	Swin-S	0.4378	0.3138	511.85 ±0.41
(Xue et al., 2022)	Ensemble	0.4615	0.359	2078.20 ±1.10
HSE-NN (Savchenko, 2022)	EfficientNet-B0	0.4018	0.3025	55.94 ±0.25
PRL (Phan et al., 2022)	RegNetY	0.3035	0.286	246.65 ±0.35
Baseline (Kollias, 2022)	VGG16	0.23	0.205	160.54 ±0.48

In practice, the gain in performance strongly depends on the relative number of exits at each level, so one can expect that the first step is enough for most easy input videos. However, it is still possible that all frames should be processed for complex examples.

4. Experimental Study

In this section, performance of the proposed approach (Fig. 1) is compared with known competitors (Section 2) and the conventional classification of videos with a fixed frame rate. We analyze two FER datasets from the third ABAW 2020 (Kollias, 2022) and EmotiW 2019 (Dhall, 2019) challenges. The average relative inference time per one frame $\bar{t}$ is measured on the CPU of MSI GP63 8RE laptop (Intel Core i7-8750H 2.2 GHz, 16 Gb RAM). The faces in all video frames were preliminary extracted by the MTCNN detector. Still, we do not report the face detection time in $\bar{t}$ because there are a lot of fast detectors with high quality, and we do not have the goal of choosing the best one. For instance, face detection using the MediaPipe library requires approximately 7 ms per frame on our MSI laptop.

4.1. ABAW Challenge

In this subsection, we describe the uni-task frame-level FER task results with eight emotional labels (anger, disgust, fear, happiness, sadness, surprise, neutral and other) from the third ABAW CVPR 2022 Workshop and Competition. The training and validation sets provided by organizers contain 585,317 and 280,532 frames, respectively. The macro-averaged F1 score P_{EXPR} (Kollias, 2022) computed on official validation and test sets, and classification time $\bar{t}$ of the top-performers of this challenge are shown in Table 1. As this paper mainly focuses on efficient video processing, we chose the family of EmotiEffNet (EfficientNets) (Savchenko, 2022; 2021a) for further experiments. Indeed, they provide one of the most excellent accuracies among existing single models with a reasonable inference time. Moreover, they did not require fine-tuning on the training set of the ABAW challenge. Only a new classifier $\mathcal{C}$ (MLP with one hidden layer) should be trained on top of the features extracted by a pre-trained network. Finally, EmotiEffNet-B2 is currently the state-of-the-art model for one of the primary FER datasets for static photos, AffectNet (Mollahosseini et al., 2017). In the remaining part of this paper, we used the pre-trained models made publicly available by its authors.

Table 2. Validation F1-score and average relative inference time per one frame $\bar{t}$ (ms) of efficient video classification methods, ABAW-3 challenge, EmotiEffNet-B0.

Method	F1-score	Time $\bar{t}$
Smoothing (all frames)	**0.4262**	55.94±0.25
AdaFrame	0.4205	42.32±0.30
LiteEval	0.4220	50.71±0.26
AR-Net	0.4051	22.39±0.25
OCSampler	0.3928	4.85±0.22
FrameExit	0.4177	5.97±0.37
Proposed approach	0.4217	**3.70±0.20**

In Table 2, we present the results of several fast video classification methods. Most of them, e.g., AdaFrame (Wu et al., 2019b), LiteEval (Wu et al., 2019a), and OCSampler (Lin et al., 2022) need a fast feature extractor for the frame selection. Hence, the MobileNet v1 (Savchenko, 2021a) trained similarly to EmotiEffNets is used as a lightweight neural network here. Other hyper-parameters of these methods were chosen to get the lowest running time, but, if possible, the F1-score should not be lower than 1% lower than the best F1-score for processing of all frames. In all cases, the fragments of videos with $T = 201$ frames are considered to decide on the facial expression of the middle frame. The baseline is a simple smoothing of all 200 predictions at the output of MLP classifier (Savchenko, 2022). In our method, $L = 5$ steps were chosen with frame rate factors 200, 100, 50, 10, and 1. The constant ratio of sequential factors is not required, but each next $FR^{(l+1)}$ should be a divider of $FR^{(l+1)}$ to use frame embeddings computed at previous steps. The FAR α in Algorithm 1 equals 0.2.

