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

Xiaofeng Liu\textsuperscript{a}, Linghao Jin\textsuperscript{a,b}, Xu Han\textsuperscript{a,b} and Jane You\textsuperscript{c}

\textsuperscript{a}Harvard University, Cambridge, MA, USA
\textsuperscript{b}John Hopkins University, Baltimore, MD, USA
\textsuperscript{c}Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Keywords:
Facial Expression Recognition
Mutual Information
Disentangled Representation
Compressed Video

ABSTRACT

How to extract effective expression representations that invariant to the identity-specific attributes is a long-lasting problem for facial expression recognition (FER). Most of the previous methods process the RGB images of a sequence, while we argue that the off-the-shelf and valuable expression-related muscle movement is already embedded in the compression format. In this paper, we target to explore the inter-subject variations eliminated facial expression representation in the compressed video domain. In the up to two orders of magnitude compressed domain, we can explicitly infer the expression from the residual frames and possibly extract identity factors from the I frame with a pre-trained face recognition network. By enforcing the marginal independence 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 \cite{1, 2}. Considering the natural dynamic property of the human face expression \cite{3}, many works propose to explore spatio-temporal features of facial expression recognition (FER) from the videos. Recently, deep neural networks (DNN) have achieved significant progress for image-based facial expression recognition \cite{4}, while the processing of expression video is still challenging.

Although the multi-frame series can inherit richer information and the temporal-correlation between consecutive frames can typically be useful for FER, the video also introduced a lot of redundancy. The signal-to-noise ratio (SNR) in FER videos is exceptionally low due to the slight muscle activity \cite{5, 4}.

The typical spatio-temporal FER DNNs explore the spatial and temporal clues in a series of frames to extract the facial expression-related information \cite{5}. The recursive neural network (RNN), 3D convolutional, and non-local networks are the most frequently used network backbones \cite{6, 7}. Unfortunately, using DNN to process many consecutive frames can be computationally costly. It can be hard to scalable for the long videos \cite{8, 4}. Furthermore, for some

\begin{itemize}
  \item The funding support from Jiangsu Youth Programme [BK20200238], PolyU Central Research Grant G-YBJW, and Hong Kong Government General Research Fund GRF (Ref. No.152202/14E) are greatly appreciated.
  \item liuxiaofengcmu@gmail.com (X. Liu); ljin23@jhu.edu (L. Jin);
  \item xhan32@jhu.edu (X. Han); csysjia@comp.polyu.edu.hk (J. You)
  \item https://liuxiaofeng.github.io/ (X. Liu);
  \item https://www.comp.polyu.edu.hk/en-us/staffs/detail/1261 (J. You)
  \item ORCID(i)s: 0000-0002-4514-2016 (X. Liu)
\end{itemize}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{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.}
\end{figure}

networks, such as RNNs, modeling long-term dependence can be difficult \cite{2}. Many FER techniques, in particular, are able to achieve good performance using image-based FER with decision-level fusion, which completely lost the tem-
poral dependence. This suggests the temporal cues can be hard to explore in the low SNR FER videos with the conventional solutions [9, 10].

We argue that the compressed domain can be suitable for the FER task for four reasons. 1) the consecutive frames in video modality have many uninformative and repeating patterns, which may drown the “interesting” and “true” signal [11]. With the standard video compression algorithms, the compression ratio can usually be hundred times [12]. Manipulating on the compressed domain can significantly reduce the cost of computation and memory. 2) the typical compression methods (e.g., MPEG-4, H.264, and HEVC) break the video to the I frame (intraencoded frames) with the first image, and follows several P frames (predictive frames) which encoded as the “change” or “movement” [13], as shown in Figure. 1. The fundamental of expression is the action of the face muscle. Many FER systems, in reality, are based on the action unit framework [14].

As a result, compressed P frames will inherit off-the-shelf but useful expression-related factors, and their pattern is substantially simpler than raw images. 3) our compressed domain exploration can also be effective since it focuses on the “true” signals rather than processing the repeatedly near-duplicates [15]. 4) Since the to-be-processed data is transmitted in compressed format, the decoding procedure is not needed in the real-world mission.

