Robust Emotion Recognition from Low Quality and Low Bit Rate Video: A Deep Learning Approach

Bowen Cheng[†], *Zhangyang Wang*[‡], *Zhaobin Zhang*[♢], *Zhu Li*[♢], *Ding Liu*[†],
Jianchao Yang[§], *Shuai Huang*[♭], *Thomas S. Huang*[†]
† Beckman Institute, University of Illinois at Urbana-Champaign
‡ Department of Computer Science and Engineering, Texas A&M University
♢ Department of Computer Science & Electrical Engineering, University of Missouri, Kansas City
§ Snap Inc, USA ♭ Department of Industrial and Systems Engineering, University of Washington
{bcheng9, dingliu2, t-huang1}@illinois.edu atlaswang@tamu.edu
{zzktb@mail., lizhu@}umkc.edu jianchao.yang@snap.com shuaih@uw.edu

***Abstract*—Emotion recognition from facial expressions is tremendously useful, especially when coupled with smart devices and wireless multimedia applications. However, the inadequate network bandwidth often limits the spatial resolution of the transmitted video, which will heavily degrade the recognition reliability. We develop a novel framework to achieve robust emotion recognition from low bit rate video. While video frames are downsampled at the encoder side, the decoder is embedded with a deep network model for joint super-resolution (SR) and recognition. Notably, we propose a novel *max-mix* training strategy, leading to a single "One-for-All" model that is remarkably robust to a vast range of downsampling factors. That makes our framework well adapted for the varied bandwidths in real transmission scenarios, without hampering scalability or efficiency. The proposed framework is evaluated on the AVEC 2016 benchmark, and demonstrates significantly improved stand-alone recognition performance, as well as rate-distortion (R-D) performance, than either directly recognizing from LR frames, or separating SR and recognition.**

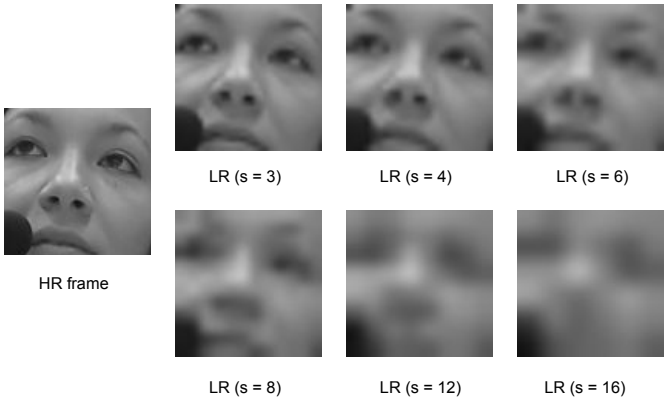


Figure 1. A HR face image (resolution: 96×96) detected from one frame in the dev 8 set of the AVEC 2016 dataset, and its downsampled LR versions with different downsampling factors s: {3, 4, 6, 8, 12, 16} (displayed after bi-cubic interpolation). Note that our proposed approach can substantially improve the emotion recognition performance, for up to $s = 8$.

1. Introduction

Emotion recognition from facial expressions mostly relies on data collected in a highly controlled environment with high resolution (HR) frontal faces. Coupled with the widespread use of smart and wearable devices, emotion recognition techniques have demonstrated the tremendous application value, in tracking human mental status and detecting mental illness, in a less obtrusive way than traditional mental healthcare monitoring approaches [1]. However, with the ever-growing use of wireless multimedia applications, the available network bandwidth is often inadequate to stream HR video. To transmit video contents over limited bandwidth networks, the encoder often compromises the spatial resolution of video frames for reducing the bit rates, by adaptive downsampling of the HR video to low resolution (LR) prior to compression [2]. It yields improved performance than coding with the original full-size video, yet at the expense of degrading quality. In particular, the LR facial images after decompression constitutes a severe challenge for facial expression analysis [3]. Figure 1 displays a few examples after downsampling, which apparently make emotion recognition increasingly difficult, or even impossible.

