Journal of Organizational and End User Computing
Volume 36 • Issue 1

User Engagement Detection-Based Financial Technology Advertising Video Effectiveness Evaluation

Qun Gao, School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai, China*

ABSTRACT

With the rapid advancement of financial technology, an increasing number of related advertisements have received widespread attention. User engagement detection during the advertisement viewing process directly reflects the effectiveness of the advertising video. Therefore, detecting user engagement during the advertisement viewing process has become a crucial issue. However, traditional engagement detection methods often require significant computational resources, significantly reducing their practicality. To address this issue, the authors propose a method to effectively detect user engagement by fully integrating multiple relatively practical models. Specifically, the authors extract key frame images from user face video and perform super-resolution reconstruction of them. Then image pyramid matching is used to achieve user engagement detection. Finally, the authors establish a reasonable database and conduct sufficient experiments based on it. Experimental results demonstrate that this proposed method has realistic engagement detection accuracy, and the design of multiple steps is also valid.

KEYWORDS

salient object detection, spatial pyramid matching, super-resolution reconstruction, user engagement detection

INTRODUCTION

With the rapid development and popularization of financial technology, a large number of related advertisements appear in our lives. More and more users gain a deeper understanding of them by watching advertisements, and the role of advertising is becoming increasingly significant. Therefore, evaluating the quality of advertisements has become an important problem. Although user-engagement detection can be applied to evaluate the quality of advertisements, traditional engagement-detection methods often consume enormous computing resources and have poor practical application value. Therefore, designing and using an effective engagement-detection method to evaluate the quality of advertisements is of great value and significance.

To address this issue, we refine the engagement detection into multiple steps. First, we extract key-frame images through salient-object detection and perform image super-resolution reconstruction

DOI: 10.4018/JOEUC.340931 *Corresponding Author

on the extracted key-frame images. Based on this, we obtain final engagement-detection results through spatial pyramid matching. Obviously, the research areas of engagement detection, salient-object detection, super-resolution reconstruction, and image-similarity matching are highly relevant to our method. Many experts have proposed multiple classic models in these fields. Quite a few models are widely used, which directly contributes to high-quality development of related fields.

Early engagement-detection methods used machine-learning methods to extract manual features using static, discrete images as datasets. In recent years, more and more people have paid attention to the concept of engagement (Shahbaznezhad et al., 2021; Nigam & Dewani, 2022). In particular, with the development of deep learning, many end-to-end engagement-detection models have been developed. As for detection of students' engagement in the classroom, DERN (Huang et al., 2019) uses a bidirectional long short-term memory (LSTM) network with added attention mechanism to analyze the features extracted by OpenFace through sequential convolution and obtains the degree of students' participation in online classes. Bhardwaj et al. (2021) paid attention to students' eye movements, using a deep-learning network to analyze students' eye focus to establish a feature matrix for judging students' concentration in online classes. Some deep-learning models effectively improve the performance of engagement detection by designing network architecture and attention mechanism in temporal and spatial domains, respectively. DFSTN (Liao et al., 2021) uses an SE-ResNet-50 network to extract spatial features of faces and uses a LSTM network with a global attention mechanism to extract temporal features. Combining the features of time and space, the model can capture the fine-grained features of face sequences better and improve detection effectively. Shen et al. (2022) proposed an engagement-detection model for massive open online courses to monitor students' learning engagement through a regionally adaptive facial expression recognition network. DenseNet (Mehta et al., 2022) improves the performance of concentration analysis by facial expression through the time, space, and spatiotemporal self-attention mechanism.

Salient-object detection can separate salient objects from the natural scene. Initially, salient-object detection methods (Li et al., 2013; Wang et al., 2016) worked by calculating the contrast difference in the neighborhood around a single or several pixels. After that, some deep learning–based salient-detection models were optimized by fully extracting multi-scale features. PAGE-Net (Wang et al., 2019) uses a pyramid enhancement module and a salient edge-detection module to optimize the salience-detection model. In the pyramid-enhancement module, the model gets multi-scale salience information and pays more attention to the content related to salient detection. In the salient edge-detection module, the authors stack multi-scale salience information and introduce attention mechanism to make the extracted feature receptive field larger and more expressive.