Furthermore, the FER task has long suffered from the high inter-subject variation caused by identity discrepancies in facial attributes [16, 8]. The learned features may capture more identity-related information than expression-related information, and are not purely related to the FER task. Noticing that the P frames may contain the relative location of face key points, which can be related to the identity [17]. Metric learning is a standard approach for extracting identity from the expression representation [16, 8, 4]. Inspired by the adversarial disentanglement, [18, 19] propose to render the identity removed face, which is inspired by adversarial disentanglement (GAN). These researchs, on the other hand, concentrate on image-based FER. [3] extend the metric learning [8] for video data by explicitly substitute the image with the video features, without taking into account the video’s characteristics. Furthermore, these methods necessitate the use of the identity label and multiple expressions of the same person, which significantly restricts their applicability to the in-the-wild FER task [20].

In this paper, we target to exploit the identity information from the I frame using a pre-trained face recognition network, e.g., FaceNet [21]. Their identity embeddings are remarkably reliable, since they achieve high accuracy over millions of identities [22], 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 marginal independence of our identity and expression feature. Instead of the complicated adversarial training [23] for disentanglement, we adopt the mutual information (MI) as the statistical measure of the independence of these two representations [24]. The MI of two random variables can usually be intractable to directly and precisely measure in a high-dimensional space [25]. Recently, some of the works illustrate that mutual information can be differentiable approximated [26]. 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 [27], while our MI regularization is concise and efficient.

This work extends our previous work [28] in the following significant ways:

- A novel expression and identity disentanglement framework based on practical mutual information minimization, which follows a min-min game with the joint and marginal distribution sampling.
- We demonstrate the generality of our framework in more FER dataset, i.e., MMI, carry out all experiments using the novel MIC framework.
- Moreover, the systematical cross-dataset evaluation, sensitivity analysis, and identity feature extraction analysis are provided.

The main contributions of this paper are summarized as follows:

- 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 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 thoroughly researched, since facial expression is a natural and universal means for human communication [5, 29]. Considering that the expression is essentially a dynamic action which should take minute muscle movements through time into account [30]. Traditionally, the handcrafted features are utilized to represent the spatio-temporal cues and for FER. Frame aggregation and spatiotemporal FER networks are being developed in parallel with the exponential growth of deep learning. Frame aggregation approaches may make use of image-based FER networks by conducting frame-wise aggregation at the decision-level [9] or feature-level [10].

The essential temporal correlation, on the other hand, is not investigated. Instead, the spatio-temporal FER networks
Several adversarial disentanglement works demonstrate that simply separate the input may result in the extracted feature has no meaningful information [24, 23, 42, 43]. The reconstruction of input can explicitly enforce the disentanglement objective.

**Mutual-information** has a long history in unsupervised learning. The infomax principle [25], as prescribed for neural networks, advocates maximizing the MI between network input and output. This can be the fundamental of many ICA algorithms, which can be nonlinear [38] but are often hard to adapt for use with deep networks. Recently, some of the works proposed to achieve unsupervised learning with MI. [39] 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. 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 [26]. In [40], the DeepInfoMax (DIM) is proposed to learn the representations based on both local and global information. In [41], Deep Graph Info-max (DGI) extends this approach to graph-structured data.

Inspired by these works, we are 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 [16, 8]. We note that previous identity-aware FER methods usually explicitly require the identity labels of FER datasets to sample the triplets [16, 8, 3], while the identity label is not common in FER. In contrast, our solution does not relies on the identity label of FER samples, but utilizes the easily available face recognition dataset.

The typical solution for FER is metric learning. Our MI regularization is also related to the triplet loss [21], which maximizes the Euclidean or cosine distance between two identities. With the development of GAN, adversarial training also can be utilized for disentanglement [23, 24]. 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 [16, 8, 3]. We can also choose adversarial training [23, 24] as a baseline to achieve identity elimination in our framework. We note that the adversarial game is notorious for its unstable model collapse [27], while our solution follows a collaborative way.

Several adversarial disentanglement works demonstrate that simply separate the input may result in the extracted feature has no meaningful information [24, 23, 42, 43]. The reconstruction of input can explicitly enforce the disentan-
3. Methodology

We propose to develop an efficient video-based FER framework that operates directly on the compressed domain. The overall framework is shown in Figure 2, which is consisted 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. The dependence of identity and expression representations is then measured using a differentiable mutual information regularized module. Furthermore, when the apex frame is annotated, the complementary constraint can be applied to stabilize the early stage training.