This paper presents a novel framework to achieve robust and reliable emotion recognition, while keeping the communication load low. At the encoder side, the video frames are adaptively downsampled before compression and transmission, in order to meet the bandwidth requirements. The core innovation of the proposed framework is a jointly optimized scheme of super resolution (SR) and recognition models based on deep learning [4], after decoding. As an important finding, we develop a novel "max-mix" training strategy, and obtain a single deep model that is verified to be robust to a vast range of downsampling factors. The "One-for-All" model is well adapted for the varied bandwidths in practical transmission. Our model demonstrates significantly superior recognition and rate-distortion (R-D) performance, than either directly recognizing from LR frames, or the two-stage pipeline where restoration and recognition are separate. Finally, we point out a few directions, towards which our framework can be further improved.

978-1-5386-0563-9/17/$31.00

2. Related Work

2.1. Emotion Recognition from Facial Expressions

Recognizing human emotion can depend upon gesture, pose, facial expression, speech, behaviors, and even brain signals [5]. In this paper, we mainly discuss emotion recognition from videos that record facial expressions. The seminal work [6] recognized fine-grained changes in facial expression by proposing the Facial Action Coding System (FACS). A large portion of research efforts tried to formulate emotion recognition as a multi-class *classification* problem. The most famous categorization system is the scheme of six "universal" atom emotions [7]: anger, disgust, fear, happiness, sadness, and surprise. Many feature engineering or feature learning approaches have been proposed for the six-emotion classification problem, e.g., [8], [9], [10], [11].

The *regression* formulation is another promising alternative to model the infinite space of possible emotions [12]. A person's emotions were found to be described by a low-dimensional representation. One simple and common choice is to decompose the emotion into two orthogonal and real-valued dimensions: *arousal* and *valence* [13]. Arousal measures how engaged or apathetic a subject appears, while valence measures how positive or negative a subject appears. The arousal-valence representation describes a larger and continuous space of emotions, which the six-emotion scheme only roughly partitions the emotion space into six regions. Moreover, the regression formulation allows for time-continuous, real-valued outputs, which is more realistic for modeling temporal emotion dynamics from video.

Several benchmarks have been constructed for the task of automatic emotion recognition, such as the extended Cohn-Kanade (CK+) dataset [14], and the MMI facial expression database [15]. Following many recent works [16], [17], [18], we develop our emotion recognition model based on the AVEC 2016 [19] dataset, whose data was originally from the RECOLA corpus [20]. Multimodal signals, including audio, video (40 ms binned frames), and physiological signals, were synchronously recorded from 27 subjects. Continuous-time and continuous-valued ratings of arousal and valence were given by human raters. In this paper, we focus on video data only, and choose the valence value as the regression goal for simplicity (same as [18], one of the state-of-the-arts on the same dataset). The proposed method can integrate other data modalities, and can be easily extended to predict arousal and valence values jointly.

2.2. Low Bit Rate Video Transmission with Adaptive Downsampling

For a variety of computer vision tasks where processing server needs to communicate with remotely deployed visual sensors, the communication costs can be prohibitive, especially for applications like city-scale visual surveillance networks, where thousands of high resolution cameras are connected. How to reduce the communication cost in the distributed vision system is an important research issue.

Extensive prior works have shown that downsampling to LR prior to encoding and upsampling after decoding can can reduce the operating cost in bit rate, and with upscaling/super-resolution, can visually beat the video compressed directly at HR using standard codecs with the same number of bits, under insufficient bit rates [2], [21], [22]. In addition, video downsampling has also been a common practice pre-processing for high-level computer vision tasks such as detection and tracking, in order to meet the computational complexity and/or latency requirements, especially on mobile devices with limited processing power [23].