Pang et al. (2020) proposed an aggregate interaction module that aggregated different scale features of adjacent convolutional layers. At the same time, a mutual learning mechanism was introduced to flexibly integrate salience information of different resolutions through interactive learning and effectively aggregate multi-scale features. In addition, some methods use depth image to assist salience detection. UC-Net (Zhang et al., 2021) uses a probabilistic model to extend the results of the salience-detection model to optimize the ambiguity of depth images in the manual annotation process. DMRA (Ji et al., 2022) uses residual connection to extract cross-modal salience information from RGB images and depth images to fully extract fine-grained features. Recently, more novel significance-detection models have been proposed. A novel group weakly supervised learning induced co-saliency detection (Qian et al., 2022) was proposed, which adopts a group of related images of the same type as the ground truth group learning scheme to look for common clues in images of the same category.

Super-resolution reconstruction algorithms are designed to restore the input low-resolution image to the corresponding high-resolution image. Recently, super-resolution reconstruction models based on deep learning have become more popular, and some schemes have used attention mechanisms to optimize deep convolutional neural networks. Zhang et al. (2018) combined the channel attention mechanism with residual blocks to propose a residual channel attention network for super-resolution

reconstruction. Based on this, SAN (Dai et al., 2019) uses a second-order attention mechanism that adaptively learned channel features that can better represent edge and texture features to adjust feature representations accordingly. Niu et al. (2020) proposed a more specific attention mechanism. In the process of feature extraction, channel information of each layer and position is weighted differently so as to improve the quality of reconstructed images.

More recently, novel models have been proposed. Swin2SR (Conde et al., 2022) improves on the basis of Swin Transformer and uses window-based multi-head self-attention to divide the feature graph into several disjointed regions during the down-sampling process and then calculate multi-head attention only within each window so that the calculation amount is effectively reduced. Chen et al. (2024) proposed a lightweight image super-resolution network that incorporates multiple layers of features. The model uses two layers of nested asymmetric residuals to better extract features while reducing the number of parameters.

Initially, image-similarity matching methods based on machine learning (Rahman et al., 2007; Shrivastava et al., 2011) usually used general algorithms such as improved support vector machines and fuzzy c-means to extract features and evaluate the similarity of images. After that, SiameseNet (Zagoruyko & Komodakis, 2015), an image-similarity matching method based on deep learning, extracts feature representations through convolutional neural networks and combines two branches of Siamese network (Bromley et al., 1993) to measure similarity based on features, effectively improving the matching accuracy. On this basis, more models improve their performance by continuously refining the feature representation corresponding to the image. Xiong et al. (2021) proposed the offset mean filtering method, which uses a feature descriptor to show the degree of self-similarity to achieve multimodal remote sensing image matching. Jiang et al. (2022) combined the structural consistency of quadratic programming and local graphs to propose a local graph that maintained topological structure and the correctness of image matching while extracting features.

Recently, more innovative ideas have emerged in the field of image matching. Inspired by the idea of neighborhood pool, Dereli and Ünsal (2023) embedded a new hardware structure into field programmable gate arrays to improve the work efficiency of hardware in facial image matching tasks. He et al. (2023) introduced the prior knowledge of topological invariance into the similarity matching criteria to obtain the consistent representation of images and realized the unsupervised clustering of 3D medical images.

It is particularly worth mentioning that in recent years, more and more deep-learning models (Feng & Chen, 2022; Hou et al., 2022; Li et al., 2022) have been applied to finance and some other popular fields, playing a very important role. Analyzing relevant fields deeply and drawing on the experience, this paper optimizes and deeply combines several classic models (Zhao et al., 2019; Ledig et al., 2017; Lazebnik et al., 2006), thus achieving high-quality engagement detection. Specifically, the EGNet model (Zhao et al., 2019) is used to judge the key-frame image from the perspective of saliency detection. After super-resolution reconstruction of the key-frame image by the SRGAN model (Ledig et al., 2017), the final engagement detection is completed through image pyramid matching (Lazebnik et al., 2006). In general, this method not only can fully combine multiple models to achieve high-quality engagement detection but also can minimize the consumption of computing resources, mainly through avoiding computing the entire video.

METHOD

First, the EGNet model is used to detect the saliency of each frame image and determine if it is a key frame in the video. Second, for the key-frame images that are determined, an image super-resolution reconstruction model is used to improve their quality. Third, an image spatial pyramid matching technique is used to determine the engagement condition of the key-frame images. By analyzing the engagement condition of multiple key-frame images, judgments about the user's engagement condition can be made.