Table 3. Validation F1-score and average relative inference time per one frame $\bar{t}$ (ms) for fixed FAR and proposed multiple testing correction, ABAW-3 challenge, EmotiEffNet-B0.

Sequence of frame rates	Thresholds estimator	F1-score	Time $\bar{t}$
(200–>100–>50–>10–>1)	Fixed FAR	0.4190	20.15±0.35
	Proposed	0.4217	3.70 ±0.20
(100–>50–>10–>1)	Fixed FAR	0.4205	23.82±0.29
	Proposed	0.4221	11.03 ±0.32
(50–>25–>1)	Fixed FAR	0.4257	26.58±0.31
	Proposed	0.4253	17.12 ±0.23
(50–>10–>1)	Fixed FAR	0.4258	25.03±0.30
	Proposed	0.4258	15.51 ±0.19
(200–>50–>1)	Fixed FAR	0.4203	29.39±0.28
	Proposed	0.4207	20.41 ±0.20
(100–>50–>1)	Fixed FAR	0.4225	27.26±0.25
	Proposed	0.4230	20.31 ±0.21

As one can notice, the proposed algorithm is 15 times faster than smoothing all frames, though the drop in F1-score is less than 0.5%. If the time for face detection is taken into account, the gain in performance will be even more noticeable. Moreover, it is the most efficient video classification method. For example, our approach is 60% faster than FrameExit, while the F1-score of the latter is 0.4% lower. Only LiteEval is slightly (0.03%) more accurate, but its running time is too high due to the usage of the lightweight MobileNet model, which is only twice faster than EfficientNet-B0.

Let us provide the ablation study results for our method. In Table 3, we compare the proposed multiple testing correction with the choice of thresholds $s^{(l)}$ (3) by using the same α_l for all steps $l = 1, 2, ..., L$. The multiple comparisons are worth using only if $L \geq 3$. Indeed, only one threshold should be estimated in our training Algorithm 1 if $L = 2$. As one can notice, the Benjamini-Hochberg correction leads to much better performance. The more the number of steps, the greater the gain in the running time of the proposed technique with conventional estimation of thresholds.

Finally, the results of several convolutional neural networks from a family of EmotiEffNets, namely, EmotiEffNet-B0 from the previous experiments, its multi-task version MT-EmotiEffNet-B0 (Savchenko, 2023) and deeper EmotiEffNet-B2 (Savchenko et al., 2022), are shown in Table 4. Though EmotiEffNet-B0 is the best model in this competition, it is essential to emphasize that our method works with an arbitrary feature extractor without the need to re-train it with our model. Moreover, our speed-up over processing of all frames is even more significant for two other neural networks: up to 40 and 25 times for MT-EmotiEffNet and EmotiEffNet-B2, respectively.

4.2. AFEW from EmotiW Challenge

This subsection provides the experimental results for the AFEW dataset from EmotiW 2019 audio-visual emotion recognition challenge (Dhall, 2019). It contains 773 train and 383 validation short clips (1-5 seconds) with known emotional labels (Anger, Disgust, Fear, Happiness, Sad and Surprise, and Neutral) for each clip. Only video modality is considered.

We reproduced the FER pipeline for EmotiEffNets from the original paper (Savchenko et al., 2022), namely, concatenation of the point-wise mean, max, min, and standard deviation in the feature pooling $Pool$ (1) for facial features extracted from each facial descriptors for a given frame rate. The classifier $\mathcal{C}$ is the LinearSVC with regularization parameter found using cross-validation on the training set.

Table 5 contains the results for various feature extractors and frame rate factors, while the classification accuracy and average inference time $\bar{t}$ are presented in Table 6. Here we presented several known single models, namely, FAN (Meng et al., 2019), DenseNet-161 (Liu et al., 2018), the best single model (IR-50) and ensemble (factorized bilinear pooling, FBP) from the paper (Zhou et al., 2019), the best single model (VGG-Face + BLSTM) of a winner of EmotiW-2019 (Li et al., 2019) and the noisy student (ResNet-18) with iterative training (Kumar et al., 2020). The efficient video classification techniques use the same features (pre-trained EmotiEffNet-B0 and EmotiMobileNet) described in the Subsection 4.1.