At the decoder side, SR techniques are often adopted as post-processing for enhancing the display quality [24], [25]. If a fixed downsampling ratio during encoding is known, the SR models can be obtained by various example-based training approaches [26], [27], [28], [29]. However, the practical bandwidth might be varied due to network load, congestion and bottleneck situations. [30], [31] pointed out that to achieve the overall optimal R-D performance, the downsampling ratio at the encoder had to be adaptively determined. In that way, the distortions caused by downsampling which reduces the number of pixels transmitted, and coding which introduces quantization noises to the pixels transmitted, could be balanced. As a result, the SR post-processing at the decoder side has to effectively cope with varied downsampling factors. One straightforward but expensive solution is to utilize an ensemble of SR models, each of which is trained dedicatedly for one downsampling factor. A more cost-effective option is to seek a single "one-for-all" SR model, whose performance keeps robust over a useful range of low resolutions. Up to our best knowledge, its viability has not been examined yet.

2.3. Low-Resolution Visual Recognition

Empirical studies [32], [33] in face recognition proved that a minimum face resolution between 32×32 and 64×64 is required for most stand-alone recognition algorithms, whose performance would be much degraded when applied with even lower resolutions [34], [35]. In the emotion recognition literature, most existing methods assumed the availability of HR frontal faces. [3] first investigated the effects of different image resolutions for facial expression analysis. The author concluded that while the performance difference was negligible when the head region resolution was 72×96 or higher, the recognition turned growingly unreliable when head region resolution was lower than 36×48. It is thus desirable to obtain more robust features for LR images and low-intensity expressions [8]

When dealing with LR subjects, the traditional two-stage pipeline tried to first apply SR algorithms before perform recognition tasks. Recently, the SR performance has been noticeably improved, with the aid of deep network models [36]. However, the recovered HR images inevitably over-smoothened details. More importantly, such a straightforward approach yields the sub-optimal performance: the artifacts introduced by the reconstruction process will undermine the final recognition. [37] presented a close-the-loop approach of image restoration and recognition, based on the assumption that the degraded image, if correctly restored, will also have a good identifiability. [38] advanced

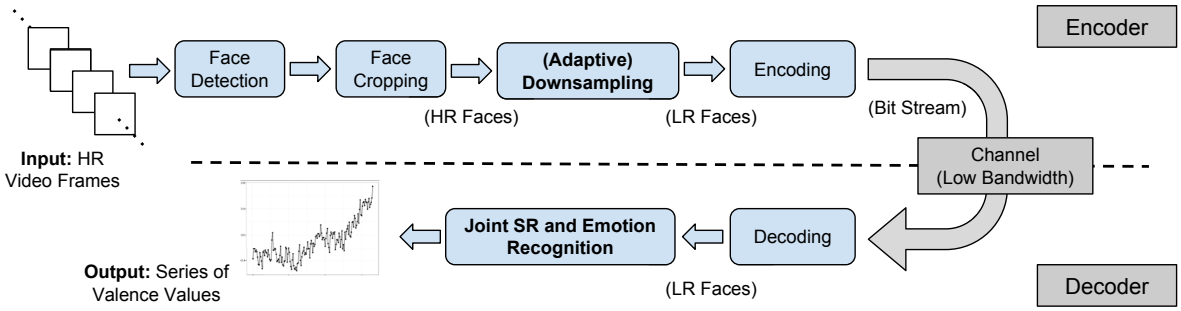


Figure 2. Pipeline of the proposed framework. The intermediate outputs are also annotated along with the pipeline.

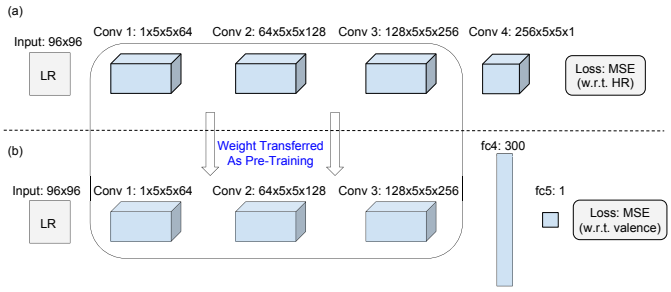


Figure 3. The network architectures for: (a) SR fully convolutional network (SR-FCN); (b) Convolutional neural network (CNN) for joint SR and emotion recognition.

the methodology using a deep network trained from end to end, and observed the possibility of robust object recognition even when the region of interests (ROI) was smaller than 16×16 pixels. However, it remains to be an open issue how much the performance degradation can be remedied in the same way for emotion recognition.