Salient-object detection was used for the detection of key frames in videos. In a salient-object detection task, there is always a problem of the target missing local edge information. Essentially, EGNet (Zhao et al., 2019) explicitly uses edge information and fuses it into the salient object to optimize the missing edge. EGNet consists of three modules: salient-object feature extraction, edge extraction, and one-to-one guidance. The progressive salient-object features extraction adopts the improved U-Net, adding three convolutional layers in each path and adding a ReLU activation function after the convolutional layer to ensure the learning of nonlinear features. This improvement enables the network to acquire richer context features.

In nonlocal salient-edge features extraction, the edge information of a salient object can be modeled and the edge information can be extracted effectively. In the process of extracting the edge features of significant objects, it is not enough to obtain only local edge information; more advanced position information also must be obtained. In the neural network, the top layer of the network has the largest receptive field, so that has the richest location information. Based on this, a top-down position propagation method, where the top-level position information is propagated layer by layer to optimize the original local edge features, is adopted. Then the optimized edge features are convolved and nonlinear calculated in a transition layer to get the final extracted edge features. Use F to represent the local features of the added location information, W to represent the parameters of the transition layer, P_r to represent the mapping of the confidence degree of the prediction salient object, and Z^- and Z^+ to represent nonsignificant class and significant class, respectively. The objective function of the edge feature extraction module can be expressed by (1).

$$L = -\sum_{j \in Z^+} logP_r\left(y_i = 1|F;W\right) - \sum_{j \in Z^-} logP_r\left(y_i = 0|F;W\right) \tag{1}$$

Finally, in order to make salient edge features guide salient object segmentation and localization effectively, a one-to-one guidance module is proposed. Sub-paths are added to some network layers. In the sub-path, the salient edge features are fused with the salient object features to optimize the segmentation details. The output of the one-to-one guidance module can be represented by (2).

$$G = T\left(F;\Theta \quad F_E\right) + F_E \tag{2}$$

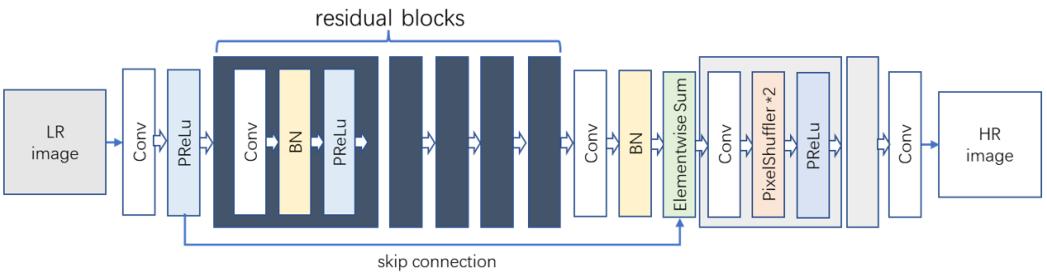
T represents the convolution layer, F represents the salient edge feature, F_E represents the salient object feature of the corresponding layer, and Θ represents the convolution layer parameter.

Once the salient object of one frame image is obtained, the size and peak signal-to-noise ratio (PSNR) of a salient object can be detected. If the size of the salient object is large and the PSNR is high, the frame image can be determined as a key-frame image directly.

Through this approach above, we have more fully integrated the EGNet model to detect a key-frame image, especially during the implementation process, and we have optimized some details based on the actual situation.

On this basis, in order to further enhance the clarity of the key-frame images, we used an image super-resolution model to reconstruct them. SRGAN (Ledig et al., 2017) is a generative adversarial network for image super-resolution reconstruction that is composed of a generator and a discriminator. A deep residual network is used as the generator. Because the deep network framework is difficult to train, SRGAN adds many skip connections to the deep residual network as the generator so that the deep network has less information loss during calculation and effectively improves the accuracy of the network. Fig. 1 illustrates some essential details more clearly.

