Mutual Information Regularized Identity-aware Facial Expression Recognition in Compressed Video

Xiaofeng Liu[a], Linghao Jin[a,b], Xu Han[a,b] and Jane You[c]

[a]Harvard University, Cambridge, MA, USA
[b]John Hopkins University, Baltimore, MD, USA
[c]Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Keywords:
Facial Expression Recognition
Mutual Information
Disentangled Representation
Compressed Video

ABSTRACT

This paper targets to explore the inter-subject variations eliminated facial expression representation in the compressed video domain. Most of the previous methods process the RGB images of a sequence, while the off-the-shelf and valuable expression-related muscle movement already embedded in the compression format. In the up to two orders of magnitude compressed domain, we can explicitly infer the expression from the residual frames and possible to extract identity factors from the I frame with a pre-trained face recognition network. By enforcing the marginal independent of them, the expression feature is expected to be purer for the expression and be robust to identity shifts. Specifically, we propose a novel collaborative min-min game for mutual information (MI) minimization in latent space. We do not need the identity label or multiple expression samples from the same person for identity elimination. Moreover, when the apex frame is annotated in the dataset, the complementary constraint can be further added to regularize the feature-level game. In testing, only the compressed residual frames are required to achieve expression prediction. Our solution can achieve comparable or better performance than the recent decoded image based methods on the typical FER benchmarks with about 3 times faster inference.

1. Introduction

The video modality is increasingly important in many computer vision applications [6, 48]. Considering the natural dynamic property of the human face expression [30, 47], many works propose to explore spatio-temporal features of facial expression recognition (FER) from the videos. Recently, the deep neural networks have achieved significant progress for the image-based facial expression recognition [49, 82, 84], while the processing of expression video is still challenging.

Although the multi-frame sequence can inherit richer information and the temporal-correlation between the consecutive frames can usually be helpful for the facial expression recognition, the video also introduced significantly redundancy. Considering the subtle muscle movement in FER videos, the signal-to-noise ratio (SNR) is extremely low [42, 49].

The conventional spatio-temporal FER networks take sequentially frames to utilize the spatial and temporal cues to represent the facial expressions [42]. The recursive neural network (RNN), 3D convolutional, and non-local network can be used to derive information from sequences [36, 74, 59, 55]. However, using deep neural networks to process many consecutive frames can be very computationally expensive when the length of video is long [51, 49]. Actually, many methods can achieve good performance using decision-level fusion of image-based FER, which totally ignores the temporal dependency. This implies the temporal cues may hard be explored in the low SNR FER videos.

We propose that the compressed domain can be suitable

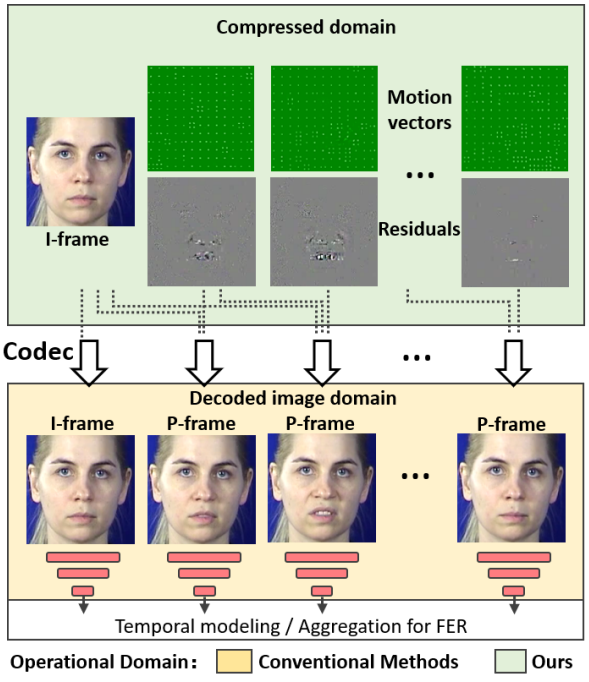

Figure 1: Illustration of the typical video compression and the scheme of conventional FER methods, which first decode the video and then feed it into a FER network.

*The funding support from Youth Innovation Promotion Association, CAS (2017264), PolyU Central Research Grant G-YBJW, and Hong Kong Government General Research Fund GRF (Ref. No.152202/14E) are greatly appreciated.

liuxiaofengcmu@gmail.com (X. Liu); ljin23@jhu.edu (L. Jin); xhan32@jhu.edu (X. Han); csyjia@comp.polyu.edu.hk (J. You)
https://liu-xiaofeng.github.io/ (X. Liu); https://www.comp.polyu.edu.hk/en-us/staffs/detail/1261 (J. You)
ORCID(s): 0000-0002-4514-2016 (X. Liu)

for FER task for four reasons. **1)** the consecutive frames in video modality have many uninformative and repeating patterns, which may drowning the "interesting" and "true" signal [86]. With the standard video compression algorithms, the compression ratio can usually be hundred times [65]. Manipulating on the compressed domain can significantly reduce the cost of computation and memory. **2)** the typical compression solution (e.g., MPEG-4, H.264, and HEVC) separate the video to the I frame (intracoded frames) with the first image, and follows several P frames (predictive frames) which encoded as the "change" or "movement" [40], as shown in Figure. 1. The movement of face muscle is the fundamental of expression. Actually, many FER system are based on the action unit system [53, 30]. Therefore, the compressed P frames can inherit the off-the-shelf yet valuable expression-related information, and its pattern is much simpler than the raw images. **3)** our compressed domain exploration can also be effective since it focus on the "true" signals rather than processing the repeatedly near-duplicates [80, 70]. **4)** we do not need the decode operation in real-world task, since the to-be processed data is usually transferred with the compressed format.

Moreover, the high inter-subject variations caused by identity differences in facial attributes is a long-lasting difficulty in FER community [66, 56, 51]. The learned features may capture more identity related information rather than expression related information, and are not purely related to the FER task. We note that the P frames can also incorporate the relative position of face key points, which are also related to the identity [15, 73]. The typical solution for eliminating identity from the expression representation is using the metric learning [56, 51, 14, 49]. Inspired by the adversarial disentanglement, [4, 13] propose to render the identity removed face with conditional generative adversarial network (GAN). However, these works focus on the image-based FER. [47] extend the metric learning protocol [51] to video data by directly replace the image to the video feature, which does not consider the characteristic of video. Besides, most of these work requires the identity label and multiple expressions of the same person, which significantly limits their applications on the in-the-wild FER dataset [17].

In this work, we propose to explore the identity factor from the I frame with the pre-trained face recognizer, e.g., FaceNet [68]. Their embeddings are remarkably reliable, since they achieve high accuracy over millions of identities [35], and robust to a broad range of nuisance factors such as expression, pose, illumination and occlusion variations.

Using the identity feature as the anchor, we can explicitly enforce the marginally independence of our identity and expression feature. Instead of the complicate adversarial training [54, 50] for disentanglement, we adopt the mutual information (MI) as the statistical measure of the independence of these two representations. The MI of two random variables can usually be intractable to directly and precisely measure in a high-dimensional space [43, 11]. Recently, some of the works illustrate that the mutual information can be differentialbel approximated [10, 58]. We propose to minimize the differentiable MI measure as the objective. Practically, it can be a latent space min-min game of an encoder-discriminator framework, which follows a collaborative fashion rather than adversarial competition. We note that GAN is notorious for its unstable model collapse [23], while our MI regularization is concise and efficient.

In summary, this paper makes the following contributions.

- We propose to inference expression from the residual frames, which explores the off-the-shelf yet valuable expression related muscle movement in the up to two orders of magnitude compressed domain.
- Targeting for the identity-aware video-based FER, the independence of expression and identity representations from P frames and I frame are enforced with the differentiable MI measure.
- The separability of expression and identity representations is maximized by a min-min game with the joint and marginal distribution sampling, which does not rely on the unstable adversarial game to achieve the identity elimination.