3. Technical Approach

3.1. System Overview

The pipeline of the proposed framework is illustrated in Figure 2. We assume that face detection and cropping has been accomplished at the encoder side as pre-processing. Only the cropped faces are to be downsampled, compressed and transmitted to the decoder side [39]. After decoding, the joint SR and emotion recognition module simultaneously enhances the spatial resolution and predicts the per-frame valence value, using an end-to-end deep network, which will be detailed in the next section. The system outputs a time series of predicted valence values.

We do not discuss how to adaptively control the downsampling factors as per the communication needs, which has been well studied in previous video coding and wireless communication literature [30], [31]. Instead, we aim to make the decoder robust to a wide range of varied downsampling factors that the encoder might adopt.

3.2. Joint SR and Emotion Recognition

Figure 3 (b) depicts the convolutional neural network (CNN) architecture for joint SR and emotion recognition, which mostly inherits the CNN+D structure in [18]. The target CNN is fed with LR video frames. It has 3 convolutional layers consisting of 64, 128, and 256 filters respectively, each of size 5×5. The first two layers are followed by 2×2 max pooling while the third layer is followed by quadrant pooling. Followed is a fully-connected layer with 300 hidden units, regularized by dropout with probability 0.5. ReLU neuron is adopted for all. A linear regression layer estimates the valence values, under the mean squared error (MSE) loss function.

As pointed out by [38], training a CNN-based recognition model over LR images is usually not robust and prone to overfitting, due to the severe information loss. On the other hand, a CNN trained on HR images will also witness degraded performance when tested on LR images, due to the domain mismatch. Our main intuition is to regularize and enhance the CNN feature extraction, by pre-training the first several convolutional layers using a SR sub-model, which reconstructs HR images from LR counterparts.

A 4-layer SR fully convolutional network (SR-FCN) is first constructed, as in Figure 3 (a). Its first three layers are configured the same as the first three layers of the target CNN, while the fourth layer reconstructs the input image from the output feature maps of the third layer. SR-FCN is trained in an unsupervised way to reconstruct the HR frames from LR inputs, under the MSE loss as well. Note that it is different from the target CNN that regresses LR frames to valence values. After that, its first three layers are exported to initialize the first layers of the target CNN. Starting from this SR-based partial initialization, the CNN is then jointly tuned for the emotion recognition task, from end to end.

3.3. Max-Mix Training for An One-for-All Model

Almost all data-driven SR approaches [26], [36] as well as some latest low-resolution recognition works [37], [38] assume one identical downsampling factor between training and testing. A SR model is only dedicated to coping with one downsampling factor. It is more desirable to train a *"One-for-All"* model, since it is robust to the vast range of downsampling factors caused by the varied transmission bandwidths, without incurring any scalability or efficiency issue. Given a range of possible downsampling factors, we propose the *max-mix* training: first pre-training SR-FCN with LR-HR pairs generated with the *maximum downsam-*