Figure 1. The deep residual network with skip connections as the generator



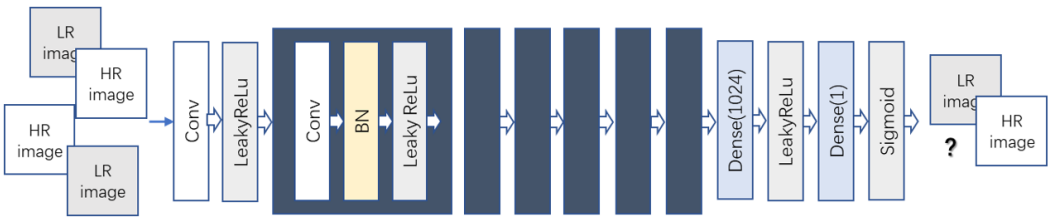
The discriminator is composed of several convolution modules in series; each module is composed of a convolution layer, a batch normalization (BN) layer, and the activation function. In the discriminator, LeakyReLU is used as the activation function after the convolution layer to avoid maximum pooling and reduce information loss. Fig. 2 illustrates some essential details more clearly.

In order to improve the speed and accuracy of image super-resolution reconstruction and to further restore image-texture details while obtaining the super-resolution image, SRGAN uses a visual loss suitable for generative adversarial network. Specifically, the loss function, called perceptual loss, consists of content loss and adversarial loss. During the process of image super-resolution reconstruction, many algorithms use mean square error (MSE) as the loss function to optimize the network. Since MSE carries out average operation in the calculation, this loss function can improve the peak signal-to-noise ratio but will lose the texture information and result in a fuzzy image. To solve this problem, a pretrained VGG network is used to extract the middle-layer features of the image and the content loss is defined on the information, which can effectively retain the high-frequency information of the image. The definition of content loss is shown in (3), where *ij* represents the feature graph obtained by the *i*th convolutional layer after the *i*th maximum pooled layer in the VGG network, W_{ij} and H_{ij} represent the length and width of the feature graph, respectively, and G represents the generator.

$$L_{ij} = \frac{1}{W_{ij}H_{ij}}\sum_{x=1}^{W_{ij}}\sum_{y=1}^{H_{ij}}\varphi_{ij}\left(I^{HR}\right)_{x,y} - \varphi_{ij}\left(G\left(I^{LR}\right)\right)_{x,y}^{2} \tag{3}$$

Through the above approach, the SRGAN model has been more fully integrated to reconstruct the key-frame images. The quality of the image would be significantly improved; especially, the details in the image would become clearer. Meanwhile, in the process of model implementation, some details have been optimized based on the actual situation.

Figure 2. The discriminator for SRGAN



After completing the super-resolution reconstruction of key-frame images, how to accurately match each key-frame image with the reference image (high engagement image) is an urgent problem to be solved. In order to achieve more comprehensive and accurate image matching, a spatial pyramid image-matching model (Lazebnik et al., 2006) was applied. Spatial pyramid matching is an image-matching method that uses subregions to get the feature pyramid, counts the number of feature points in each layer of the pyramid, and finally calculates the matching degree of the image. Specifically, spatial pyramid matching divides the image into several subregions, calculates the features of each subregion, and then splices the block features as the matching features of the whole image. The idea of multi-scale segmentation is adopted in the process of segmentation so that the granularity of extracted features changes with the size of the segmentation, forming a pyramid structure.

As shown in Fig. 3, the image is divided into different subregions. The extraction level of features is different according to the number of divided regions. Use points to represent the key points in each layer. For the layers that have been divided into regions, count the frequency of the key points in each area and then count the feature distribution in each layer to obtain the pyramid of the image features.

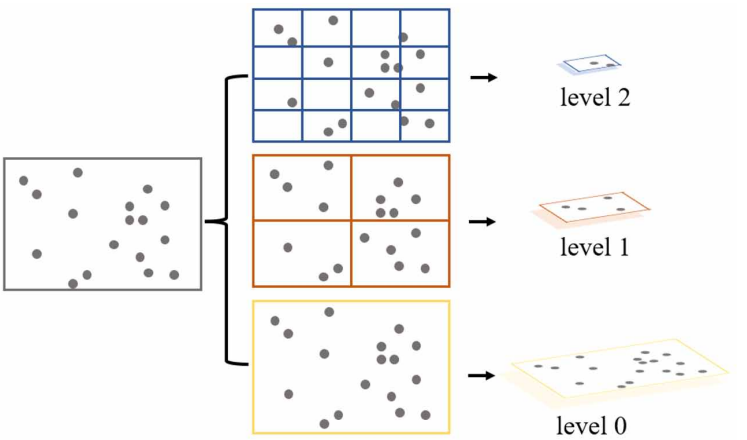
After obtaining the feature pyramid of the image, calculate the corresponding number of matching points according to the number of features in each layer. Specifically, use D to represent the number of subregions and H_X to represent the number of subregion features of image X in the lth layer of the pyramid. The matching points of two images, X and Y, in the lth-layer pyramid can be obtained using (4).

