Enhancing Continuous Emotion Recognition via Visually Diverse Frame Selection

Soki Saigo
Graduate School of FSE,
Waseda University
Tokyo, Japan
sokeen@moegi.waseda.jp

Taiga Hayami
Graduate School of FSE,
Waseda University
Tokyo, Japan
hayatai17@fuji.waseda.jp

Hiroshi Watanabe
Graduate School of FSE,
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—Recent advances in artificial intelligence and computer vision have spurred growing interest in understanding human behavior and emotional states from a third-person perspective. In real-world applications such as security, public guidance, and AI-based scene understanding, it is often necessary to infer emotional states without direct interaction, relying instead on surrounding contextual and interpersonal cues. However, many existing studies rely solely on facial expressions and fail to capture emotions embedded in complex social or situational contexts. Although prior research has attempted to incorporate contextual information, the use of temporal dynamics remains limited. To address this, we propose a method that dynamically selects frames based on inter-frame visual dissimilarity. Experimental results show that our approach effectively captures temporal transitions and enhances the accuracy of continuous emotion recognition in third-person scenarios. Evaluations on the VEATIC dataset, which provides valence and arousal annotations in realistic and socially grounded settings, demonstrate that our method outperforms the baseline that uses uniformly sampled frame sequences.**

***Index Terms*—Continuous emotion recognition, frame selection, context modeling, valence-arousal estimation.**

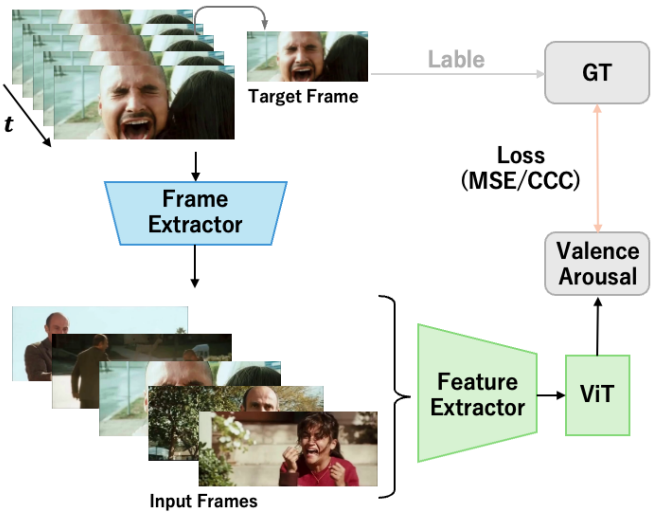

Fig. 1. Overview of the proposed emotion recognition model. The model takes five temporally and visually diverse frames as input and predicts continuous valence and arousal.

I. Introduction

Video-based continuous emotion recognition has gained increasing attention in areas such as surveillance, public guidance, and non-verbal human-computer interaction. In third-person scenarios in particular, emotional state estimation must be performed without relying on direct interaction or speech. Instead, models must infer emotional cues from visual signals embedded in the scene, such as facial expressions, body posture, environmental context, and interpersonal dynamics. To facilitate research in this direction, context-rich datasets such as VEATIC [1] have been introduced. VEATIC provides video clips drawn from real-world, socially grounded settings, offering a valuable benchmark for emotion recognition from a third-person perspective.

Despite these advances, most existing methods, including those evaluated on VEATIC, continue to rely predominantly on facial features and employ frame sampling strategies based on fixed temporal intervals. These uniformly sampled frames are typically fed into deep learning models without consideration of their visual relevance or semantic diversity. Consequently, the input sequences often exhibit high redundancy, making it difficult for models to capture meaningful emotional transitions or subtle contextual shifts.

To overcome this limitation, we propose a novel frame selection strategy based on inter-frame visual dissimilarity. For each target frame, our method dynamically selects visually distinct frames from both the past and future temporal windows. These frames are chosen based on either pixel-wise differences or the structural similarity index (SSIM), and are temporally ordered to form a five-frame input set. In contrast to conventional approaches that emphasize uniform temporal continuity, our method purposefully incorporates frames that are complementary or contrasting in visual content, thereby enriching contextual understanding. This enables the model to better capture emotional dynamics and nuanced situational changes, ultimately improving its generalizability and interpretability in third-person emotion recognition tasks.