	HR	s = 3				s = 4			
		LR-3	Non-Joint-3	Joint-3	Joint-OA	LR-4	Non-Joint-4	Joint-4	Joint-OA
RMSE	0.146	0.142	**0.121**	0.132	0.129	0.155	**0.123**	0.127	0.131
CC	0.430	0.392	0.363	0.396	**0.399**	0.381	0.354	0.380	**0.391**
CCC	0.325	0.302	0.293	0.323	**0.328**	0.283	0.281	0.319	**0.327**
	HR	s = 6				s = 8			
		LR-6	Non-Joint-6	Joint-6	Joint-OA	LR-8	Non-Joint-8	Joint-8	Joint-OA
RMSE	0.146	0.149	0.128	**0.127**	0.134	0.161	**0.129**	0.134	0.130
CC	0.430	0.300	0.344	0.325	**0.375**	0.323	0.317	0.320	**0.358**
CCC	0.325	0.263	0.280	0.274	**0.309**	0.238	0.265	0.266	**0.285**
	HR	s = 12				s = 16			
		LR-12	Non-Joint-12	Joint-12	Joint-OA	LR-16	Non-Joint-16	Joint-16	Joint-OA
RMSE	0.146	0.143	0.126	**0.125**	0.132	0.137	**0.124**	0.137	0.132
CC	0.430	0.291	0.287	0.246	**0.308**	**0.316**	0.244	0.219	0.273
CCC	0.325	0.224	**0.235**	0.204	0.223	**0.212**	0.191	0.192	0.172

TABLE 1. THE OVERALL RMSE, CC AND CCC COMPARISONS AT DIFFERENT FACTORS s (BEST RESULTS IN EACH CASE ARE IN BOLD).

pling factor, followed by fine-tuning the CNN model, on a mixture of LR frames that are generated from HR frames using *the range of all downsampling factors*[1]. As verified by our experiments, the resulting CNN is able to achieve even better performance, than dedicatedly trained SR models at a specific downsampling factor.

4. Experiments

For all AVEC video data, we first convert color frames to gray-scale, and crop the face from each video frame using the given bounding box. All face regions are then normalized to 96×96 pixels, and are treated as the HR subjects to be downsampled, compressed and transmitted. We generate LR frames using a range of downsampling factors s: [3, 4, 6, 8, 12, 16]. Such a range is intentionally set to be vast: while s = 3 causes only mild degradations, s = 16 leads to 6×6 facial regions whose expressions are unlikely to be identified even by human viewers.

All CNNs were trained using stochastic gradient descent with batch size of 128, momentum of 0.9, and weight decay of 5×10^{-4}. We apply mean subtraction and contrast normalization prior to passing each face image through the CNN. We train the SR-SCN for 30,000 iterations, using a constant learning rate of 0.01 is used, and . To fine-tune the target CNN, a learning rate of 0.001 is used for the first three pre-trained layers, and the remaining layers are initialized randomly and trained with a learning rate of 0.01: the learning rates are both divided by 10 when we observe that the validation set performance stops to improve.

We use the AVEC development set of 9 sequences as our testing set. Three metrics are measured for the emotion recognition performance [19]: (i) Root Mean Square Error (RMSE); (ii) Pearson Correlation Coefficient (CC); and (iii) Concordance Correlation Coefficient (CCC), which combines CC with the RMSE between the mean of the two compared time series. A good recognition result will likely favor *lower RMSE*, as well as *higher CC and CCC*. Note that CCC is the *most reliable* measure among the three, and was thus used to choose AVEC competition winners.

4.1. Performance Evaluation and Analysis of Low-Resolution Emotion Recognition

We consider the following comparison methods:

- **HR:** a CNN baseline trained and tested on HR data.
- **LR-s:** a CNN baseline trained and tested on LR data, with the downsampling factor s.
- **Non-Joint-s:** a SR-FCN is first trained to up-scale LR frames to HR. A separate fully-connected neural network is then trained to regress predict valence values from up-scaled HR images. The SR and emotion recognition modules are not jointly tuned.
- **Joint-s:** the joint SR and emotion recognition model described in Section 3.2, trained dedicatedly for a specific downsampling factor s.
- **Joint-OA:** the joint SR and emotion recognition model, training with the max-mix strategy.