3.1. Modeling Compressed Representations

To illustrate the format of the input video, we choose the MPEG-4 as an example [12]. 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, respectively. 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 \( T^t \in \mathbb{R}^{h \times w \times 2} \).

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

The majority of compression algorithms are programmed primarily to minimize file size, and the encoded format can vary greatly from RGB images in terms of statistical and structural properties. As a consequence, a specially built processing network is needed to manage the compressed format. Considering the structure of residual images \( \Delta^t \) 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 [6, 44, 1]. Practically, we follow the CNN in the typical CNN-LSTM FER structure [44, 6, 1], but with fewer layers to explore the information in \( \Delta^t \). 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 [15] independently concatenate the paired \( \Delta^t \) and \( T^t \) at each time step and predict an action score of each P-frame. The temporal cues and their development patterns are important for the FER task [5]. We simply choose the LSTM in [44] to model the sequential development of residual frames and summarize the information to an 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 [34, 35, 45].

For the I frame with raw image format, we simply use the FaceNet [21] pre-trained on millions of identities [22] as our identity feature extractor \( f_I : \mathbb{R}^{h \times w \times 3} \rightarrow \mathbb{R}^{1024} \) denotes the identity feature. We note that \( Z_E \) and \( z_I \) do not need to have the same dimension for MI Regularization. Several datasets’ FER videos begin with a neutral expression, which may help with 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 \( z_E \).

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 [23, 24]. For example, \( z_E \) can achieve 1) using the conventional C.E. 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 latter objective can be challenging.

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.

\[
MI(z_E; z_I) = \int_{\mathcal{X}} \int_{\mathcal{Z}_E} \int_{\mathcal{Z}_I} \frac{dP_{z_E z_I}}{dP_{z_E} \otimes P_{z_I}} \log \frac{dP_{z_E z_I}}{dP_{z_E} \otimes P_{z_I}}
\]

(1)

where \( z_E \) and \( z_I \) are the random variables follows the distribution \( \mathcal{E} \) and \( I \) respectively. \( P_{z_E z_I} \) indicates the joint prob-

---

X Liu et al.: Accepted to Pattern Recognition
that where ̂ z

Then, we can evaluate the bias corrected gradients (e.g., moving average) [26].

We have three to be minimized objectives, i.e., cross-entropy loss, mutual information and \( \mathcal{L}_1 \) loss, which works the FER branch to maximize the discrepancy between the 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 adversarial training).

Besides, MI is a symmetric measure, while the conditional entropy \( H(\hat{z}_1|z_E) = H(z_1) - I(z_1; z_E) \) optimized in conventional disentanglement works [23, 24] is asymmetric and essentially we should calculate both \( H(\hat{z}_1|z_E) \) and \( H(z_1|z_E) \) as supervision signal [23, 24]. To maximize the discrepancy of \( z_E \) and \( \hat{z}_1 \), we can also apply the adversarial disentanglement solutions [23, 24]. Nevertheless, with the above-mentioned limitations, such methods can be hard to optimize and lead to inferior performance.

3.3. Complementary constraint

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+ [47, 14] consists of a sequence that shifts from the neutral expression to an apex facial expression. The last frame usually is 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 [48], the video frames usually start from the neutral face and develop to the apex around the middle of the video, and returning back to the neutral at the end of the 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 \( \mathcal{L}_2 \) loss.

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

where \( \hat{I}_{\text{Apex}} = \text{Dec}(z_1, z_E) \). The complementary restriction is not necessary for our system since the FER loss is heavily weighted in the FER branch. It requires to maintain sufficient information w.r.t. expression and not easy to have nothing meaningful. However, 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 at the start of a 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 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. Specifically, we update our \( f_E \) and LSTM modules with

\[
\mathcal{L}_{CE} + \alpha MI(\widehat{z_E}; z_F)_n + \beta \mathcal{L}_1
\]

For the MI estimator \( T_{\theta} \), we update it with \( MI(\widehat{z_E}; z_F)_n \). Moreover, the decoder module \( \text{Dec} \) is updated with \( \mathcal{L}_1 \). 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.