We evidenced its effectiveness on several video-based FER benchmarks with much faster inference. The promising performance evidenced its generality and scalability.

2. Related Works

Video-based FER has been extensively studied, since the facial expression is a natural and universal means for human communication [42, 31]. Considering that the expression is essentially a dynamic action which should take minute muscle movements through time into account [67]. Traditionally, the handcrafted features is utilized to represent the spatio-temporal cues and for FER. For instance the LBP-TOP [91], 3D-HOG [38], Gabor-DCT [19] and 3D-SIFT [69]. With the fast development of deep learning, both the frame aggregation and spatiotemporal FER networks are developed. The frame aggregation methods can utilize the image-based FER networks by making the decision-level [33] or the feature-level frame-wise aggregation [83]. However, the important temporal correlation is not investigated. Instead, the spatio-temporal FER networks take sequentially images to utilize both the spatial/textural and temporal information [2]. Moreover, the cascaded networks propose to combine the perceptual vision representations learned from CNNs with the RNNs for variable-length video[18, 29]. Besides, when taking the non-local network for video analysis [79, 55], the possible connections and the corresponding computation costs grows exponentially to the number of frames. The 3D convolutional has shared kernel-weights along the time axis, which has also been widely used for video-based FER [8, 92]. However, the previous works only consider the image domain, and the spatiotemporal FER does not significantly outperforms aggregation methods [42]. To the best of our knowledge, this is the first effort to investigate the compressed video FER, which is orthogonal to these advantages and can be easily added to each other.

Video compression convert the digital video into a specific

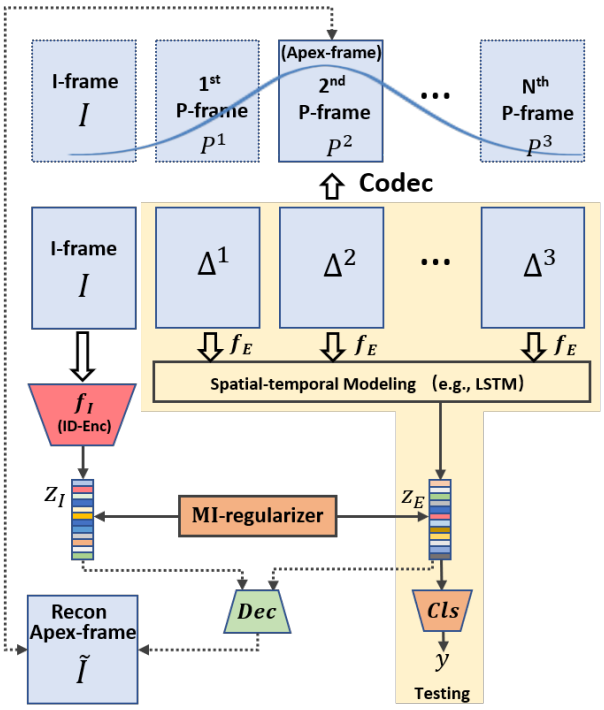

Figure 2: The illustration of our **M**utual **I**nformation Regularized in **C**ompressed Video (MIC) framework for identity-aware FER in the compressed video domain. Dec and Cls indicate decoder and classifier, respectively.

format that is suitable for recording and distribution of this video [65, 64]. Conventionally, the H.264/MPEG-4 and the Advanced Video Coding are the typical algorithms [86, 72]. Usually, the video codecs separates a video into several Group Of Pictures (GOP). A GOP is composed of an I frame, and followed by several P-frames or B-frames. I-frame indicates a self-contained RGB frame with full visual representation, while the P-frame or B-frame is the inter frames which hold motion vectors and residuals w.r.t. the previous frame [40]. [88, 87] propose to replace the optical flow by the motion vector. However, decoding the RGB images is still necessary in this method. More recently, in the action recognition task, [80, 9, 1] propose to fuse all of the available modalities in the compressed video, e.g., I-frame, residuals and motion vectors, to avoid the decoding of images. Besides, [78] proposes to achieve fast video-based objection detection in compressed domain. Although the expression shares some similarities with action recognition, the movement range and the utilization of I frame are essentially different. The I frame is directly used to predict the action and combine with the result of P frames [80], while the I frame in FER usually be the neutral face (different from the video label). Besides, the low resolution motion vector can not well encode the expression.

Mutual-information has a long history in unsupervised learning. The infomax principle [43, 11], as prescribed for neural networks, advocates to maximize the MI between network input and output. This can be the fundamental of many ICA algorithms, which can be nonlinear [28, 5] but are often hard to adapt for use with deep networks. Recently, some of the works propose to achieve unsupervised learning with MI. [12] proposes a generative adversarial network to minimize MI with positive and negative samples for Independent Component Analysis (ICA). It introduces a strategy to draw samples from the joint distribution and the product of marginal distributions and proposed to train an encoder and a discriminator to minimize the Jansen-Shannon divergence. Similarly, [58] proposes the Contrastive Predicting Coding (CPC). This work utilizes the auto-regressive network which is trained by the probabilistic contrastive loss. Moreover, it has recently been shown that the GAN framework can be extended not only to maximize or minimize MI but also to explicitly compute it using the Mutual Information Neural Estimation (MINE) proposed in [10]. In [25], the DeepInfoMax (DIM) is proposed to learn the representations based on both local and global information. In [75], Deep Graph Infomax (DGI) extends this approach to graph-structured data.

Inspired by these works, we targeting to utilize the MI as the justified measure of independence, and minimize it directly as our disentanglement objective.

Eliminating identity can benefit to extract more "pure" expression feature [56, 51]. The typical solution for FER is the metric learning. Our MI regularization is also related with the triplet loss [68], which maximize the Euclidean or cosine distance between two identities. With the development of GAN, the adversarial training also can be utilized for disentanglement [54, 50].

Instead, we consider the mutual information to be a more meaningful divergence to capture complex non-linear relationships, between the identity and expression representations. Besides, the identity label is required in these methods [56, 51, 47]. We can also choose adversarial training [50] as a baseline to achieve the identity elimination in our framework. We note that the adversarial game is notorious by its unstable model collapse [23], while our solution follows a collaborative way.

3. Methodology

Our goal is to design a fast and effective video-based FER framework operates directly on the compressed domain. The overall frame work is shown in Figure. 2, which are consist of four core modules. The pre-trained identity branch and FER branch (frame embedding network f_E, aggregation module and Classifier) work on the I frame and undecoded P frames respectively. Then, a differentiable mutual information regularization module is applied to measure the dependence of identity and expression representations. Moreover, the complementary constraint can be applied when the apex frame is annotated to stabilize the early stage training.

3.1. Modeling Compressed Representations

To illustrate the format of input video, we choose the MPEG-4 as an example [64, 72]. The compressed domain has two typical frames, i.e., I frame and P frames. Specifically, the I frame $I \in \mathbb{R}^{h \times w \times 3}$ is a complete RGB image. We use h and w to denote its height and width respec-

tively. Besides, the P frame at time t $P^t \in \mathbb{R}^{h\times w\times 3}$ can be reconstructed with the stored offsets, called residual errors $\Delta^t \in \mathbb{R}^{h\times w\times 3}$ and motion vectors $\mathcal{T}^t \in \mathbb{R}^{h\times w\times 2}$.

Noticing that the motion vectors $\mathcal{T}^t$ has much lower resolution, since its values within the same macroblock are identical. Considering the micro movements of facial expression in each frame, the low resolution $\mathcal{T}^t$ usually not helpful for the FER. For P frame reconstruction $P_i^t = P_{i-\mathcal{T}_i^t}^{t-1} + \Delta_i^t$, where index all the pixels and $P^0 = I$. Then, $\mathcal{T}^t$ and Δ^t are processed by discrete cosine transform and entropy-encoded.