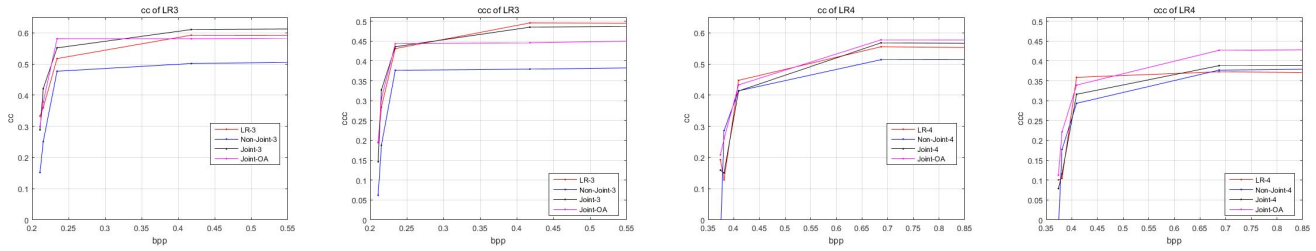
For fair comparison, we carefully ensure all models to have the same amount of parameters. Table 1 presents the overall RMSE, CC and CCC comparison results on the AVEC development set [2]. Comparison HR and LR-s certifies the notable impact of low resolution on emotion recognition.

If we look at RMSEs only, then non-joint methods achieve best in almost all cases (even better than HR). However, RMSE results display little consistency with CC/CCCs, implying that RMSE may not be a reliable measure. For s = 12 and 16, little improvement seems attainable over the LR baselines, since all recognizable information are almost lost at such low resolutions (see Fig. 1). For s = 3, 4, 6 and 8, the CC and CCC results are fairly consistent: the recognition benefits from joint training in most cases. What is more, the Joint-OA model consistently outperforms Joint-s, with the largest margins of 0.050 (CC) and 0.035 (CCC) at s = 6. With surprise, we notice that for s = 3, 4, the Joint-OA results even slightly surpass HR in terms of CCC.

Two questions arise naturally: (1) why joint training can help; and (2) why a "distracted" Joint-OA model can

1. In our experiments, we find that mixing all LR frames of s = [3, 4, 6, 8, 12, 16] does not lead to the optimal performance. We conjecture that"bad" s values such as 12, 16 lead to un-recognizable LR samples that perturb training. Instead, we mix a "reasonable" range of LR samples of s = [3, 4, 6] for fine-tuning. It is the default way to obtain the Joint-OA model, and is verified to be better than fine-tuning with any single s.

2. We follow [18] to first concatenate all nine sequences into one long sequence, and then compute its RMSE/CC/CCC as the overall results.



(a) s = 3, CC (b) s = 3, CCC (c) s = 4, CC (d) s = 4, CCC

Figure 4. The CC and CCC comparisons at different QPs, with s = 3 and 4.

beat "dedicated" Joint-s models? *For Question 1*, the SR hallucinated details help discover subtle features, which are otherwise prone to be overlooked in LR frames [38]. However, the restoration-driven pre-training non-selectively enhances all visual details, which may also include artifacts that hamper recognition. The joint tuning step introduces extra information (the valence values) to reinforce the learning of more task-related features, while suppressing other unrelated components. *For Question 2*, we conjecture that pre-training SR-FCN with maximum s helps its low-level filters to capture more robust mappings, boosting the (implicit) feature enhancement. Further, the mixture s fine-tuning may correspond to re-scaling training data, which is a popular type of data augmentation for classification tasks [4] and helps learn scale-invariant features.

4.2. Rate-Distortion Performance Comparison

For bandwidth constrained applications, achieving robust facial expression recognition from low bit rate video can be attractive for many security and surveillance applications. The problem is coupled with the low resolution sensor problem, but has its own peculiar challenges. In addition to the loss of pixels from resolution limitations, video coding may also introduce quantization errors that can affect the emotion recognition performance. Indeed, compression of visual features for visual recognition has been an active research topic with many interesting results for key point feature compressions [40], [41].