$$I\left(H_X^l, H_Y^l\right) = \sum_{i=1}^{D} \min\left(H_X^l\left(i\right), H_Y^l\left(i\right)\right) \tag{4}$$

The matching points of each layer are calculated separately, and then the matching degree of the two images is calculated by the increment of matching points between the two adjacent layers. L is used to represent the number of layers of the image pyramid, and the matching degree between X and Y images can be expressed by (5).

$$K^L\left(X,Y\right) = I^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}}\left(I^l - I^{l+1}\right) \tag{5}$$

Figure 3. Division of the image into subregions of different sizes to form a pyramid structure



Furthermore, if the similarity between a key-frame image and the reference image (high-engagement image) is high, the key-frame image is directly determined to have high engagement. If the similarity between a key-frame image and the reference image is low, the key frame is determined to have low engagement. Meanwhile, in the process of model implementation, we have optimized some details based on the actual situation. In this manner, each key-frame image is judged. If the majority of key-frame images are determined as having high engagement, it is concluded that the user is highly engaged when watching the advertising video.

We can see that multiple models above support and combine with each other, effectively promoting engagement detection. The essential parameters of the models could be adjusted or trained, which could better ensure the adaptability and robustness and thus achieve realistic experimental results even for face videos under different conditions.

EXPERIMENTS

Based on the actual situation, a large number of videos of faces watching advertisements were collected and a database was built. To make the database more comprehensive and objective, some low-light, high-noise, and occluded-face videos were purposely added. In order to test the accuracy and further evaluate whether there was randomness in the experimental results, the face video database was grouped and experiments were conducted on each group. The experimental results in Figs. 4 and 5 show that the method achieved ideal detection accuracy and there is no significant difference in the experimental results of each group, which also proves the reliability of the experimental results from another perspective.

To verify the effectiveness of each step, an ablation experiment was conducted. The ablation-experiment results in Figs. 6 and 7 demonstrate that each step can play a relatively pivotal role. The experimental results are generally in line with expectations. On the one hand, the method can achieve realistic detection accuracy, and group experiments have also proven that realistic experimental results are not random. On the other hand, valid ablation experiments provide sufficient evidence for the effectiveness of each essential step. Although there may occasionally be some unsatisfactory experimental results (e.g., the precision of engagement detection would obviously decline for challenging face videos), this method still shows superiority from multiple perspectives.

Figure 4. Comparison of AP values using the proposed method on different face videos

Note. The AP values are experimental results based on the whole database and other randomly divided groups of face videos.

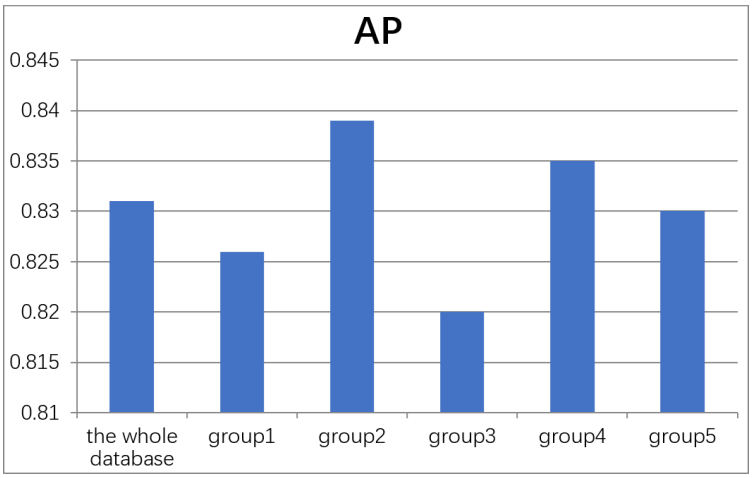


Figure 5. Comparison of AUC values using the proposed method on different face videos
Note. The AUC values are experimental results based on the whole database and other randomly divided groups of face videos.