The typical compression algorithms are only developed to compress the file size, and the encoded format can be very different with the RGB images w.r.t. the statistical and structural properties. Therefore, a tailored processing network is necessary to accommodate the compressed format. Considering the structure of residual images Δ^t are much simpler than the decoded images, it is possible to utilize simpler and faster CNNs $f_E : \mathbb{R}^{h\times w\times 3} \rightarrow \mathbb{R}^{512}$ to extract the feature of each frame [36, 7, 6]. Practically, we follow the CNN in the typical CNN-LSTM FER structure [7, 36, 6], but with fewer layers to explore the information in Δ^t. Noticing that f_E is shared for all frames, and only needs to store one f_E in processing.

Besides, most existing action recognition methods with compressed video [80, 70] independently concatenate the paired Δ^t and $\mathcal{T}^t$ at each time step and predict an action score of each P-frame. The temporal cues and its development patterns are important for FER task [42]. We simply choose the LSTM in [7] to model the sequential development of residual frames and summarize the information to a expression feature z_E. Since our LSTM is applied to 512-dim features, the computation burden is largely smaller than work on the raw images. Noticing that more advanced RNN, 3D CNN or attention networks can potentially be utilized to replace our LSTM model to further boost the performance [8, 92, 39].

For the I frame with raw image format, we simply use the FaceNet [68] pre-trained on millions of identities [35] as our identity feature extractor $f_I : \mathbb{R}^{h\times w\times 3} \rightarrow z_I$, where $z_I \in \mathbb{R}^{1024}$ denotes the identity feature. We note that Z_E and z_I do not need to have the same dimension for MI Regularization. Actually, the FER videos in many datasets start from the neutral expression which can further facilitate identity recognition.

3.2. MI regularization

To eliminate the identity related factors in our FER representation, we propose to utilize the identity feature from pre-trained face recognizer z_I as anchor, and explicitly inspect the information w.r.t. z_I in z_E. We note that previous identity-aware FER methods usually explicitly require the identity labels of FER datasets to sample the triplets [56, 51, 47], while the identity label is not common in FER. In contrast, our solution does not relies on identity label of FER samples, but utilize the easily available face recognition dataset.

Achieving the disentanglement of different factors requires two major objectives, i.e., 1) each factor has its specific information, and 2) does not incorporate the information of the other factors [50]. For example, z_E can achieve 1) using the conventional CE loss minimization w.r.t. the expression label y, and z_I is from the pre-trained identity extractor, which inherently has identity information. However, how to explicitly measure the dependency between these factors and minimize this metric to achieve the later objective can be challenging.

Algorithm 1 Training scheme of our framework

Initialize network parameters of f_E, $LSTM$, cls, T_θ and Dec.
while Not Converged **do**
  Randomly sample n compressed FE videos.
  Extract expression and identity feature z_E and z_I.
  Draw n (z_E, z_I) for the joint distribution.
  Draw n z_I for the marginal distribution.
  Evaluate lower bound $\widehat{MI(z_E; z_I)}_n$
  $= \frac{1}{n}\sum_{i=1}^{n} T(z_E, z_I, \theta) - \log(\frac{1}{n}\sum_{i=1}^{n} e^{T(z_E, \hat{z_I}, \theta)})$.
  Calculate the cross entropy loss for n samples
  $\mathcal{L}_{CE} = \frac{1}{n}\sum_{i=1}^{n}\{-\sum_{c=1}^{C} y_c \log(Cls(z_E)_c)\}$.
  Reconstruct $\hat{I}_{Apex}$ using z_E and z_I.
  Calculate the cross entropy loss for n samples
  $\mathcal{L}_1 = \frac{1}{n}\sum_{i=1}^{n} ||I_{Apex} - \hat{I}_{Apex}||_2^2$.
  Compute gradients and update network parameters.
  Cls $\leftarrow \nabla\mathcal{L}_{CE}$
  f_E and LSTM $\leftarrow \nabla\mathcal{L}_{CE} + \alpha\nabla\widehat{MI(z_E; z_I)}_n + \nabla\beta\mathcal{L}_1$.
  $T_\theta \leftarrow \nabla\widehat{MI(z_E; z_I)}_n$.
  $Dec \leftarrow \nabla\mathcal{L}_1$.

Actually, the mutual information (MI) is the exact metric to measure the amount of information obtained about one random variable through observing another random variable.

$$MI(z_E; z_I) = \int_{\mathcal{E}\times\mathcal{I}} \log \frac{d\mathbb{P}_{z_E z_I}}{d\mathbb{P}_{z_E} \otimes \mathbb{P}_{z_I}} d\mathbb{P}_{z_E z_I} \tag{1}$$

where z_E and z_I are the random variables follows the distribution $\mathcal{E}$ and $\mathcal{I}$ respectively. $\mathbb{P}_{z_E z_I}$ indicates the joint probability distribution of (z_E, z_I), $\mathbb{P}_{z_E} = \int_{\mathcal{I}} d\mathbb{P}_{z_E z_I}$ and $\mathbb{P}_{z_I} = \int_{\mathcal{E}} d\mathbb{P}_{z_E z_I}$ are the marginals. $\mathbb{P}_{z_E} \otimes \mathbb{P}_{z_I}$ is the product of the marginals.

MI minimization explicitly enforce the joint distribution to be equal to the product of marginals, which lead to the statistically independence of two vectors. Instead, the MI maximization can result in two vectors have the same information, and the MI is simply equal to the entropy of a variable.

MI is an important metric for a wide range of applications [10]. It can capture the non-linear statistical dependence of two variables [37]. However, it is mostly used as an evaluation metric rather than an optimization objective. This is because the fast computation of MI is limited to discrete variables [61]. For the continuous random variables, its complexity is quadratic to the number of samples, which is not desired for a loss function. Therefore, we propose to utilize the mutual information neural estimator (MINE) [10]

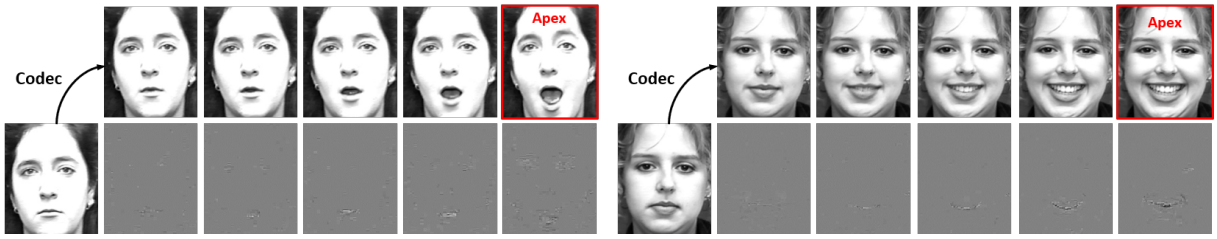

Figure 3: The illustration of the compressed and decoded frames in CK+ dataset.

to provide the unbiased estimation of MI on n independent and identically distributed (i.i.d.) samples. It is linearly scalable w.r.t. dimensionality and sample size, by leveraging a gradient descent over neural network $T_\theta : \mathcal{E} \times \mathcal{I} \rightarrow \mathbb{R}$. MINE propose to approximate MI by exploiting a lower-bound based on the *Donsker-Varadhan* representation of the Kullback-Leibler divergence. Therefore, the neural information measure can be formulated as

$$\widehat{MI(z_E; z_I)}_n = \sup_{\theta \in \Theta} \left\{ \mathbb{E}_{\mathbb{P}^n_{z_E z_I}}[T_\theta] - \log(\mathbb{E}_{\mathbb{P}^n_{z_E} \otimes \hat{\mathbb{P}}^n_{z_E}}[e^{T_\theta}]) \right\} \quad (2)$$