In this experiment, we encode the actively downsampled testing video at different quality-rate levels to mimic real world transmissions, where video is usually coded subject to a rate constraint. We then fed the decoded videos to the joint SR and recognition models (same as Section 4.1 *without re-training*), and calculate the CC and CCC results. The observations in Figure 4 are mostly consistent with the uncompressed case, showing our models' robustness to coding qualities. For s = 3, Joint-OA gains more advantages with larger quantization parameters (QPs)[3] , while for s = 4 Joint-OA outperforms other methods for most bit rates. For the Rate-Distortion (RD) operating range with good to excellent visual quality, the loss of recognition performance is negligible. As coding-introduced distortion becomes more pronounced at larger QPs, the recognition starts to suffer.

3. the smaller the QP is, the better the reconstruction quality would be.

With s = 3, 4, the CC and CCC starts to saturate for QPs smaller than 24, which operate at approximately 0.24 bits per pixel (bpp) for s = 3, and 0.4 bpp for s = 4. The loss of coding efficiency in LR4, compared to LR3, is due to the fixed overhead from video coding headers and structures, that is shared among all pixels. The efficiency decreases as the number of the pixels is reduced. Notice that the pixels fed into the recognition algorithm are 8-bit. This compression is indeed effective on top of the active downsampling in conserving the bandwidth.

In summary, actively downsampling reduces the number of pixels to be transmitted, while coding with larger QPs enforces heavier quantization of the pixels remaining. Both will contribute to saving the bandwidth, and there exists an interesting tradeoff in-between.

5. Conclusion and Discussion

This paper presents a novel framework for robust emotion recognition from low bit rate video, and demonstrates its promising performance as well as strong robustness to both pixel reduction and pixel quantization. There is apparent room for its further performance improvement. From the *system perspective*, we expect to incorporate more building blocks (e.g., the video encoding and decoding steps) into the joint optimization scheme, and make the pipeline in Figure 2 more end-to-end. From the *model perspective*, so far we have not utilized any temporal information for video-based recognition. The previous work [17], [18] exploited recurrent neural networks to capture the temporal coherence, and obtained additional performance gains. Since adjusting the *temporal resolution* (a.k.a., frame rate) [42] is also a common means to reduce video bit rates, our future work may also extend to adaptive temporal downsampling, followed by temporal-spatial joint video SR and recognition. Finally, as we observe that CC/CCC are evidently better evaluation metrics than RMSE, it is a noteworthy option to train our emotion recognition model under CC/CCC-based loss functions rather than the current MSE loss.

Acknowledgments

Bowen Cheng, Ding Liu and Thomas Huang's research works are supported in part by US Army Research Office grant *W911NF-15-1-0317*. The authors sincerely acknowledge the valuable efforts of the AVEC challenge organizers [19]. The authors would also like to acknowledge the helpful discussions with Dr. Pooya Khorrami and Dr. Thomas Paine.

References

[1] D. Zhou, J. Luo, V. M. Silenzio, Y. Zhou, J. Hu, G. Currier, and H. A. Kautz, "Tackling mental health by integrating unobtrusive multimodal sensing." in *AAAI*, 2015, pp. 1401–1409.

[2] V.-A. Nguyen, Y.-P. Tan, and W. Lin, "Adaptive downsampling/upsampling for better video compression at low bit rate," in *ISCAS*. IEEE, 2008, pp. 1624–1627.

[3] Y.-l. Tian, "Evaluation of face resolution for expression analysis," in *IEEE CVPRW*, 2004.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[5] Z. Wang, D. Liu, S. Chang, F. Dolcos, D. Beck, and T. Huang, "Image aesthetics assessment using deep chatterjee's machine," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 941–948.

[6] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE TPAMI*, 2001.

[7] R. Krause, "Universals and cultural differences in the judgments of facial expressions of emotion," *Journal of Personality and Social Psychology*, vol. 5, no. 3, pp. 4–712, 1987.

[8] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[9] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of ACM International conference on multimodal interaction*, 2013, pp. 543–550.

[10] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *IEEE CVPR*, 2014.

[11] P. Khorrami, T. Paine, and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *Proceedings of the IEEE ICCV Workshops*, 2015, pp. 19–27.

[12] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, 1977.

[13] E. A. Kensinger, "Remembering emotional experiences: The contribution of valence and arousal," *Reviews in the Neurosciences*, 2004.