Since we are using the neural network \( T_{\theta} \) for mutual information estimation, it is scalable, flexible, and completely trainable via back-propagation. Moreover, the decoder for reconstruction is used for complementary constraints. In contrast, FLF [23, 24] uses a discriminator and a decoder for adversarial disentanglement and complementary constraint. Considering the \( T_{\theta} \) in MIC and discriminator in FLF [23, 24] has a similar structure, there is no significant difference for the network complexity. The mutual information calculation with Eq. 4 has the complexity of \( O(n) \), where \( n \) is the number of sampled data. The linear complexity can be simple for implementation. Conventionally, the fast computation of MI is limited to discrete variables [49]. For the continuous random variables, its complexity is quadratic to the number of samples, which is not desired for a loss function.

### 4. Experiments

In this section, we first detail our experimental setup, present a 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** [47, 14] is referring to the Cohn-Kanade AU-Coded Expression dataset, which is a widely accepted FER benchmark [5, 45]. The video is collected in a restricted environment, in which the participate subjects are facing the recorder with an empty background. The video in CK+ consists of a sequence that shifts from the neutral expression to an apex facial expression. The last frame usually is the apex frame, which has the most strong expression intensity. The expression included in this dataset is anger, contempt, disgust, fear, happiness, sadness, and surprise. There are 327 facial expression videos collected from 118 subjects. Following the previous works, we use subject independent 10-folds cross-validation [5, 3].

Many FER datasets follow a well-defined collection protocol, which usually starts from the neutral face and develops to an expression. Specifically, the image-based FER methods select the last three frames to construct their training and testing datasets. In Fig. 3, we show the compressed and decoded frames in CK+ dataset. Similarly, in MMI [48], the video frames usually start from the neutral face and develop to the apex around the middle of the video, and returning back to the neutral at the end of the 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.

#### Table 1

Experimental results on the CK+ dataset. Note that in order to make the comparison fair, we do not consider image-based and 3D geometry based experiment setting and models [8, 16, 4].*Additional FER+ dataset is used. †Additional body language dataset is used.
Null
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 [44], and achieve better accuracy.

 Besides, [3] 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 tuples usually makes the training not scalable [4], 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 -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 network and the disentanglement with mutual information is feasible. Compared with using the conventional adversarial training based disentanglement [23, 24] as an alternative (i.e., MIC-MI+Adv), our MI regularizer is easy to train and can converge 1.8 times faster in training.

 We can also follow the action recognition method [60] to concatenate the motion vector and residual as the input, and denote as MIC+T’. However, we do not achieve significant performance on all datasets, but the inference speed in testing can be slower. This may be related to the coarse resolution of the motion vector can not well describe the fine-grain muscle movement of the face. More appealingly, the performance of our MIC can be further improved With the ResNet backbone. With the simplified mode variational LSTM-based [44] to denoise identity with \( z_F \). Besides, we can directly use mutual information as the metric of independence. We can see that the residual frame itself can incorporate much less identity information as the metric of independence. We can see that the MI regularization can explicitly remove the identity factors and outperforms the adversarial training [23].

### 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 to 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 networks. There is a good performance for the expression class of fear, happiness, sadness, and surprise. In contrast, the accuracy of expression class anger and disgust is relatively limited. Especially, there is a high degree of confusion between anger and disguise. This may be 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 the 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 a fair comparison. Besides, the audio modality in AFEW can be used to boost the recognition performance [61, 62, 63]. 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 be added on our framework following the multimodal methods [61].

With the simplified mode variational LSTM-based [44]
Table 4
Experimental results on AFEW dataset. *Additional FER+ dataset is used. †Optical flow is used. ††Additional body language dataset is used.

| Method      | Backbone          | Accuracy | Model type | Ave Test |
|-------------|-------------------|----------|------------|----------|
| Undirectional LSTM (2017) [62] | VGGFace | 45.04 | Dynamic | 32fps |
| DenseNet-161 (2018) [64] | VGGFace | 51.44 | Static | - |
| C3D-GRU (2019) [51] |  VGGFace | 51.2 | ✓ | - |
| DSTA (2019)† [65] |  VGGFace | 42.98 | Dynamic | 4fps |
| NST (2020)†† |  VGGFace | 99.69 | Dynamic | - |
| Mode VLSTM (2019) [44] |  VGGFace | 51.44 | Dynamic | 11fps |
| MIC | CNN-LSTM | 53.18 | Dynamic | 34fps |
| MIC-MI | ResNet18-LSTM | 52.62 | Dynamic | 34fps |
| MIC+Adv† [23] |  ResNet18-LSTM | 53.01 | Dynamic | 34fps |
| FAN† [7] |  ResNet18-LSTM | 53.18 | Dynamic | 30fps |
| MIC+FAN |  ResNet18-LSTM | 53.72 | Dynamic | 30fps |