Given the distribution $\mathbb{P}$, $\hat{\mathbb{P}}^n_{z_E}$ denotes the empirical distribution associated to n i.i.d. samples. Since the supremum is taken over all functions of T, the two expectations are finite. Then, the MI in Eq. (2) can be estimated as follows:

$$\widehat{MI(z_E; z_I)}_n = \int\int \mathbb{P}^n_{z_E z_I}(z_E, z_I) T(z_E, z_I, \theta) - \log(\int\int \mathbb{P}^n_{z_E}(z_E) \mathbb{P}^n_{z_I}(z_I) [e^{T(z_E, z_I, \theta)}]) \quad (3)$$

Besides, we leverage Monte-Carlo integration to avoid computing the integrals to compute $\widehat{MI(z_E; z_I)}_n$ as

$$\frac{1}{n}\sum_{i=1}^{n} T(z_E, z_I, \theta) - \log(\frac{1}{n}\sum_{i=1}^{n} e^{T(z_E, \hat{z_I}, \theta)}) \quad (4)$$

where $\hat{z_I}$ is sampled from the marginal distribution. Note that (z_E, z_I) are sampled from the joint distribution $\mathbb{P}_{z_E z_I}$. Then, we can evaluate the bias corrected gradients (e.g., moving average) [10].

The estimated $MI(z_E; z_I)$ is used as the supervision to update the FER branch. By utilizing MI regularization, the adversarial discriminator [81, 50] is no longer needed in our new framework, which makes the balance of each module easier. Note that we need the additional neural network T_θ to measure the MI, but it is collaboratively trained with the FER branch to maximize the discrepancy between two features. Essentially, we are playing a *min-min* game instead of a *min-max* game. Therefore, it is easier to stabilize the training (compared to the adversarial training).

Besides, MI is a symmetric measure, while the conditional entropy $H(z_I|z_E) = H(z_I) - MI(z_I; z_E)$ optimized in conventional disentanglement works [81, 50] is asymmetric and essentially we should calculate both $H(z_I|z_E)$ and $H(z_I|z_E)$ as supervision signal [81, 50].

To maximize the discrepancy of z_E and z_I, we can also apply the adversarial disentanglement solutions [81, 50]. Nevertheless, with the above-mentioned limitations, such methods can be hard to optimize and lead to an inferior performance.

3.3. Complementary constraint

Several adversarial disentanglement works demonstrate that simply separate the input may result in the extracted feature has no meaningful information [54, 50, 52, 24]. The reconstruction of input can explicitly enforce the disentangled factors to be complementary to each other. However, reconstructing the video can be hugely underconstrained.

Many FER datasets follow a well-defined collection protocol, which usually starts from the neutral face and then develops to an expression. Specifically, the video in CK+ [34, 53] consists of a sequence which shift from the neutral expression to a apex facial expression. The last frame usually be the apex frame, which has the most strong expression intensity. Actually, the image-based FER methods select the last three frames to construct their training and testing datasets. Similarly, in MMI [62], the video frames are usually start from the neutral face and develop to the apex around the middle of video, and returning back to the neutral in the end of video. Noticing that the apex frame (i.e., last frame in CK+ or middle frame in MMI) can clearly incorporate both the identity and expression information. Therefore, we are possible to utilize the apex frame as a reference of reconstruction, and simply apply the $\mathcal{L}_2$ loss.

$$\mathcal{L}_1 = ||I_{Apex} - \hat{I}_{Apex}||_2^2 \quad (5)$$

where $\hat{I}_{Apex} = Dec(z_I, z_E)$. We note that the complementary constraint is not a necessity in our framework, since the FER loss takes a large weight in the FER branch. It requires to maintain sufficient information w.r.t. expression and not easy to have nothing meaningful. But the complementary constraint does helpful for the convergence in the early stage. When the apex is annotated, we only need to decode the apex frame in the decoded image domain in the start of few training epochs.

3.4. Overall objectives

We have three to be minimized objectives, i.e., cross-entropy loss, mutual information and $\mathcal{L}_1$ loss, which works collaboratively to update each module. The expression classification is the main task of the FER. We choose the typical cross-entropy loss $\mathcal{L}_{CE} = -\sum_{c=1}^{C} y_c \log(Cls(z_E)_c)$ to

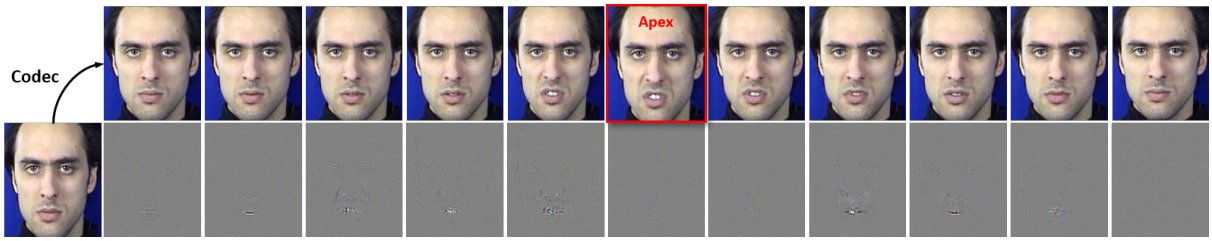

Figure 4: The illustration of the compressed and decoded frames in MMI dataset.

ensure z_E contains sufficient expression information and finally have a good performance on C-class expression classification. We use y_c and $Cls(z_E)_c$ indicate the c^{th} class probability of the label and classifier softmax predictions respectively. Since the FER branch can be updated with all of the losses, we assign the balance parameter $\alpha \in [0, 1]$ and $\beta \in [0, 1]$ to mutual information and $\mathcal{L}_1$ loss minimization objectives respectively.

The detailed training flow is shown in Algorithm 1. We note that only the FER branch, i.e., f_E, LSTM and Cls, is used for testing.

4. Experiments

In this section, we first detail our experimental setup, present quantitative analysis of our model, and finally compare it with state-of-the-art methods. The good FER accuracy and high inference speed in testing demonstrate its effectiveness.

4.1. Description of the datasets

CK+ Dataset [34, 53] is referring to the Cohn-Kanade AU-Coded Expression dataset. This dataset is a widely accepted FER benchmark [42, 39]. The video is collected in a restricted environment, in which the participate subjects are facing to the recorder with the empty background. The video in CK+ consists of a sequence which shift from the neutral expression to a apex facial expression. The last frame usually be the apex frame, which has the most strong expression intensity. The expression included in this dataset are anger, contempt, disgust, fear, happiness, sadness, and surprise. There are 327 facial expression videos that collected from 118 subjects. Following the previous works, we use subject independent 10-folds cross validation [46, 42, 47].

Many FER datasets follow a well-defined collection protocol, which usually starts from the neutral face and then develops to an expression. Specifically, the image-based FER methods select the last three frames to construct their training and testing datasets. Similarly, in MMI [62], the video frames are usually start from the neutral face and develop to the apex around the middle of video, and returning back to the neutral in the end of video. Noticing that the apex frame (i.e., last frame in CK+ or middle frame in MMI) can clearly incorporate both the identity and expression information. Therefore, we are possible to utilize the apex frame as a reference of reconstruction, and simply apply the $\mathcal{L}_1$ loss.