In our pipeline (Fig. 1), we started with frame rate factor $FR^{(1)} = 18$ as higher values are recognized with too many mistakes. As a result, it is only twice faster than the classification of all frames for each feature extractor. Though the running time of the proposed method is approximately equal to $\bar{t}$ for the best techniques (ARNet, OCSampler, and FrameExit), our approach is much more accurate.

5. Conclusion

This paper presents the novel framework (Fig. 1) that implements efficient video-based FER using sequential analysis of various frames. Its most remarkable feature is the multiple testing correction (4) that makes it possible to automatically reach a balance between efficiency and accuracy (Table 3). Our method focuses on the most critical aspects of affective behavior analysis (Makarov et al., 2016), namely, dirty video datasets that limit the usage of complex models in cross-dataset settings, short video sequences (low number of frames with the same class label), and the need to perform face detection before actual classification. Indeed, only $L - 1$ parameters (thresholds) should be estimated in the proposed method, given the available training set. The deterministic refinement of frame rates lets us obtain more

Table 4. Validation F1-score and average relative inference time per one frame $\bar{t}$ (ms) for various neural networks and sequences of frame rate factors, ABAW-3 challenge.

Sequence of	EmotiEffNet-B0		MT-EmotiEffNet-B0		EmotiEffNet-B2	
frame rates	F1-score	Time $\bar{t}$	F1-score	Time $\bar{t}$	F1-score	Time $\bar{t}$
(200)	0.3624	0.55±0.05	0.3323	0.56±0.04	0.3062	1.15 ±0.12
(1)	0.4262	55.94±0.25	0.3913	56.68±0.25	0.3532	116.04 ±0.30
(200–>100–>50–>10–>1)	0.4217	**3.70±0.20**	0.3820	**1.34±0.08**	0.3503	**4.63 ±0.13**
(50–>25–>1)	0.4253	17.12±0.23	0.3861	1.81±0.06	0.3518	19.09 ±0.19
(50–>10–>1)	0.4258	15.51±0.19	**0.3898**	3.02±0.12	0.3521	14.58 ±0.13
(200–>50–>1)	0.4207	20.41±0.20	0.3771	1.15±0.09	0.3488	24.82 ±0.22
(100–>50–>1)	0.4230	20.31±0.21	0.3787	1.07±0.05	0.3503	24.57 ±0.23
(200–>1)	0.4205	48.27±0.37	0.3832	31.37±0.28	0.3477	74.01 ±0.27
(100–>1)	0.4228	36.48±0.19	0.3840	14.43±0.07	0.3505	47.49 ±0.18
(50–>1)	**0.4258**	33.01±0.21	0.3885	12.65±0.09	**0.3528**	43.73 ±0.14

Table 5. Validation accuracy and average relative inference time per one frame $\bar{t}$ (ms) for various neural networks and sequences of frame rate factors, AFEW dataset.

Sequence of	EmotiEffNet-B0		MT-EmotiEffNet-B0		EmotiEffNet-B2	
frame rates	Accuracy	Time $\bar{t}$	Accuracy	Time $\bar{t}$	Accuracy	Time $\bar{t}$
(18)	0.5085	3.60±0.03	0.5013	3.65±0.03	0.5040	7.48 ±0.06
(1)	0.5927	55.94±0.19	0.5699	56.68±0.20	0.5937	116.04 ±0.29
(18–>9–>1)	0.5850	**29.75±0.15**	0.5515	**27.55±0.14**	0.5778	**53.74 ±0.21**
(18–>6–>1)	**0.5927**	32.79±0.17	0.5515	30.09±0.15	0.5831	54.00 ±0.20
(9–>3–>1)	0.5903	38.70±0.18	0.5831	37.41±0.17	0.5989	73.03 ±0.23
(6–>3–>1)	0.5903	40.01±0.17	0.5726	38.93±0.17	0.5937	76.06 ±0.22
(18–>1)	0.5824	31.02±0.17	0.5541	30.30±0.16	0.5778	58.00 ±0.20
(9–>1)	0.5903	34.53±0.17	**0.5752**	33.01±0.17	0.5910	63.15 ±0.23
(6–>1)	0.5877	34.31±0.16	0.5726	34.04±0.16	0.5910	61.30 ±0.22
(3–>1)	0.5903	40.38±0.19	0.5726	39.45±0.19	**0.5937**	72.75 ±0.28