Table 5
Comparison of the cross-dataset results. We use CK+ training set for training and test on MMI testing set.

| Method      | Backbone          | Accuracy | Model type | Ave Test |
|-------------|-------------------|----------|------------|----------|
| Mode variational LSTM [44] | CNN-LSTM | 67.28% | Dynamic | 34fps |
| MIC | CNN-LSTM | 65.85% | Dynamic | 30fps |
| FAN† [7] | ResNet18-LSTM | 66.42% | Dynamic | 30fps |
| MIC+FAN | ResNet18-LSTM | 68.53% | Dynamic | 30fps |

Table 6
Comparison of the cross-dataset results. We use MMI training set for training and test on CK+ testing set.

| Method      | Backbone          | Accuracy | Model type | Ave Test |
|-------------|-------------------|----------|------------|----------|
| Mode variational LSTM [44] | CNN-LSTM | 52.34% | Dynamic | 30fps |
| MIC | CNN-LSTM | 55.85% | Dynamic | 30fps |
| FAN† [7] | ResNet18-LSTM | 54.72% | Dynamic | 30fps |
| MIC+FAN | ResNet18-LSTM | 56.43% | Dynamic | 30fps |

Table 7
Comparison of the cross-dataset results. We use AFEW training set for training and test on MMI testing set for the shared classes.

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 [6].

In addition, we note that the complementary constraint requires the apex frame in training, which is not applicable for the AFEW dataset. We do not apply the reconstruction loss in the AFEW task. Although the requirement of the apex frame imposes some limitations on the training, it does not affect the generality of the testing or implementation of the trained model with CK+ and MMI.

Some of the works propose to improve the image-based FER networks and combine the frame-wise scores for video-based FER [64, 7]. The image-based FER methods [64] achieves high performance, but [64] uses a very deep network DenseNet-161 and pretrained it on the private Situ dataset. Moreover, [64] utilize the sophisticated post-processing. Actually, an intuition of a 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.

[62] uses VGGFace as the backbone of $f_E$ and an RNN model with LSTM units to capture the temporal dynamic cues of the videos. Moreover, [52, 51, 65] 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 decoding processing and needs to handle much more complicated data. With the 6-layer ResNet backbone as an expression feature extractor, the performance of our MIC model can be further improved without an additional FER dataset for pre-training. Overall, the proposed MIC can improve the testing speed by a large margin and can achieve the SOTA accuracy as the previous models.
Table 8
Comparison of the cross-dataset results. We use AFEW training set for training and test on CK+ testing set for the shared classes.

| Method         | Backbone            | Accuracy  |
|----------------|---------------------|-----------|
| Mode variational LSTM [44] | CNN-LSTM            | 64.60%    |
| MIC            | CNN-LSTM            | 67.47%    |
| FAN [7]        | ResNet18-LSTM       | 66.29%    |
| MIC+FAN        | ResNet18-LSTM       | 68.62%    |

Table 9
Facial expression recognition with the extracted identity feature from I frame.

| $\beta$ | CK+     | MMI     | AFEW   |
|---------|---------|---------|--------|
| $z_I$   | 14.29%  | 16.68%  | 14.28% |
| Chance  | 14.29%  | 16.67%  | 14.29% |

4.4. Identity feature extraction

We adopt the pre-trained face recognizer FaceNet [21] to extract the identity factor from the I frame. The feature embedded with the convolutional layers and the first fully connected layer can be robust to a broad range of nuisance factors such as expression, pose, illumination, and occlusion variations, since it achieves high accuracy over millions of identities [22]. To check the expression information in the extracted identity feature $z_I$, we use the feature for expression classification as [23]. The results are shown in 9. We note that achieve zero FER accuracy with $z_I$ does not mean $z_I$ has no information about expression. Instead, approaching the chance probability, i.e., uniform distribution w.r.t. expression classes, indicates the expression is well disentangled from the identity feature $z_I$ with FaceNet.