Method	Accuracy	Landmarks	Ave Test
STM-ExpLet (2014) [46]	94.19	×	-
LOMo (2016) [71]	95.10	✓	-
DTAGN (2015) [32]	97.25	✓	-
PHRNN-MSCNN (2017) [89]	98.50	✓	-
C3D-GRU (2019) [41]	97.25	×	-
CTSLSTM (2019) [26]	93.9	✓	-
(N+M)-tuplet (2019) [47]	93.90	✓	12fps
SC (2019) [76]	97.60	✓	-
G2-VER (2019) [3]	97.40	×	-
LBVCNN (2019) [39]	97.38	×	-
Mode VLSTM (2019) [7]	97.42	×	11fps
MIC	**98.95**	×	**35fps**
MIC-MI	97.84	×	**35fps**
MIC-MI+Adv[50]	98.78	×	**35fps**
MIC-$\hat{I}$	98.72	×	**35fps**
MIC+$\mathcal{T}^t$	98.93	×	29fps

Table 1
Comparison of various methods on the CK+ dataset in terms of average recognition accuracy of seven expressions. Note that in order to make the comparison fair, we do not consider image-based and 3D geometry based experiment setting and models [51, 56, 49].

MMI Dataset [62] consist of a total of 326 facial expression videos from 32 participants. There are 213 labeled videos with the expression label angry, disgust, fear, happy, sad, and surprise. The video frames are start from the neutral face. Then the expression is developed to the apex in the middle of video, and returning back to the neutral in the end of video. In our experiments, we follow the previous works to use subject independent 10-folds cross validation [42, 47].

AFEW Dataset [17] is more close to the uncontrolled real-world environment. It is consist of the video clips of movies [63]. The video in AFEW has the spontaneous facial expression. The AFEW has seven expressions: anger, disgust, fear, happiness, sadness, surprise and neutral. Following the evaluation protocol in EmotiW [16], there are training, validation and testing sets. Since its testing label is not available, we follow the previous work to use the validation set for comparison [7]. Noticing that the validation set is not used in training stage for parameter or hyper-parameter tuning.

4.2. Implementation details

We preprocess video frames and augment the data according to [7, 36, 6] for fair comparison. For these three datasets, the videos only has one GOP and not need to segment the video. We utilize the Pytorch deep learning platform for our framework. In the training stage, on all datasets, we set the batch size to 48. All of the modules use the Adam

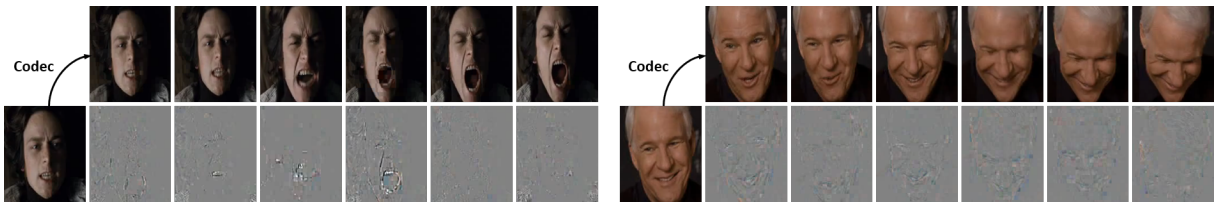

Figure 5: The illustration of the compressed and decoded frames in AFEW dataset.

Method	ID accuracy	MI
Mode variational LSTM (2019) [7]	28.4	1.57
MIC	**0.8**	**0.01**
MIC-MI	5.7	1.22
MIC-MI+Adv[50]	1.4	0.08
MIC-$\hat{I}$	0.9	0.10
MIC+$\mathcal{T}^t$	**0.8**	0.09

Table 2
Comparison of the identity eliminating on CK+ dataset w.r.t. the identity recognition accuracy using z_E, and the mutual information between z_E and z_I.

Method	Accuracy	Landmarks	Ave Test
3D CNN-DAP (2014) [45]	63.4	✓	
DTAGN (2015) [32]	70.24	✓	
CNN+LSTM (2017) [36]	78.61	×	
CNN+DBN (2019) [90]	71.43	optical flow	3fps
CTSLSTM (2019) [26]	78.40	✓	8fps
Mode VLSTM (2019) [7]	79.33	×	10fps
MIC	**81.29**	×	**32fps**
MIC-MI	80.25	×	**32fps**
MIC-MI+Adv[50]	80.98	×	**32fps**
MIC-$\hat{I}$	80.94	×	**32fps**
MIC+$\mathcal{T}^t$	81.24	×	28fps

Table 3
Comparison of various methods on the MMI dataset in terms of average recognition accuracy of seven expressions. Note that in order to make the comparison fair, we do not consider image-based and 3D geometry based experiment setting and models [51, 56, 49].

optimizer with momentum 0.9, and a weight decay of 1e-5 for 100 training epochs. On the CK+ and MMI datasets, the learning rate is initialized to 1e-1, and be modified to 1e-2 for the 30^{th} epoch. For the AFEW dataset, we initialize the learning rate to 1e-4, and modify it to 8e-6 for the 30^{th} epoch and 1e-7 for the 60^{th} epochs.

All of our training/testing use an NVIDIA Titan X GPU. We note that the calculation of the accumulated residuals to recover apex frame is measured on Intel E5-2698 v4 CPUs, but we do not need this operation in testing. For the testing speed, we measure the average frame per second (fps) according to the average running time which is the sum of the data pre-processing time and the FER branch forward pass time.

Practically, our f_E, LSTM and Cls follow the CNN-LSTM structure in [36, 7, 6] for fair comparison. Considering the relatively simpler residual data, we use part of convolutional layer for f_E (i.e., C1, C2 and F4 layers as in [36, 7, 6]) and a fully connected layer for Cls (i.e., $\mathbb{R}^{512} \rightarrow \mathbb{R}^6 or \mathbb{R}^7$).

For the CK+ and MMI dataset, we choose the last or middle frame as the apex reference image respectively. Since the complementary constraint is only used to stabilize the initial training of disentanglement, we uniformly decrease β to 0 until the 30^{th} epoch. Practically, we use grid search to find the optimal α and set it to 0.1, 0.1, 0.2 on CK+, MMI and AFEW datasets respectively.

4.3. Evaluation and ablation study

Results on CK+ dataset. The 10-fold cross-validation performance of our proposed method is shown in the Table 1. For fair comparison, the image-based experiment settings are not incorporated in the tables. Besides, only the state-of-the-art (SOTA) accuracy obtained by the single-models (non-ensemble model) are listed.

Many models, e.g., DTAGN [32], LOMo [71], PHRNN-MSCNN [89], CTSLSTM [26] and SC [76], achieved the SOTA performance by utilizing the facial landmarks. However, this operation highly relies on the fine-grained landmark detection, which itself is a challenging task [42, 49], and unavoidably introduced additional computation.

Based on the mode variational LSTM [7], our proposed MIC achieves the SOTA result without the facial landmarks, 3D face models or optical flow. It worth noticing that the much simpler CNN encoding network makes more residual frames can be processed parallel than [7]. Moreover, our efficiency is also benefit from avoiding the decompress of the video. Since the videos are typically stored and transmitted with the compressed version, and the residuals are off-the-shelf. As a result, the proposed compressed domain MIC can speed up the testing about 3 times over [7], and achieve better accuracy.

Besides, [47] is a typical metric-learning based identity removing method. Our solution can significantly outperform it with respect to both speed and accuracy. Actually, the sampling of tuplets usually makes the training not scalable [49], while our identity eliminating scheme is concise and effective.

When we remove some modules from our framework, the performances have different degrees of decline. We use -MI and -$\hat{I}$ to denote the MIC without MI regularizer or complementary constraint respectively. The performance drop of MIC is significant when we remove the MI regularization module, which further evidenced that the identity can be a notorious factor for FER. The result also implies that the identity can be well encoded by the face recognition net-

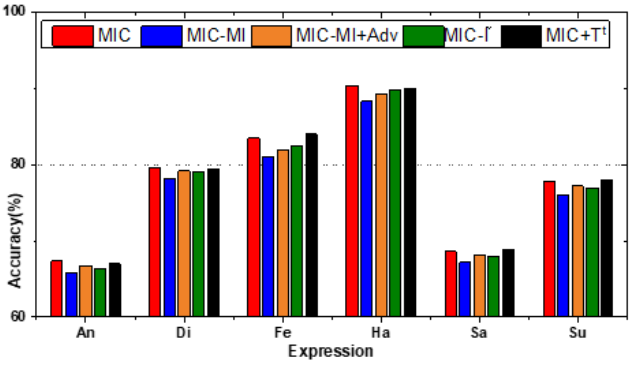

Figure 6: Comparison of accuracy according to each emotion among five networks on MMI dataset.