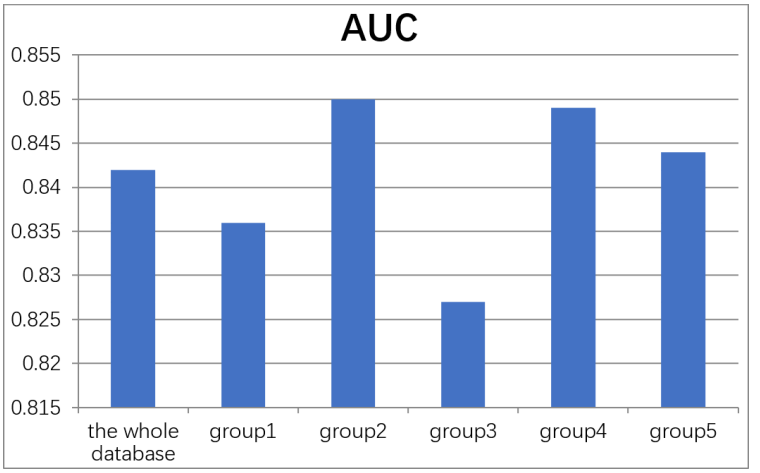
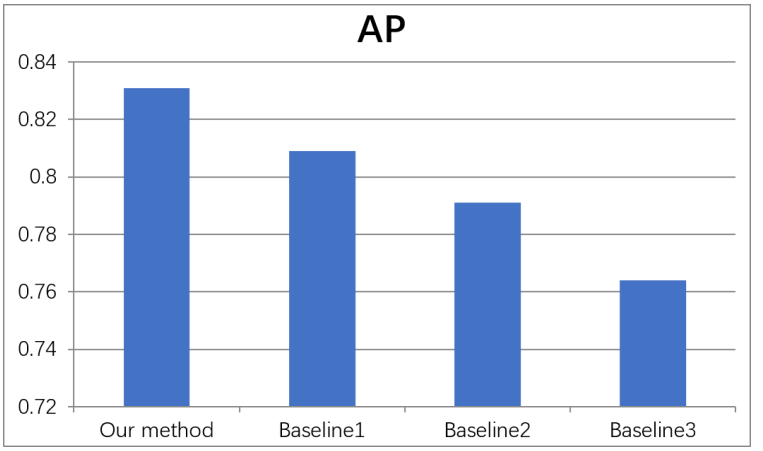


Figure 6. Comparison of AP values by the proposed model and baseline models
Note. The Baseline1 model is to detect the engagement without using salient-object detection. The Baseline2 model is to detect the engagement without using super-resolution reconstruction. The Baseline3 model is to detect the engagement without using spatial pyramid matching.



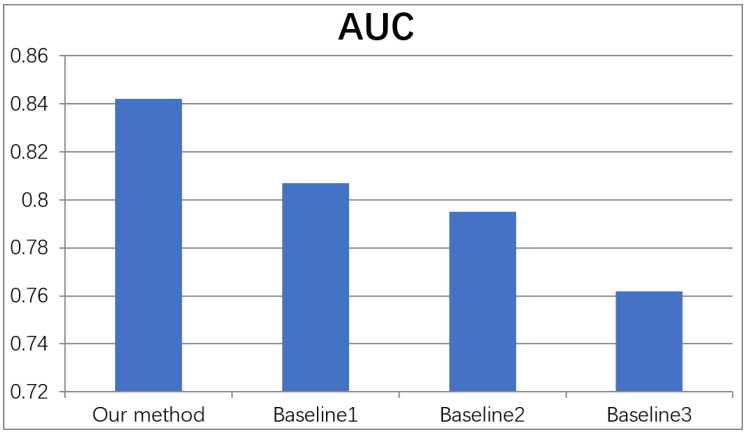
CONCLUSION

This paper proposes an innovative approach to detect user engagement, which could be effectively applied to advertising video effectiveness evaluation. Specifically, the method in this paper consists of multiple essential models. These models are fully integrated and mutually supportive, significantly improving the robustness of the engagement-detection method. In particular, this method can significantly reduce the consumption of computing resources, which directly contributes to its practical application value.