4.5. Sensitivity analysis

We use $\alpha$ and $\beta$ to balance the MI and complementary constraint terms and choose the best value with grid searching. In Tab. 10, we provide the sensitivity analysis of using different $\alpha$ in three datasets. We can see that we can achieve the best performance on CK+ and MMI with $\alpha = 0.1$. For the AFEW dataset, the performance is relatively stable for $\alpha$ from 0.15 to 0.25. We simply use 0.2 for all of our MIC models on the AFEW dataset.

$\beta$ is used to balance the complementary constraint term, which can be helpful for stabilizing the training. In Tab. 11, both the fixed $\beta$ and linear changing $\beta$ are compared. Decreasing $\beta$ from 1 to 0 for 30 or 50 epochs can usually achieve the best performance.

4.6. Cross-database validation

Since the subjects are different across CK+ and MMI datasets, the cross-dataset evaluation can be used to evidence if the trained model is affected by identity [16]. In Tab. 5, the FER model trained CK+ training sets in the previous experiment is directly implemented to the MMI testing set. We can see that our MIC can outperform the other methods with the same backbone. In Tab. 6, we use MMI as training data and test on CK+ dataset. The proposed MIC outperforms the other methods with the same backbone consistently.

We note that FAN [7] uses ResNet18 as a feature extractor and sequentially pre-trained on MSceleb-1M face recognition dataset and FER+ expression dataset.

In Tab. 7 and Tab. 8, we use the training set of AFEW for training and test on MMI and CK+, respectively. Considering the large domain shift between AFEW and MMI/CK+, there is a significant performance drop. We note that the proposed method can also achieve better performance than its backbones [44, 7].

4.7. Critical discussion and future work

The proposed MIC framework has demonstrated its effectiveness w.r.t. accuracy and testing speed. However, MIC highly relies on the separation of I and P frames in compressed video. Although MPEG-4, H.264, and HEVC formats are widely used, some of the compression solutions do not follow the motion compensation with I and P frames. For example, the YUV format compresses the video by considering the different changes of brightness and chromaticity. Moreover, frame loss can be common in real-world video transfer. How will the frame loss affect the FER performance is underexplored.

Identity can be the most challenging variation for FER [8], while the FER performance can also be affected by pose and illuminations. It can be promising to take the other variations into account.

The recent work [55] proposes to utilize the additional unlabeled face dataset to boost the performance, which can be a powerful means for alleviating the scale issue of video-based FER.

For the large domain gap, e.g., AFEW to MMI/CK+, the domain adaptation methods [67, 68] can be used to achieve better cross-dataset performance.

5. Conclusion

In this paper, we target to explore the facial expression cues directly on the compressed video domain. We are motivated by our practical observation that 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...
Mutual Information Regularized Identity-aware FER in Compressed Video

| α    | 0   | 0.01 | 0.05 | 0.1  | 0.15 | 0.2  | 0.25 | 0.3  | 0.5  | 1    |
|------|-----|------|------|------|------|------|------|------|------|------|
| CK+  | 97.84% | 98.71% | 98.90% | 98.95% | 98.95% | 98.94% | 98.90% | 98.88% | 98.54% | 98.46% |
| MMI  | 80.25% | 81.02% | 81.23% | 81.29% | 81.27% | 81.26% | 81.13% | 81.10% | 81.08% | 81.04% |
| AFEW | 52.62% | 52.92% | 53.04% | 53.15% | 53.18% | 53.18% | 53.16% | 53.11% | 53.07% | 53.01% |

Table 10
Sensitivity analysis of hyperparameter α in CK+, MMI and AFEW datasets.

| β    | CK+ | MMI | AFEW |
|------|-----|-----|------|
| Keep 1 | 97.87% | 80.69% | 52.76% |
| Keep 0.5 | 97.90% | 80.98% | 52.91% |
| Keep 0.1 | 97.92% | 81.02% | 53.05% |
| 1→0:30 | 98.95% | 81.29% | 53.18% |
| 1→0:50 | 98.95% | 81.26% | 53.18% |
| 1→0:100 | 98.72% | 81.17% | 53.02% |
| 1→0:5:30 | 97.85% | 80.96% | 52.88% |

Table 11
Sensitivity analysis of hyperparameter β in CK+, MMI and AFEW datasets. Keep 1 indicates using a fixed β = 1. We denote linear decreasing β from 1 to 0 until the 30th epoch as 1→0:30.

be a good alternative to adversarial training [23] for many disentanglement tasks.