II. Related Work

A. Emotion Recognition Methods

Early studies on continuous emotion recognition primarily employed recurrent neural networks (RNNs) and convolutional

979-8-3315-2415-9/25/$31.00 ©2025 IEEE

neural networks (CNNs) to model temporal variations in facial expressions [2]. For example, a combined CNN–LSTM architecture [3] was proposed to capture the dynamic evolution of emotional states over time. Additionally, 3D-CNN-based approaches [4] have been utilized to simultaneously extract spatial and temporal features for emotion prediction. In recent years, Transformer-based architectures have gained traction due to their strong capability to model long-range dependencies. Vision Transformer (ViT)-based models [5], for instance, have achieved competitive performance by learning frame-level temporal transitions. Moreover, Transformers augmented with attention mechanisms [6] have proven effective in recognizing subtle affective patterns, further advancing the field.

Another line of research has addressed frame selection strategies to improve efficiency and reduce redundancy. Approaches based on information-theoretic saliency [7] and multi-scale deep learning [8] have been applied to select informative frames from facial videos. However, most of these methods focus on discrete emotion classification and have not been extensively adapted for continuous emotion estimation.

In contrast to prior work, our method introduces a dynamic frame selection approach that selects visually dissimilar frames around a target frame to better capture contextual transitions. While the VEATIC baseline selects uniformly spaced frames, our method deliberately constructs input sequences with visual diversity. This strategy is particularly beneficial for third-person scenarios, where facial cues may be limited or unavailable, and peripheral contextual information becomes essential for affective inference.

B. Emotion Recognition Datasets

To support continuous emotion estimation, various benchmark datasets have been developed. AFEW-VA [9] provides valence and arousal labels for video clips from movies, but the dataset focuses primarily on facial regions and lacks diverse contextual information. Likewise, Aff-Wild and Aff-Wild2 [10] offer emotion annotations in naturalistic settings but are also centered on facial expressions.

In contrast, VEATIC provides continuous valence and arousal annotations for full-body videos derived from films and documentaries. Beyond facial expressions, it includes rich contextual elements such as body posture, background scenery, and interpersonal interactions. These features make VEATIC particularly suitable for third-person affective understanding, enabling the development and evaluation of models that rely on non-verbal cues beyond facial features. This aspect is often underrepresented in datasets that focus primarily on facial information.

III. Proposed Method

We propose a model for continuous emotion recognition that processes temporally and visually diverse frame sets to enhance affect estimation from third-person perspective video data. As illustrated in Fig. 1, the proposed model consists of two main components: a ResNet-50 backbone that extracts spatial features from each input frame, and a Vision Transformer (ViT) that models temporal dependencies and outputs continuous valence and arousal values.

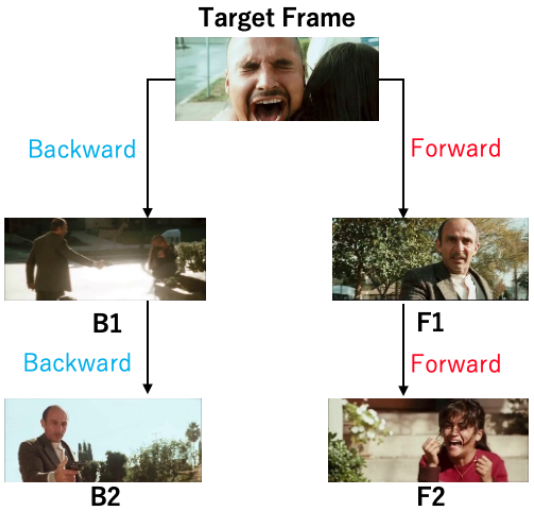

Fig. 2. Dissimilarity-based frame selection process. For each target frame, two past and two future frames are selected based on pixel-wise dissimilarity.

To construct the input for this model, we introduce a dynamic frame selection strategy based on inter-frame visual dissimilarity. Instead of uniformly sampling consecutive frames, our method selects frames that are visually distinct from a given target frame, as illustrated in Fig. 2. This enables the model to capture contextual cues and dynamic scene transitions that may be overlooked by traditional sampling approaches.

For each target frame, we perform directional searches into both the past and the future to locate frames whose similarity to the target is below a predefined threshold. Each discovered frame becomes an anchor for an additional recursive search in the same direction. The final five-frame set includes the target frame, two visually dissimilar frames from the past, and two from the future, arranged in chronological order. This configuration not only captures temporal transitions but also reflects visual contrasts surrounding the emotional moment. Moreover, we adopt this five-frame input setting to ensure consistency with the VEATIC benchmark protocol, which also evaluates models based on sequences of five frames.

The similarity between two RGB frames I_1 and I_2 is computed as:

$$S(I_1, I_2) = 1 - \frac{1}{255} \cdot \text{mean}(|I_1 - I_2|), \tag{1}$$

where the pixel-wise absolute difference is averaged across all spatial and color channels, and normalized by 255.