Method	Model type	Accuracy	Ave Test
CNN-RNN (2016) [22]	45.43	Dynamic	-
Undirectional LSTM (2017) [77]	48.60	Dynamic	-
HoloNet (2016) [85]	44.57	Static	-
DSN-HoloNet (2017) [27]	46.47	Static	-
DenseNet-161 (2018) [44]	51.44	Static	-
DSN-VGGFace (2018) [21]	48.04	Static	-
FAN (2019) [55]	51.18	Static	-
CTSLSTM (2019) [26]	51.2	✓	-
C3D-GRU (2019) [41]	49.87	Dynamic	-
DSTA (2019)* [60]	42.98	Dynamic	-
E-ConvLSTM (2019)* [57]	45.29	Dynamic	4fps
Mode VLSTM (2019) [7]	51.44	Dynamic	11fps
MIC	**53.18**	Dynamic	**34fps**
MIC-MI	52.62	Dynamic	**34fps**
MIC-MI+Adv[50]	53.01	Dynamic	**34fps**
MIC+$\mathcal{T}^t$	**53.18**	Dynamic	30fps

Table 4
Comparison of various methods on the AFEW dataset in terms of average recognition accuracy of seven expressions. *optical flow is used.

work and the disentanglement with mutual information is feasible. Compared with using the conventional adversarial training based disentanglement [50] as an alternative (i.e., MIC-MI+Adv), our MI regularizer is easy to train and can converge with 1.8 times faster in the training.

We can also follow the action recognition method [70] to concatenate the motion vector and residual as the input, and denote as MIC+$\mathcal{T}^t$. However, we do not achieve significant performance on all datasets, but the inference speed in testing can be slower. This maybe related to the coarse resolution of the motion vector can not well describe the fine-grain muscle movement of the face.

The confusion matrix of our proposed MIC method on the CK+ is reported in Figure 7 (left). The accuracy for the expression class happiness, disgust, angry, surprise and contempt are almost perfect.

In Table 2, we investigate the identity eliminating performance. The first metric is following [50] to use recognise identity with z_E. Besides, we can directly use the mutual information as the metric of independence. We can see that the residual frame itself can incorporate much less identity information than the decoded images as in [7], while it is still possible to detect identity with the facial contours. The MI regularization can explicitly remove the identity factors and outperforms the adversarial training [50].

Results on MMI dataset.

The evaluation results on the MMI dataset are shown in Table 3. The performance is also consistent with the CK+ dataset, which evidenced its effectiveness and generality. All of our methods achieve comparable performance than the landmark-based STOA methods. It is more promising that MIC can be significantly better than the methods without the landmarks. Our MIC is efficient, since we explore the correlation in video frames in the compressed domain.

In Figure 6, we give a comparison of accuracy w.r.t. each emotion among five MIC baselines, and the confusion matrix of our MIC is reported in Figure 7 (middle). There are good performance for the expression class of fear, happiness, sadness and surprise. While the accuracy of expression class anger and disgust is relatively limited. Especially, there is a high degree of confusion between the anger and disguise. This maybe related to the subtle movements between these expressions are relatively in the residual frames.

Results on AFEW dataset. The evaluation of the proposed MIC on AFEW dataset is shown in Table 4. We note that only the SOTA accuracy obtained by the single-models (non-ensemble model) are listed for fair comparison. Besides, the audio modality in AFEW can be used to boost the recognition performance [20, 22, 77]. We note that we only focus on the image compression in this paper, and the audio/video data are stored in separate tracks, but the additional modality can also potentially to be add on our framework following the multi-modal methods[20, 22].

With the simplified Mode variational LSTM-based [7] backbone, the exploration in the compressed domain can achieve comparable or even better recognition performance. More promisingly, our MIC can also achieve real-time processing for the uncontrolled environment, which evidenced its generality. We note that the typical time resolution in FER is 24fps [36].

Some of the works propose to improve the image-based FER networks and combine the frame-wise scores for video-based FER [85, 27, 44, 21, 55]. [85, 27] input both the LBP maps and the image to the CNNs. [27] and [21] utilize the additional supervision on intermediate layers. The image-based FER methods [44] achieves high performance, but [44] uses a very deep network DenseNet-161 and pre-trains it on the private Situ dataset. Moreover, [44] utilize the sophisticated post-processing. Actually, an intuition of statistic based solution is to avoid LSTM and speed up the processing. However, with the super deep and complicated structure, their processing can be much slower than our solution.

Both [22] and [77] use VGGFace as the backbone of f_E and a RNN model with LSTM units. They target to capture the temporal dynamic cues of the videos. Moreover, [26, 41, 60] also propose to modify the LSTM model for the spatial-temporal modeling. However, all of the above solutions are applied to the decoded space, which requires

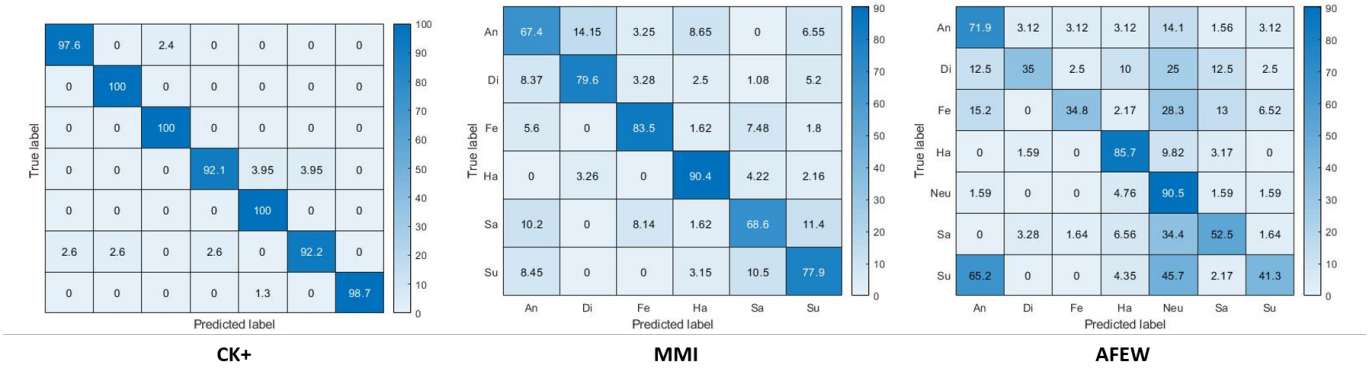

Figure 7: Confusion matrix of MIC on CK+, MMI and AFEW datasets.

decoding processing and needs to handle much more complicated data. Overall, the proposed MIC can improve the testing speed by a large margin and can achieves the SOTA accuracy as the previous models.

5. Conclusion

In this work, we propose to explore the facial expression cues directly on the compressed video domain. We are motivated by our practical observation that the facial muscle movements can be well encoded in the residual frames which can be informative and free of cost. Besides, the video compression can reduce the repeating boring patterns in the videos, which rendering the representation to be robust. The increased relevance and reduced complexity or redundancy in FER videos make computation much more effective. We extract the identity and expression factor from the I frame and P frame respectively, and explicitly enforce their independence with concise and effective mutual information regularization. When the apex frame label is available in training, the complementary constraint can further stabilize the training. In three video-based FER benchmarks, our MIC can improve the performance without the additional identity, face model or facial landmarks labels. The processing speed of the test stage is promising for real-time FER. Moreover, our mutual information regularization can potentially be a good alternative of adversarial training [50] for many disentanglement tasks.