[14] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *CVPRW*. IEEE, 2010, pp. 94–101.

[15] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Workshop on EMOTION: Corpora for Research on Emotion and Affect*, 2010.

[16] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 65–72.

[17] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proceedings of ACM International Conference on Multimodal Interaction*, 2015, pp. 467–474.

[18] P. Khorrami, T. Le Paine, K. Brady, C. Dagli, and T. S. Huang, "How deep neural networks can improve emotion recognition on video data," in *ICIP*. IEEE, 2016, pp. 619–623.

[19] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[20] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *IEEE FG*, 2013, pp. 1–8.

[21] A. M. Bruckstein, M. Elad, and R. Kimmel, "Down-scaling for better transform compression," *IEEE TIP*, 2003.

[22] W. Lin and L. Dong, "Adaptive downsampling to improve image compression at low bit rates," *IEEE TIP*, vol. 15, no. 9, pp. 2513–2521, 2006.

[23] C. Chen, W. Ji, S. Rho, B.-W. Chen, and Y. Chen, "Evaluate mobile video quality in hybrid spatial and temporal domain," *Multimedia Tools and Applications*, vol. 75, no. 22, pp. 14 507–14 524, 2016.

[24] S. Ma, L. Zhang, X. Zhang, and W. Gao, "Block adaptive super resolution video coding," *Advances in Multimedia Information Processing-PCM 2009*, pp. 1048–1057, 2009.

[25] M. Shen, P. Xue, and C. Wang, "Down-sampling based video coding using super-resolution technique," *IEEE TCSVT*, 2011.

[26] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE TIP*, 2010.

[27] Z. Wang, Y. Yang, Z. Wang, S. Chang, J. Yang, and T. S. Huang, "Learning super-resolution jointly from external and internal examples," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4359–4371, 2015.

[28] Z. Wang, Y. Yang, Z. Wang, S. Chang, W. Han, J. Yang, and T. Huang, "Self-tuned deep super resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 1–8.

[29] D. Liu, Z. Wang, B. Wen, J. Yang, W. Han, and T. S. Huang, "Robust single image super-resolution via deep networks with sparse prior," *IEEE Transactions on Image Processing*, 2016.

[30] R.-J. Wang, C.-W. Huang, and P.-C. Chang, "Adaptive downsampling video coding with spatially scalable rate-distortion modeling," *IEEE TCSVT*, 2014.

[31] J. Dong and Y. Ye, "Adaptive downsampling for high-definition video coding," *IEEE TCSVT*, vol. 24, no. 3, pp. 480–488, 2014.

[32] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. J. Phillips, "A meta-analysis of face recognition covariates," in *IEEE BTAS*, 2009.

[33] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE TPAMI*, 2008.

[34] P. H. Hennings-Yeomans, S. Baker, and B. V. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *CVPR*. IEEE, 2008.

[35] W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *IEEE TIP*, 2012.

[36] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*, 2014.

[37] H. Zhang, J. Yang, Y. Zhang, N. M. Nasrabadi, and T. S. Huang, "Close the loop: Joint blind image restoration and recognition with sparse representation prior," in *ICCV*. IEEE, 2011, pp. 770–777.

[38] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[39] F. A. Andaló, O. A. Penatti, and V. Testoni, "Transmitting what matters: Task-oriented video composition and compression," in *Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2016, pp. 72–79.

[40] A. Nagar, Z. Li, G. Srivastava, and K. Park, "Akula–adaptive cluster aggregation for visual search," in *IEEE DCC*, 2014, pp. 13–22.

[41] X. Xin, Z. Li, and A. K. Katsaggelos, "Laplacian embedding and key points topology verification for large scale mobile visual identification," *Signal Processing: Image Communication*, 2013.

[42] Z. Yu, H. Li, Z. Wang, Z. Hu, and C. W. Chen, "Multi-level video frame interpolation: Exploiting the interaction among different levels," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 7, pp. 1235–1248, 2013.