To optimize the model, we adopt a composite loss function that accounts for both frame-wise accuracy and sequence-level consistency. Following the benchmark formulation in the VEATIC dataset, the total loss is defined as:

$$\mathcal{L} = \frac{1}{2} \left(\mathcal{L}^{\text{valence}} + \mathcal{L}^{\text{arousal}} \right), \tag{2}$$

where each dimension-specific loss $\mathcal{L}^{(*)}$ is a weighted sum of the concordance correlation coefficient (CCC) loss and the mean squared error (MSE):

$$\mathcal{L}^{(*)} = \mathcal{L}^{(*)}_{\text{CCC}} + \lambda \cdot \mathcal{L}^{(*)}_{\text{MSE}}, \tag{3}$$

where $* \in \{\text{valence}, \text{arousal}\}$, and λ is a balancing parameter (set to $\lambda = 0.1$ in our experiments).

The CCC loss is based on the concordance correlation coefficient ρ_c, and is defined as:

$$\mathcal{L}_{\text{CCC}} = 1 - \rho_c, \tag{4}$$

with ρ_c computed as:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{5}$$

where: - μ_x, μ_y: the means of predictions and ground truth, - σ_x, σ_y: the standard deviations, - ρ: the Pearson correlation coefficient between predicted values x and ground truth values y.

The MSE component is defined as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{T}\sum_{t=1}^{T}(x_t - y_t)^2, \tag{6}$$

where T is the number of frames in a sequence, and x_t, y_t are the predicted and ground truth values at time step t, respectively.

This formulation encourages the model to make accurate predictions at each frame while maintaining overall temporal consistency across the video.

IV. Experiment

We evaluate the effectiveness of the proposed method using the VEATIC dataset, which also serves as the baseline method. The following evaluation metrics are used: Concordance Correlation Coefficient (CCC), Pearson Correlation Coefficient (PCC), Root Mean Squared Error (RMSE), and Sign Agreement Metric (SAGR). All models share the same architecture, and the weighting parameter λ in the loss function is fixed to 0.1.

SAGR is defined as follows:

$$\text{SAGR} = \frac{1}{N}\sum_{i=1}^{N}\delta(\text{sign}(\hat{y}_i), \text{sign}(y_i)), \tag{7}$$

where $\hat{y}_i$ and y_i denote the predicted and ground truth labels of the i-th sample, respectively, and δ is the Kronecker delta function, which equals 1 when the predicted and ground truth signs agree and 0 otherwise.

As baselines, we implement VEATIC's uniform downsampling method that samples five consecutive frames at fixed temporal intervals ($k = 5$, 25, 50). In contrast, our proposed method selects five contextually diverse frames based on visual dissimilarity with respect to a target frame. We evaluate two types of similarity metrics: (1) pixel-wise difference with a threshold of 0.80, and (2) SSIM (Structural Similarity Index Measure) with thresholds 0.75, 0.80, and 0.85. SSIM is a perceptual metric that quantifies image quality degradation based on changes in luminance, contrast, and structural information. While it is useful for assessing visual similarity, it may not effectively capture the subtle temporal variations essential for emotion estimation.

TABLE I
Quantitative performance comparison of baseline and proposed methods on the VEATIC benchmark

Dimension	Method	CCC	PCC	RMSE	SAGR
Valence	VEATIC (k=5)	0.609	0.644	0.303	0.789
	VEATIC (k=25)	<u>0.624</u>	0.670	0.293	**0.798**
	VEATIC (k=50)	0.609	0.655	0.301	0.785
	Ours (Diff@0.80)	**0.687**	**0.750**	**0.258**	<u>0.797</u>
	Ours (SSIM@0.75)	0.606	0.688	0.285	0.769
	Ours (SSIM@0.80)	0.612	<u>0.691</u>	<u>0.282</u>	0.771
	Ours (SSIM@0.85)	0.599	0.679	0.288	0.766
Arousal	VEATIC (k=5)	0.630	0.668	0.210	0.779
	VEATIC (k=25)	<u>0.641</u>	0.684	0.202	0.768
	VEATIC (k=50)	0.622	0.653	0.214	0.764
	Ours (Diff@0.80)	**0.685**	**0.746**	**0.182**	**0.804**
	Ours (SSIM@0.75)	0.622	<u>0.693</u>	0.206	<u>0.785</u>
	Ours (SSIM@0.80)	0.608	0.692	0.205	0.772
	Ours (SSIM@0.85)	0.607	0.691	<u>0.200</u>	0.780

Table I shows the quantitative comparison. Our method using pixel difference with a threshold of 0.80 outperforms all baselines and SSIM-based variants across all metrics. Although SSIM variants show moderate improvements over the VEATIC baselines, they fail to match the performance of the pixel-difference-based method. We attribute this to the sensitivity of SSIM-based frame selection to threshold settings and its limited ability to capture dynamic emotional cues. In particular, visual comparisons in Fig. 3 demonstrate that SSIM-selected frames tend to have minimal changes compared to the target frame, resulting in suboptimal temporal diversity. We also observed that in the case of SSIM-based frame selection, the threshold setting had a significant impact on model performance, indicating a need for more fine-grained optimization. Furthermore, to achieve more semantically and perceptually effective frame selection, it would be promising to incorporate similarity measures based on deep features or to explore selection strategies within feature embedding spaces derived from CNNs.