References

[1] Abdari, A., Amirjan, P., Mansouri, A., 2019. Action recognition in compressed domain using residual information, in: 2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA), IEEE. pp. 130–134.

[2] AL CHANTI, D.A., Caplier, A., 2018. Deep learning for spatio-temporal modeling of dynamic spontaneous emotions. IEEE Transactions on Affective Computing .

[3] Albrici, T., Fasounaki, M., Salimi, S.B., Vray, G., Bozorgtabar, B., Ekenel, H.K., Thiran, J.P., 2019. G2-ver: Geometry guided model ensemble for video-based facial expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE. pp. 1–6.

[4] Ali, K., Hughes, C.E., 2019. All-in-one: Facial expression transfer, editing and recognition using a single network. arXiv preprint arXiv:1911.07050 .

[5] Almeida, L.B., 2003. Misep–linear and nonlinear ica based on mutual information. Journal of Machine Learning Research 4, 1297–1318.

[6] Baddar, W.J., Lee, S., Ro, Y.M., 2019. On-the-fly facial expression prediction using lstm encoded appearance-suppressed dynamics. IEEE Transactions on Affective Computing .

[7] Baddar, W.J., Ro, Y.M., 2019. Mode variational lstm robust to unseen modes of variation: Application to facial expression recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3215–3223.

[8] Barros, P., Wermter, S., 2016. Developing crossmodal expression recognition based on a deep neural model. Adaptive behavior 24, 373–396.

[9] Battash, B., Barad, H., Tang, H., Bleiweiss, A., 2019. Mimic the raw domain: Accelerating action recognition in the compressed domain. arXiv preprint arXiv:1911.08206 .

[10] Belghazi, M.I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, R.D., 2018. Mine: mutual information neural estimation. arXiv preprint arXiv:1801.04062 .

[11] Bell, A.J., Sejnowski, T.J., 1995. An information-maximization approach to blind separation and blind deconvolution. Neural computation 7, 1129–1159.

[12] Brakel, P., Bengio, Y., 2017. Learning independent features with adversarial nets for non-linear ica. arXiv preprint arXiv:1710.05050 .

[13] Cai, J., Meng, Z., Khan, A.S., Li, Z., O'Reilly, J., Tong, Y., 2019. Identity-free facial expression recognition using conditional generative adversarial network. arXiv preprint arXiv:1903.08051 .

[14] Cai, J., Meng, Z., Khan, A.S., Li, Z., O'Reilly, J., Tong, Y., 2018. Island loss for learning discriminative features in facial expression recognition, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE. pp. 302–309.

[15] Calder, A.J., Young, A.W., 2005. Understanding the recognition of facial identity and facial expression. Nature Reviews Neuroscience 6, 641–651.

[16] Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., Gedeon, T., 2017. From individual to group-level emotion recognition: Emotiw 5.0, in: Proceedings of the 19th ACM international conference on multimodal interaction, pp. 524–528.

[17] Dhall, A., Goecke, R., Joshi, J., Sikka, K., Gedeon, T., 2014. Emotion recognition in the wild challenge 2014: Baseline, data and protocol, in: Proceedings of the 16th international conference on multimodal interaction, pp. 461–466.

[18] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634.

[19] Dosodia, P., Poonia, A., Gupta, S.K., Agrwal, S.L., 2015. New gabor-dct feature extraction technique for facial expression recognition, in: 2015 Fifth International Conference on Communication Systems and Network Technologies, IEEE. pp. 546–549.

[20] Fan, X., Tjahjadi, T., 2017. A dynamic framework based on local zernike moment and motion history image for facial expression recognition. Pattern recognition 64, 399–406.

[21] Fan, Y., Lam, J.C., Li, V.O., 2018. Video-based emotion recognition using deeply-supervised neural networks, in: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 584–588.

[22] Fan, Y., Lu, X., Li, D., Liu, Y., 2016. Video-based emotion recognition using cnn-rnn and c3d hybrid networks, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 445–450.

[23] Goodfellow, I., 2016. Nips 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160 .

[24] Hadad, N., Wolf, L., Shahar, M., 2018. A two-step disentanglement method, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 772–780.

[25] Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y., 2018. Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670 .

[26] Hu, M., Wang, H., Wang, X., Yang, J., Wang, R., 2019. Video facial emotion recognition based on local enhanced motion history image and cnn-ctslstm networks. Journal of Visual Communication and Image Representation 59, 176–185.

[27] Hu, P., Cai, D., Wang, S., Yao, A., Chen, Y., 2017. Learning supervised scoring ensemble for emotion recognition in the wild, in: Proceedings of the 19th ACM international conference on multimodal interaction, pp. 553–560.

[28] Hyvärinen, A., Pajunen, P., 1999. Nonlinear independent component analysis: Existence and uniqueness results. Neural Networks 12, 429–439.

[29] Jain, D.K., Zhang, Z., Huang, K., 2017. Multi angle optimal pattern-based deep learning for automatic facial expression recognition. Pattern Recognition Letters .

[30] Jiang, B., Valstar, M.F., Pantic, M., 2011. Action unit detection using sparse appearance descriptors in space-time video volumes, in: Face and Gesture 2011, IEEE. pp. 314–321.

[31] Jiyoung, L., Sunok, K., Seungryong, K., Kwanghoon, S., 2020. Multi-modal recurrent attention networks for facial expression recognition. IEEE Transactions on Image Processing .

[32] Jung, H., Lee, S., Yim, J., Park, S., Kim, J., 2015. Joint fine-tuning in deep neural networks for facial expression recognition, in: Proceedings of the IEEE international conference on computer vision, pp. 2983–2991.

[33] Kahou, S.E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., Jean, S., Froumenty, P., Dauphin, Y., Boulanger-Lewandowski, N., et al., 2016. Emonets: Multimodal deep learning approaches for emotion recognition in video. Journal on Multimodal User Interfaces 10, 99–111.

[34] Kanade, T., Cohn, J.F., Tian, Y., 2000. Comprehensive database for facial expression analysis, in: Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), IEEE. pp. 46–53.

[35] Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E., 2016. The megaface benchmark: 1 million faces for recognition at scale, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4873–4882.

[36] Kim, D.H., Baddar, W.J., Jang, J., Ro, Y.M., 2017. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Transactions on Affective Computing 10, 223–236.

[37] Kinney, J.B., Atwal, G.S., 2014. Equitability, mutual information, and the maximal information coefficient. Proceedings of the National Academy of Sciences 111, 3354–3359.

[38] Klaser, A., Marszałek, M., Schmid, C., 2008. A spatio-temporal descriptor based on 3d-gradients.

[39] Kumawat, S., Verma, M., Raman, S., 2019. Lbvcnn: Local binary volume convolutional neural network for facial expression recognition from image sequences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0.

[40] Le Gall, D., 1991. Mpeg: A video compression standard for multimedia applications. Communications of the ACM 34, 46–58.

[41] Lee, M.K., Choi, D.Y., Kim, D.H., Song, B.C., 2019. Visual scene-aware hybrid neural network architecture for video-based facial expression recognition, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE. pp. 1–8.

[42] Li, S., Deng, W., 2020. Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing .

[43] Linsker, R., 1988. Self-organization in a perceptual network. Computer 21, 105–117.

[44] Liu, C., Tang, T., Lv, K., Wang, M., 2018a. Multi-feature based emotion recognition for video clips, in: Proceedings of the 20th ACM International Conference on Multimodal Interaction, pp. 630–634.

[45] Liu, M., Li, S., Shan, S., Wang, R., Chen, X., 2014a. Deeply learning deformable facial action parts model for dynamic expression analysis, in: Asian conference on computer vision, Springer. pp. 143–157.