References

[1] W. J. Badar, S. Lee, Y. M. Ro, On-the-fly facial expression prediction using lstm encoded appearance-suppressed dynamics, IEEE Transactions on Affective Computing (2019).
[2] X. Liu, Z. Guo, S. Li, L. Kong, P. Jia, J. You, B. Kumar, Permutation-invariant feature restructuring for correlation-aware image set-based recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4986–4996.
[3] X. Liu, Y. Ge, C. Yang, P. Jia, Adaptive metric learning with deep neural networks for video-based facial expression recognition, Journal of Electronic Imaging 27 (1) (2018) 013022.
[4] X. Liu, B. V. Kumar, P. Jia, J. You, Hard negative generation for identity-disentangled facial expression recognition, Pattern Recognition 88 (2019) 1–12.
[5] S. Li, W. Deng, Deep facial expression recognition: A survey, IEEE Transactions on Affective Computing (2020).
[6] D. H. Kim, W. J. Badar, J. Jang, Y. M. Ro, Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition, IEEE Transactions on Affective Computing 10 (2) (2017) 223–236.
[7] D. Meng, X. Peng, K. Wang, Y. Qiao, Frame attention networks for facial expression recognition in videos, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 3866–3870.
[8] X. Liu, B. Vijaya Kumar, J. You, P. Jia, Adaptive deep metric learning for identity-aware facial expression recognition, in: CVPRW, 2017, pp. 522–531.
[9] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Frouinemy, Y. Dauphin, N. Boulanger-Lewandowski, et al., Emnets: Multimodal deep learning approaches for emotion recognition in video, Journal on Multimodal User Interfaces 10 (2) (2016) 99–111.
[10] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, L. Sigal, Video emotion recognition with transferred deep feature encodings, in: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, 2016, pp. 15–22.
[11] C. Yeo, P. Ahammad, K. Ramchandran, S. S. Sastry, Compressed domain real-time action recognition, in: 2006 IEEE Workshop on Multimedia Signal Processing, IEEE, 2006, pp. 33–36.
[12] I. E. Richardson, H264/mpeg-4 part 10 white paper: prediction of intra macroblocks, Internet Citation, Apr 30 (2003).
[13] D. Le Gall, Mpeg: A video compression standard for multimedia applications, Communications of the ACM 34 (4) (1991) 46–58.
[14] P. Lacey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, 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, 2010, pp. 94–101.
[15] C.-Y. Wu, M. Zaheer, H. Hu, R. Mannathu, A. J. Smola, P. Krähenbühl, Compressed video action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6026–6035.
[16] Z. Meng, P. Liu, J. Cai, S. Han, Y. Tong, Identity-aware convolutional neural network for facial expression recognition, in: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, 2017, pp. 558–565.
[17] A. J. Calder, A. W. Young, Understanding the recognition of facial identity and facial expression, Nature Reviews Neuroscience 6 (8) (2005) 641–651.
[18] K. Ali, C. E. Hughes, All-in-one: Facial expression transfer, editing and recognition using a single network, arXiv preprint arXiv:1911.07050 (2019).
[19] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O’Reilly, Y. Tong, Identity-free facial expression recognition using conditional generative adversarial network, arXiv preprint arXiv:1903.08051 (2019).
[20] A. Dhall, R. Goecke, J. Joshi, K. Sikka, T. Gedeon, Emotion recognition in the wild challenge 2014: Baseline, data and protocol, in: Proceedings of the 16th international conference on multimodal interaction, 2014, pp. 461–466.
[21] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.
[22] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, E. Brossard, The megaface benchmark: 1 million faces for recognition at scale, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4873–4882.
[23] X. Liu, S. Li, L. Kong, W. Xie, P. Jia, J. You, B. Vijaya Kumar, Feature-level frankenstein: Eliminating variations for discriminative recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 637–646.
[24] X. Liu, Y. Chao, J. J. You, C.-C. J. Kuo, B. Vijayakumar, Mutual information regularized feature-level frankenstein for discriminative recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).
[25] R. Linsker, Self-organization in a perceptual network, Computer Vision, 2019, pp. 4986–4996.
Transactions on Image Processing (2020).