accurate results when compared to traditional complex techniques based on RL (Kim et al., 2021; Wu et al., 2019b). Secondly, our approach can be used with an arbitrary frame pooling (aggregator) and facial feature extractor, including lightweight architectures (Savchenko, 2021a). Hence, there is no need to train additional lightweight models for fast frame selection (Raviv et al., 2022; Wang et al., 2021; Wu et al., 2019a).

It is important to emphasize that applying our approach to other video-based object classification problems is possible. Indeed, we can improve performance when compared to existing fast video classification techniques if at least one of the following conditions holds: (1) an object should be preliminary detected and tracked with the computationally expensive method before classification of objects attributes (e.g., faces are detected and recognized, cars are detected and mark/model is classified, the license plate is detected and characters are recognized, etc.); (2) the video for an observed object is short or object attributes can be rapidly changed (our primary task of emotional intelligence); and (3) the training set is too small to train complex RL-based policies for skipping frames and the domain significantly differs with other domains, so the transfer learning/domain adaptation of RL-based techniques is impossible. One example is face recognition in video surveillance systems or facial recognition payment systems. Each subject is in front of a camera for 1 to 5 seconds, and processing on embedded devices is desirable. Though many facial quality assessment tools exist to choose the best frames, our method can be an intense but straightforward competitor. Another example application is video-based traffic analysis, searching for traffic violations, estimating a vehicle's speed, etc.

The main disadvantage of our method is the need to know the number of frames T to predict facial expression in the whole video fragment. For example, only every 200-th frame was analyzed at the coarsest level in the ABAW challenge (Subsection 4.1), but $FR^{(l)}$ cannot be greater than 20 (18 in our experiments) for the AFEW dataset (Subsection 4.1). In the future, it is essential to extend the proposed approach for online decision-making with automatic (maybe, not very accurate) detection of change points in emotions and an adaptive choice of the first frame rate factor $FR^{(l)}$. For example, it will be meaningful to include a scalable approach, detecting positive/negative changes of

Table 6. Validation accuracy and average relative inference time per one frame $\bar{t}$ (ms), AFEW dataset.

METHOD	F1-SCORE	TIME $\bar{t}$
FAN (RESNET-18)	0.5118	35.18±0.08
DENSENET-161	0.5144	170.61±0.31
IR-50	0.5378	92.64±0.24
VGG-FACE + BLSTM	0.5391	165.90±0.45
NOISY STUDENT	0.5517	29.26±0.06
FBP FUSION	0.6550	232.02±0.33
EmotiEffNet-B0		
ALL FRAMES	0.5927	55.94±0.19
ADAFRAME	0.5906	49.95±0.25
LITEEVAL	0.5927	52.20±0.31
AR-NET	0.5526	32.43±0.23
OCSAMPLER	0.5530	30.27±0.18
FRAMEEXIT	0.5726	31.89±0.34
PROPOSED APPROACH	0.5910	29.75±0.15

valence/arousal and then going at the finer detail of facial expressions. However, the main challenge should be resolved: estimating the evolution of emotional state without significant time delay.

Acknowledgements

This work was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated November 2, 2021 No. 70-2021-00142.

References

Belova, N. S. and Savchenko, A. V. Statistical testing of segment homogeneity in classification of piecewise–regular objects. *International Journal of Applied Mathematics and Computer Science*, 25(4):915–925, 2015.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Dhall, A. EmotiW 2019: Automatic emotion, engagement and cohesion prediction tasks. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, pp. 546–550. ACM, 2019.