[46] Liu, M., Shan, S., Wang, R., Chen, X., 2014b. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1749–1756.

[47] Liu, X., Ge, Y., Yang, C., Jia, P., 2018b. Adaptive metric learning with deep neural networks for video-based facial expression recognition. Journal of Electronic Imaging 27, 013022.

[48] Liu, X., Guo, Z., Li, S., Kong, L., Jia, P., You, J., Kumar, B., 2019a. Permutation-invariant feature restructuring for correlation-aware image set-based recognition, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4986–4996.

[49] Liu, X., Kumar, B.V., Jia, P., You, J., 2019b. Hard negative generation for identity-disentangled facial expression recognition. Pattern Recognition 88, 1–12.

[50] Liu, X., Li, S., Kong, L., Xie, W., Jia, P., You, J., Vijaya Kumar, B., 2019c. Feature-level frankenstein: Eliminating variations for discriminative recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 637–646.

[51] Liu, X., Vijaya Kumar, B., You, J., Jia, P., 2017. Adaptive deep metric learning for identity-aware facial expression recognition, in: CVPRW, pp. 522–531.

[52] Liu, Y., Wei, F., Shao, J., Sheng, L., Yan, J., Wang, X., 2018c. Exploring disentangled feature representation beyond face identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2080–2089.

[53] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I., 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: 2010 ieee computer society conference on computer vision and pattern recognition-workshops, IEEE. pp. 94–101.

[54] Mathieu, M.F., Zhao, J.J., Zhao, J., Ramesh, A., Sprechmann, P., LeCun, Y., 2016. Disentangling factors of variation in deep representation using adversarial training, in: NIPS, pp. 5040–5048.

[55] Meng, D., Peng, X., Wang, K., Qiao, Y., 2019. frame attention networks for facial expression recognition in videos, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 3866–3870.

[56] Meng, Z., Liu, P., Cai, J., Han, S., Tong, Y., 2017. Identity-aware convolutional neural network for facial expression recognition, in: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE. pp. 558–565.

[57] Miyoshi, R., Nagata, N., Hashimoto, M., 2019. Facial-expression recognition from video using enhanced convolutional lstm, in: 2019 Digital Image Computing: Techniques and Applications (DICTA), IEEE. pp. 1–6.

[58] Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 .

[59] Ouyang, X., Kawaai, S., Goh, E.G.H., Shen, S., Ding, W., Ming, H., Huang, D.Y., 2017. Audio-visual emotion recognition using deep transfer learning and multiple temporal models, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 577–582.

[60] Pan, X., Ying, G., Chen, G., Li, H., Li, W., 2019. A deep spatial and temporal aggregation framework for video-based facial expression recognition. IEEE Access 7, 48807–48815.

[61] Paninski, L., 2003. Estimation of entropy and mutual information. Neural computation 15, 1191–1253.

[62] Pantic, M., Valstar, M., Rademaker, R., Maat, L., 2005. Web-based database for facial expression analysis, in: 2005 IEEE international conference on multimedia and Expo, IEEE. pp. 5–pp.

[63] Perveen, N., Roy, D., Mohan, C.K., 2018. Spontaneous expression recognition using universal attribute model. IEEE Transactions on Image Processing 27, 5575–5584.

[64] Richardson, I.E., 2003. H264/mpeg-4 part 10 white paper-prediction of intra macroblocks. Internet Citation, Apr 30.

[65] Richardson, I.E., 2004. H. 264 and MPEG-4 video compression: video coding for next-generation multimedia. John Wiley & Sons.

[66] Rifai, S., Bengio, Y., Courville, A., Vincent, P., Mirza, M., 2012. Disentangling factors of variation for facial expression recognition, in: European Conference on Computer Vision, Springer. pp. 808–822.

[67] Sandbach, G., Zafeiriou, S., Pantic, M., Yin, L., 2012. Static and dynamic 3d facial expression recognition: A comprehensive survey. Image and Vision Computing 30, 683–697.

[68] Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823.

[69] Scovanner, P., Ali, S., Shah, M., 2007. A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th ACM international conference on Multimedia, pp. 357–360.

[70] Shou, Z., Lin, X., Kalantidis, Y., Sevilla-Lara, L., Rohrbach, M., Chang, S.F., Yan, Z., 2019. Dmc-net: Generating discriminative motion cues for fast compressed video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1268–1277.

[71] Sikka, K., Sharma, G., Bartlett, M., 2016. Lomo: Latent ordinal model for facial analysis in videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5580–5589.

[72] Sofokleous, A., 2005. H. 264 and mpeg-4 video compression: Video coding for next-generation multimedia.

[73] Tian, Y., Kanade, T., Cohn, J.F., 2011. Facial expression recognition, in: Handbook of face recognition. Springer, pp. 487–519.

[74] Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, pp. 4489–4497.

[75] Veličković, P., Fedus, W., Hamilton, W.L., Liò, P., Bengio, Y., Hjelm, R.D., 2018. Deep graph infomax. arXiv preprint arXiv:1809.10341 .

[76] Verma, M., Kobori, H., Nakashima, Y., Takemura, N., Nagahara, H., 2019. Facial expression recognition with skip-connection to leverage low-level features, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 51–55.

[77] Vielzeuf, V., Pateux, S., Jurie, F., 2017. Temporal multimodal fusion for video emotion classification in the wild, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp. 569–576.

[78] Wang, S., Lu, H., Deng, Z., 2019. Fast object detection in compressed video, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 7104–7113.

[79] Wang, X., Girshick, R., Gupta, A., He, K., 2018. Non-local neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7794–7803.

[80] Wu, C.Y., Zaheer, M., Hu, H., Manmatha, R., Smola, A.J., Krähenbühl, P., 2018. Compressed video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6026–6035.

[81] Xie, Q., Dai, Z., Du, Y., Hovy, E., Neubig, G., 2017. Controllable invariance through adversarial feature learning, in: NIPS, pp. 585–596.

[82] Xie, S., Hu, H., Wu, Y., 2019. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. Pattern Recognition 92, 177–191.

[83] Xu, B., Fu, Y., Jiang, Y.G., Li, B., Sigal, L., 2016. Video emotion recognition with transferred deep feature encodings, in: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pp. 15–22.

[84] Yan, H., 2018. Collaborative discriminative multi-metric learning for facial expression recognition in video. Pattern Recognition 75, 33–40.

[85] Yao, A., Cai, D., Hu, P., Wang, S., Sha, L., Chen, Y., 2016. Holonet: towards robust emotion recognition in the wild, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, pp. 472–478.

[86] Yeo, C., Ahammad, P., Ramchandran, K., Sastry, S.S., 2006. Compressed domain real-time action recognition, in: 2006 IEEE Workshop on Multimedia Signal Processing, IEEE. pp. 33–36.

[87] Zhang, B., Wang, L., Wang, Z., Qiao, Y., Wang, H., 2016. Real-time action recognition with enhanced motion vector cnns, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2718–2726.

[88] Zhang, B., Wang, L., Wang, Z., Qiao, Y., Wang, H., 2018. Real-time action recognition with deeply transferred motion vector cnns. IEEE Transactions on Image Processing 27, 2326–2339.

[89] Zhang, K., Huang, Y., Du, Y., Wang, L., 2017. Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Transactions on Image Processing 26, 4193–4203.

[90] Zhang, S., Pan, X., Cui, Y., Zhao, X., Liu, L., 2019. Learning affective video features for facial expression recognition via hybrid deep learning. IEEE Access 7, 32297–32304.

[91] Zhao, G., Pietikainen, M., 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE transactions on pattern analysis and machine intelligence 29, 915–928.

[92] Zhao, J., Mao, X., Zhang, J., 2018. Learning deep facial expression features from image and optical flow sequences using 3d cnn. The Visual Computer 34, 1461–1475.