However, the method still has some limitations. Specifically, it involves multiple essential models and they are interrelated intimately, so if an error occurs in a certain model, the error may

Figure 7. Comparison of AUC values by the proposed model and baseline models

Note. The Baseline1 model is to detect the engagement without using salient-object detection. The Baseline2 model is to detect the engagement without using super-resolution reconstruction. The Baseline3 model is to detect the engagement without using spatial pyramid matching.



be significantly increased, which directly affects the final engagement-detection accuracy. In future work, it will be important to further reduce the detection error, thus achieving higher quality of engagement detection.

AUTHOR NOTE

This work was supported by the Yantai Science and Technology Innovation Development Plan Project (Grant No. 2022XDRH015).

REFERENCES

Bhardwaj, P., Gupta, P. K., Panwar, H., Siddiqui, M. K., Morales-Menendez, R., & Bhaik, A. (2021). Application of deep learning on student engagement in e-learning environments. *Computers & Electrical Engineering*, *93*, 107277. doi:10.1016/j.compeleceng.2021.107277 PMID:36567679

Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., & Shah, R. (1993). Signature verification using a "Siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, *7*(4), 669–688. doi:10.1142/S0218001493000339

Chen, Y., Xia, R., Yang, K., & Zou, K. (2024). MFFN: Image super-resolution via multi-level features fusion network. *The Visual Computer*, *40*(2), 489–504. doi:10.1007/s00371-023-02795-0

Conde, M. V., Choi, U. J., Burchi, M., & Timofte, R. (2022). Swin2SR: SwinV2 transformer for compressed image super-resolution and restoration. In L. Karlinsky, T. Michaeli, & K. Nishino (Eds.), Lecture Notes in Computer Science: Vol. 13802. *Computer Vision—ECCV 2022 Workshops*. Springer., doi:10.1007/978-3-031-25063-7_42

Dai, T., Cai, J., Zhang, Y., Xia, S. T., & Zhang, L. (2019). Second-order attention network for single image super-resolution. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11057–11066). IEEE. doi:10.1109/CVPR.2019.01132

Dereli, S., & Ünsal, A. R. (2023). A new microarchitecture hardware proposal for the use of the improved template matching method in face similarity detection. *Microprocessors and Microsystems*, *101*, 104908. doi:10.1016/j.micpro.2023.104908

Feng, Z., & Chen, M. (2022). Platformance-based cross-border import retail e-commerce service quality evaluation using an artificial neural network analysis. [JGIM]. *Journal of Global Information Management*, *30*(11), 1–17. doi:10.4018/JGIM.306271

He, Y., Yang, G., Ge, R., Chen, Y., Coatrieux, J. L., Wang, B., & Li, S. (2023). Geometric visual similarity learning in 3D medical image self-supervised pre-training. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9538–9547). IEEE. doi:10.1109/CVPR52729.2023.00920

Hou, H., Tang, K., Liu, X., & Zhou, Y. (2022). Application of artificial intelligence technology optimized by deep learning to rural financial development and rural governance. [JGIM]. *Journal of Global Information Management*, *30*(7), 1–23. doi:10.4018/JGIM.289220

Huang, T., Mei, Y., Zhang, H., Liu, S., & Yang, H. (2019). Fine-grained engagement recognition in online learning environment. In *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)* (pp. 338–341). IEEE. doi:10.1109/ICEIEC.2019.8784559

Ji, W., Yan, G., Li, J., Piao, Y., Yao, S., Zhang, M., Cheng, L., & Lu, H. (2022). DMRA: Depth-induced multi-scale recurrent attention network for RGB-D saliency detection. *IEEE Transactions on Image Processing*, *31*, 2321–2336. doi:10.1109/TIP.2022.3154931 PMID:35245195

Jiang, X., Xia, Y., Zhang, X. P., & Ma, J. (2022). Robust image matching via local graph structure consensus. *Pattern Recognition*, *126*, 108588. doi:10.1016/j.patcog.2022.108588

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (pp. 2169–2178). IEEE. doi:10.1109/CVPR.2006.68

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4681–4690). IEEE. doi:10.1109/CVPR.2017.19

Li, G., Wang, X., Bi, D., & Hou, J. (2022). Risk measurement of the financial credit industry driven by data: Based on DAE-LSTM deep learning algorithm. [JGIM]. *Journal of Global Information Management*, *30*(11), 1–20. doi:10.4018/JGIM.308806