Dutson, M., Li, Y., and Gupta, M. Event neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 276–293. Springer, 2022.

Fan, H., Xu, Z., Zhu, L., Yan, C., Ge, J., and Yang, Y. Watching a small portion could be as good as watching all: Towards efficient video classification. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

Gao, R., Oh, T.-H., Grauman, K., and Torresani, L. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10457–10467, 2020.

Ghodrati, A., Bejnordi, B. E., and Habibian, A. Frame-Exit: Conditional early exiting for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15608–15618, 2021.

Gowda, S. N., Rohrbach, M., and Sevilla-Lara, L. SMART frame selection for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1451–1459, 2021.

Grachev, A. M., Ignatov, D. I., and Savchenko, A. V. Neural networks compression for language modeling. In *Proceedings of International Conference on Pattern Recognition and Machine Intelligence (PReMI)*, pp. 351–357. Springer, 2017.

Hochberg, J. and Tamhane, A. *Multiple comparison procedures*. Wiley, 2009.

Jeong, J.-Y., Hong, Y.-G., Kim, D., Jeong, J.-W., Jung, Y., and Kim, S.-H. Classification of facial expression in-the-wild based on ensemble of multi-head cross attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2353–2358, 2022.

Jillela, R. R. and Ross, A. Adaptive frame selection for improved face recognition in low-resolution videos. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, pp. 1439–1445. IEEE, 2009.

Karpov, A. and Makarov, I. Exploring efficiency of vision transformers for self-supervised monocular depth estimation. In *Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 711–719. IEEE, 2022.

Khokhlova, Y. I. and Savchenko, A. About neural-network algorithms application in viseme classification problem with face video in audiovisual speech recognition systems. *Optical Memory and Neural Networks*, 23(1):34–42, 2014.

Kim, H., Jain, M., Lee, J.-T., Yun, S., and Porikli, F. Efficient action recognition via dynamic knowledge propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13719–13728, 2021.

Kollias, D. ABAW: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2328–2336, 2022.

Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., and Zafeiriou, S. Deep affect prediction in-the-wild: Aff-Wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6):907–929, 2019.

Korbar, B., Tran, D., and Torresani, L. SCSampler: Sampling salient clips from video for efficient action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6232–6242, 2019.

Kullback, S. *Information Theory and Statistics*. Dover Publications, Mineola, New York, 1997.

Kumar, V., Rao, S., and Yu, L. Noisy student training using body language dataset improves facial expression recognition. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 756–773. Springer, 2020.

Li, S., Zheng, W., Zong, Y., Lu, C., Tang, C., Jiang, X., Liu, J., and Xia, W. Bi-modality fusion for emotion recognition in the wild. In *Proceedings of International Conference on Multimodal Interaction (ICMI)*, pp. 589–594. ACM, 2019.

Lim, J., Baek, Y., and Chae, B. Temporal early exiting with confidence calibration for driver identification based on driving sensing data. *IEEE Access*, 10:132095–132107, 2022.

Lin, J., Duan, H., Chen, K., Lin, D., and Wang, L. OCSampler: Compressing videos to one clip with single-step sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13894–13903, 2022.

Liu, C., Tang, T., Lv, K., and Wang, M. Multi-feature based emotion recognition for video clips. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI)*, pp. 630–634, 2018.

Makarov, I., Zyuzin, P., Polyakov, P., Tokmakov, M., Gerasimova, O., Guschenko-Cheverda, I., and Uriev, M. Modelling human-like behavior through reward-based approach in a first-person shooter game. In *Proceedings of the 3rd Workshop on Experimental Economics and Machine Learning (EEML'16)*, pp. 24–33. CEUR Workshop Proceedings, 2016.

Meng, D., Peng, X., Wang, K., and Qiao, Y. Frame attention networks for facial expression recognition in videos. In *Proceedings of the International Conference on Image Processing (ICIP)*, pp. 3866–3870. IEEE, 2019.

Meng, Y., Lin, C.-C., Panda, R., Sattigeri, P., Karlinsky, L., Oliva, A., Saenko, K., and Feris, R. AR-Net: Adaptive frame resolution for efficient action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 86–104. Springer, 2020.