[30] G. Sandbach, S. Zafeiriou, M. Pantic, L. Yin, Static and dynamic 3d facial expression recognition: A comprehensive survey, Image and Vision Computing 30 (10) (2012) 683–697.

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

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

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

[34] P. Barros, S. Wermter, Developing crossmodal expression recognition based on a deep neural model, Adaptive behavior 24 (5) (2016) 373–396.

[35] J. Zhao, X. Mao, J. Zhang, Learning deep facial expression features from image and optical flow sequences using 3d cnn, The Visual Computer 34 (10) (2018) 1461–1475.

[36] B. Zhang, L. Wang, Z. Wang, Y. Qiao, H. Wang, Real-time action recognition with deeply transferred motion vector cnns, IEEE Transactions on Image Processing 27 (5) (2018) 2326–2339.

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

[38] A. Hyvärinen, P. Paunen, Nonlinear independent component analysis: Existence and uniqueness results, Neural Networks 12 (3) (1999) 429–439.

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

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

[41] P. Velickovic, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, R. D. Hjelm, Deep graph infomax, arXiv preprint arXiv:1809.10341 (2018).

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

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

[44] W. J. Baddar, Y. M. Ro, Mode variational lstm robust to unseen modes of variation: Application to facial expression recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 3215–3223.

[45] S. Kumawat, M. Verma, S. Raman, 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, 2019, pp. 0–0.

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

[47] T. Kanade, J. F. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in: Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR005800), IEEE, 2000, pp. 46–53.

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

[49] L. Paninski, Estimation of entropy and mutual information, Neural Computation 15 (6) (2003) 1191–1253.

[50] K. Zhang, Y. Huang, Y. Du, L. Wang, Facial expression recognition based on deep evolutionary spatial-temporal networks, IEEE Transactions on Image Processing 26 (9) (2017) 4193–4203.

[51] M. K. Lee, D. Y. Choi, D. H. Kim, B. C. Song, 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, 2019, pp. 1–8.

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

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

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

[55] V. Kumar, S. Rao, L. Yu, Noisy student training using body language dataset improves facial expression recognition, in: European Conference on Computer Vision, Springer, 2020, pp. 756–773.

[56] N. Perveen, D. Roy, C. K. Mohan, Spontaneous expression recognition using universal attribute model, IEEE Transactions on Image Processing 27 (11) (2018) 5575–5584.

[57] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, T. Gedeon, From individual to group-level emotion recognition: Emotiv 5.0. in: ICMI, 2017, pp. 524–528.

[58] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.

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

[60] Z. Shou, X. Lin, Y. Kalantidis, L. Sevilla-Lara, M. Rohrbach, S.-F. Chang, Z. Yan, Dmc-net: Generating discriminative motion cues for fast compressed video action recognition, in: CVPR, 2019, pp. 1268–1277.

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

[62] V. Vielzeuf, S. Pateux, F. Jurie, Temporal multimodal fusion for video emotion classification in the wild, in: ICMI, 2017, pp. 569–576.

[63] C. Lu, W. Zheng, C. Li, C. Tang, S. Liu, S. Yan, Y. Zong, Multiple spatio-temporal feature learning for video-based emotion recognition in the wild, in: Proceedings of the 20th ACM International Conference on Multimedia Interaction, 2018, pp. 646–652.

[64] C. Liu, T. Tang, K. Lv, M. Wang, Multi-feature based emotion recognition for video clips, in: ICMI, 2018, pp. 630–634.

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

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

[67] G. He, X. Liu, F. Fan, J. You, Image2audio: Facilitating semi-supervised audio emotion recognition with facial expression image, in: CVPR Workshops, 2020, pp. 912–913.

[68] X. Liu, X. Liu, B. Hu, W. Ji, F. Xing, J. Lu, J. You, C.-C. J. Kuo, G. E. Fakhri, J. Woo, Subtype-aware unsupervised domain adaptation for medical diagnosis, AAAI (2021).