Li, J., Tian, Y., Duan, L., & Huang, T. (2013). Estimating visual saliency through single image optimization. *IEEE Signal Processing Letters*, *20*(9), 845–848. doi:10.1109/LSP.2013.2268868

Liao, J., Liang, Y., & Pan, J. (2021). Deep facial spatiotemporal network for engagement prediction in online learning. *Applied Intelligence*, *51*(10), 6609–6621. doi:10.1007/s10489-020-02139-8

Mehta, N. K., Prasad, S. S., Saurav, S., Saini, R., & Singh, S. (2022). Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement. *Applied Intelligence*, *52*(12), 13803–13823. doi:10.1007/s10489-022-03200-4 PMID:35340984

Nigam, A., & Dewani, P. P. (2022). Consumer engagement through conditional promotions: An exploratory study. [JGIM]. *Journal of Global Information Management*, *30*(5), 1–19. doi:10.4018/JGIM.290364

NiuB.WenW.RenW.ZhangS.YangL.WangS.ZhangK.CaoX.ShenH. (2020). Single image super-resolution via a holistic attention network. arXiv. https://arxiv.org/abs/2008.08767

Pang, Y., Zhao, X., Zhang, L., & Lu, H. (2020). Multi-scale interactive network for salient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9413–9422). IEEE. doi:10.1109/CVPR42600.2020.00943

Qian, X., Zeng, Y., Wang, W., & Zhang, Q. (2022). Co-saliency detection guided by group weakly supervised learning. *IEEE Transactions on Multimedia*, *25*, 1810–1818. doi:10.1109/TMM.2022.3167805

Rahman, M. M., Bhattacharya, P., & Desai, B. C. (2007). A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback. *IEEE Transactions on Information Technology in Biomedicine*, *11*(1), 58–69. doi:10.1109/TITB.2006.884364 PMID:17249404

Shahbaznezhad, H., Dolan, R., & Rashidirad, M. (2021). The role of social media content format and platform in users' engagement behavior. *Journal of Interactive Marketing*, *53*(1), 47–65. doi:10.1016/j.intmar.2020.05.001

Shen, J., Yang, H., Li, J., & Cheng, Z. (2022). Assessing learning engagement based on facial expression recognition in MOOC's scenario. *Multimedia Systems*, *28*(2), 469–478. doi:10.1007/s00530-021-00854-x PMID:34690439

Shrivastava, A., Malisiewicz, T., Gupta, A., & Efros, A. A. (2011). Data-driven visual similarity for cross-domain image matching. *ACM Transactions on Graphics*, *30*(6), 1–10. doi:10.1145/2070781.2024188

Wang, W., Zhao, S., Shen, J., Hoi, S. C. H., & Borji, A. (2019). Salient object detection with pyramid attention and salient edges. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1448–1457). IEEE. doi:10.1109/CVPR.2019.00154

Wang, X., Ma, H., & Chen, X. (2016). Salient object detection via fast R-CNN and low-level cues. In *2016 IEEE International Conference on Image Processing (ICIP)* (pp. 1042–1046). IEEE. doi:10.1109/ICIP.2016.7532516

Xiong, X., Jin, G., Xu, Q., & Zhang, H. (2021). Self-similarity features for multimodal remote sensing image matching. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *14*, 12440–12454. doi:10.1109/JSTARS.2021.3131489

Zagoruyko, S., & Komodakis, N. (2015). Learning to compare image patches via convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4353–4361). IEEE. doi:10.1109/CVPR.2015.7299064

Zhang, J., Fan, D. P., Dai, Y., Anwar, S., Saleh, F., Aliakbarian, S., & Barnes, N. (2021). Uncertainty inspired RGB-D saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5761–5779. doi:10.1109/TPAMI.2021.3073564 PMID:33856982

ZhangY.LiK.LiK.WangL.ZhongB.FuY. (2018). Image super-resolution using very deep residual channel attention networks. arXiv. https://arxiv.org/abs/1807.02758 10.1007/978-3-030-01234-2_18

Zhao, J., Liu, J. J., Fan, D. P., Cao, Y., Yang, J., & Cheng, M. M. (2019). EGNet: Edge guidance network for salient object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 8779–8788). IEEE. doi:10.1109/ICCV.2019.00887