Mollahosseini, A., Hasani, B., and Mahoor, M. H. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

Phan, K. N., Nguyen, H.-H., Huynh, V.-T., and Kim, S.-H. Facial expression classification using fusion of deep neural network in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2507–2511, 2022.

Raviv, A., Dinai, Y., Drozdov, I., Zehngut, N., Goldin, I., and Center, S. I. R. D-step: Dynamic spatio-temporal pruning. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2022.

Ryumina, E., Dresvyanskiy, D., and Karpov, A. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing*, 514: 435–450, 2022.

Savchenko, A. V. Sequential analysis with specified confidence level and adaptive convolutional neural networks in image recognition. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Savchenko, A. V. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *Proceedings of International Symposium on Intelligent Systems and Informatics (SISY)*, pp. 119–124. IEEE, 2021a.

Savchenko, A. V. Fast inference in convolutional neural networks based on sequential three-way decisions. *Information Sciences*, 560:370–385, 2021b.

Savchenko, A. V. Video-based frame-level facial analysis of affective behavior on mobile devices using EfficientNets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2359–2366, 2022.

Savchenko, A. V. MT-EmotiEffNet for multi-task human affective behavior analysis and learning from synthetic data. In *Proceedings of the European Conference on Computer Vision (ECCV 2022) Workshops*, pp. 45–59. Springer, 2023.

Savchenko, A. V., Savchenko, L. V., and Makarov, I. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13(4):2132–2143, 2022.

Sun, X., Panda, R., Chen, C.-F. R., Oliva, A., Feris, R., and Saenko, K. Dynamic network quantization for efficient video inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7375–7385, 2021.

Tao, H. and Duan, Q. An adaptive frame selection network with enhanced dilated convolution for video smoke recognition. *Expert Systems with Applications*, 215:119371, 2023.

Teerapittayanon, S., McDanel, B., and Kung, H. BranchyNet: Fast inference via early exiting from deep neural networks. In *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR)*, pp. 2464–2469. IEEE, 2016.

Wald, A. *Sequential Analysis*. Dover Publications, New York, 2013. ISBN 9780486615790.

Wang, Y., Chen, Z., Jiang, H., Song, S., Han, Y., and Huang, G. Adaptive focus for efficient video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16249–16258, 2021.

Wang, Y., Sun, Y., Huang, Y., Liu, Z., Gao, S., Zhang, W., Ge, W., and Zhang, W. FERV39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20922–20931, 2022.

Wu, Z., Xiong, C., Jiang, Y.-G., and Davis, L. S. LiteEval: A coarse-to-fine framework for resource efficient video recognition. *Advances in Neural Information Processing Systems*, 32, 2019a.

Wu, Z., Xiong, C., Ma, C.-Y., Socher, R., and Davis, L. S. AdaFrame: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1278–1287, 2019b.

Xia, B., Wu, W., Wang, H., Su, R., He, D., Yang, H., Fan, X., and Ouyang, W. NSNet: Non-saliency suppression sampler for efficient video recognition. In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 705–723. Springer, 2022.

Xue, F., Tan, Z., Zhu, Y., Ma, Z., and Guo, G. Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2412–2418, 2022.

Yeung, S., Russakovsky, O., Mori, G., and Fei-Fei, L. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2678–2687, 2016.

Zhang, W., Qiu, F., Wang, S., Zeng, H., Zhang, Z., An, R., Ma, B., and Ding, Y. Transformer-based multimodal information fusion for facial expression analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2428–2437, 2022.

Zhang, W., Ma, B., Qiu, F., and Ding, Y. Facial affective analysis based on mae and multi-modal information for 5th ABAW competition. *arXiv preprint arXiv:2303.10849*, 2023.

Zhou, H., Meng, D., Zhang, Y., Peng, X., Du, J., Wang, K., and Qiao, Y. Exploring emotion features and fusion strategies for audio-video emotion recognition. In *Proceedings of International Conference on Multimodal Interaction (ICMI)*, pp. 562–566. ACM, 2019.