To further assess the learning setup, we compare joint and separate training of valence and arousal. Table II shows that joint training consistently outperforms separate models across all metrics. This implies that modeling shared temporal dynamics between valence and arousal fosters richer and more robust emotional representations.

These findings indicate that selecting visually diverse frames enhances emotional context modeling more effectively than uniformly sampled frames. Additionally, joint training of valence and arousal promotes capturing shared temporal patterns, yielding superior performance across all evaluation metrics.

Fig. 3. Visual comparison of input frames selected by different methods for sample 0499. Each row represents a different method: from top to bottom: Diff@0.80, SSIM@0.75, SSIM@0.80, and SSIM@0.85. Each column corresponds to a different time step: from left to right: B2, B1, Target Frame, F1, F2. Diff@0.80 selects more diverse frames with greater temporal change, while SSIM-selected frames show high similarity with the target, indicating suboptimal diversity. The frame indices are overlaid in each image for reference.

TABLE II
PERFORMANCE COMPARISON BETWEEN SEPARATE AND JOINT TRAINING (DIFF@0.80)

Dimension	Training Type	CCC	PCC	RMSE	SAGR
Valence	Separate	0.367	0.485	0.351	0.697
	Joint	**0.687**	**0.750**	**0.258**	**0.797**
Arousal	Separate	0.403	0.558	0.254	0.729
	Joint	**0.685**	**0.746**	**0.182**	**0.804**

V. CONCLUSION

In this study, we proposed a novel method for continuous emotion recognition that dynamically selects five frames around a target frame based on visual dissimilarity. By incorporating temporally and contextually diverse frames, our approach enhances the model's ability to capture dynamic emotional changes. Experimental results on the VEATIC benchmark demonstrate that our frame selection strategy outperforms conventional uniform sampling methods in predicting both valence and arousal. In particular, frame selection based on pixel-wise differences effectively captures rapid scene changes and helps reflect subtle emotional transitions. On the other hand, the SSIM-based approach failed to ensure sufficient diversity due to suboptimal threshold settings, often selecting frames too similar to the target frame. Furthermore, our results show that jointly learning valence and arousal in a multi-task setting consistently improves performance, suggesting the utility of modeling their latent correlation.

For future work, we plan to explore more semantically rich similarity metrics such as LPIPS (Learned Perceptual Image Patch Similarity) to better capture human-perceived differences. Additionally, we aim to model higher-level social cues, such as the presence of others and inter-person interactions, to improve third-person emotion understanding in real-world scenarios.

REFERENCES

[1] Z. Ren, C. Pan, C. Zhao, Y. Chen, M. Xu, and Y. Zhuang, "VEATIC: Video-Based Emotion and Affect Tracking in Context Dataset," *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2484–2494, Jan. 2024.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[3] M. Mollahosseini, B. Hasani, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, 2016.

[4] T. Almaev and M. Valstar, "Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild," *IEEE CVPR Workshops*, pp. 1–10, 2017.

[5] A. Kollias and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-Wild database and challenge, deep architectures, and beyond," *International Journal of Computer Vision*, vol. 127, pp. 984–1011, 2019.

[6] H. Song, W. Zhang, Z. Lin, and Y. Liu, "Affective computing using attention-based transformer networks," *IEEE Transactions on Affective Computing*, 2023.

[7] T. Chen and P. C. Yuen, "Automatic facial expression recognition using information-theoretic saliency," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 1963–1975, 2018.

[8] X. Liu, B. Yang, and S. Luo, "Emotion recognition from facial expressions using multiscale deep learning," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 389–401, 2018.

[9] J. Kossaifi, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "AFEW-VA database for valence and arousal estimation in-the-wild," *Image and Vision Computing*, vol. 65, pp. 23–36, 2017.

[10] A. Kollias, G. Tzirakis, M. A. Nicolaou, A. Papaioannou, and S. Zafeiriou, "Aff-Wild: Valence and Arousal 'in-the-wild' Challenge," *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 34–41, 2017.