IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 1, February, 2024, pp. 225-235
Journal homepage: http://iieta.org/journals/ts

Optimizing Remote Teaching Interaction Platforms Through Multimodal Image Recognition Technology

Qin Fang, Yawen Zhang*

School of Education Science, Xinjiang Normal University, Urumqi 830017, China

Corresponding Author Email: xjsdzyw@xjnu.edu.cn

https://doi.org/10.18280/ts.410118

Received: 3 August 2023
Revised: 23 December 2023
Accepted: 5 January 2024
Available online: 29 February 2024

Keywords:
remote teaching, multimodal image recognition, self-attention mechanism, encoder-decoder model, image annotation, visual saliency, education technology optimization

ABSTRACT

In the context of the digital era, remote teaching has become an integral part of the global education system. Effective remote teaching relies on the high interactivity of interaction platforms and the precise delivery of teaching content, with multimodal image recognition technology playing a key role. This technology enhances the intelligence level of remote teaching platforms by integrating visual and textual information, providing a richer and more intuitive interactive experience for teachers and students. However, existing multimodal image recognition technologies still face challenges in accuracy, real-time performance, and semantic understanding, especially in complex teaching scenarios where the understanding and feedback on teaching content are not accurate enough, limiting the effectiveness of remote teaching interaction platforms. Addressing these limitations, this paper proposes a multimodal image alignment method based on a self-attention mechanism that effectively integrates visual information into an encoder-decoder model to achieve high consistency between images and teaching content. Additionally, a novel multimodal image annotation and recognition algorithm is introduced, considering both semantic information and visual saliency to achieve higher recognition accuracy and practicality. Experimental validation shows significant improvements in the accuracy and real-time performance of multimodal image recognition, providing strong technical support for remote teaching interaction platforms, optimizing the allocation of teaching resources, and enhancing the quality and efficiency of education.

1. INTRODUCTION

With the rapid development of information technology and the trend towards the globalization of education, remote teaching, as an emerging educational model, is gradually changing the traditional ways of teaching and learning [1-4]. However, the construction of an effective remote teaching interaction platform not only requires efficient transmission protocols and a stable network environment but also advanced image recognition technology to enhance the real-time nature and interactive experience of teaching [5]. Based on multimodal image recognition technology, a more precise understanding and presentation of teaching content can be achieved, thus providing strong support for teacher-student interaction in remote teaching scenarios [6].

Currently, the research on multimodal image recognition technology has significant implications in the field of remote teaching. It can not only enhance the interactivity and teaching effectiveness of remote teaching platforms but also help break through the geographical and temporal constraints of traditional teaching models, providing learners with a personalized and convenient learning experience [7-9]. Furthermore, multimodal technology promotes the optimization of educational resources and the creation of high-quality content, contributing to educational equity and quality improvement [10-12].

However, despite the application of existing multimodal image recognition technology in the field of remote teaching, there are still some defects and shortcomings [13-17]. For example, in dealing with complex teaching scenarios, existing technologies often struggle to accurately align and recognize relevant information across different modalities, leading to a discrepancy in the interactive experience compared to the intuitive feeling in real teaching environments [18, 19]. In addition, regarding the automatic annotation and understanding of image content, current methods cannot adequately recognize and express the teaching semantics implied in images, limiting the functionality of remote teaching platforms [20].

In response to the above issues, this paper proposes a series of innovative research methods. First, we design a multimodal alignment method based on the self-attention mechanism, which can effectively integrate visual information into the encoder-decoder model, thereby achieving a highly consistent synchronization of images with teaching content. Secondly, in response to the specific needs of remote teaching interaction, this paper proposes a multimodal image annotation and recognition algorithm that combines semantic information and visual saliency to improve the accuracy of understanding image content and the practicality of the interaction platform.

Through these studies, this paper aims to deepen the technical framework of the remote teaching interaction platform, optimize the user experience, and promote the advancement of remote education technology.

2. MULTIMODAL ALIGNMENT OF IMAGES FOR REMOTE TEACHING INTERACTIONS

In the construction of modern remote teaching interaction platforms, optimizing multimodal image recognition technology is key to improving teaching quality and interaction efficiency. This paper develops a novel multimodal alignment method based on a self-attention mechanism, innovatively addressing the issue of imprecise information alignment in complex teaching scenarios encountered by traditional technologies. By integrating an encoder-decoder model, it significantly enhances the accuracy of synchronous expression between images and teaching content, making remote education more interactive and adaptable to teaching.

In remote teaching environments, teachers rely on images and other multimedia resources to assist in teaching, while traditional single-text translation models fail to fully utilize this visual information. In remote teaching interaction platforms based on multimodal image recognition technology, effectively aligning image and text information is key to achieving efficient teaching interactions. Existing research often focuses on the utilization of visual information in the decoding phase, which, to some extent, improves the quality of text translation but does not fully explore the potential correlation between visual and textual features, resulting in unfulfilled modality complementarity. Considering the complexity of teaching content in remote teaching scenarios, it is necessary to more closely integrate key visual elements of images with textual information to ensure high semantic consistency. The encoder-decoder model proposed in this paper, designed with a self-attention mechanism, integrates visual information directly at the encoding stage, allowing the model to capture the subtle connections between image and text earlier, laying a solid multimodal semantic foundation for the subsequent decoding phase.

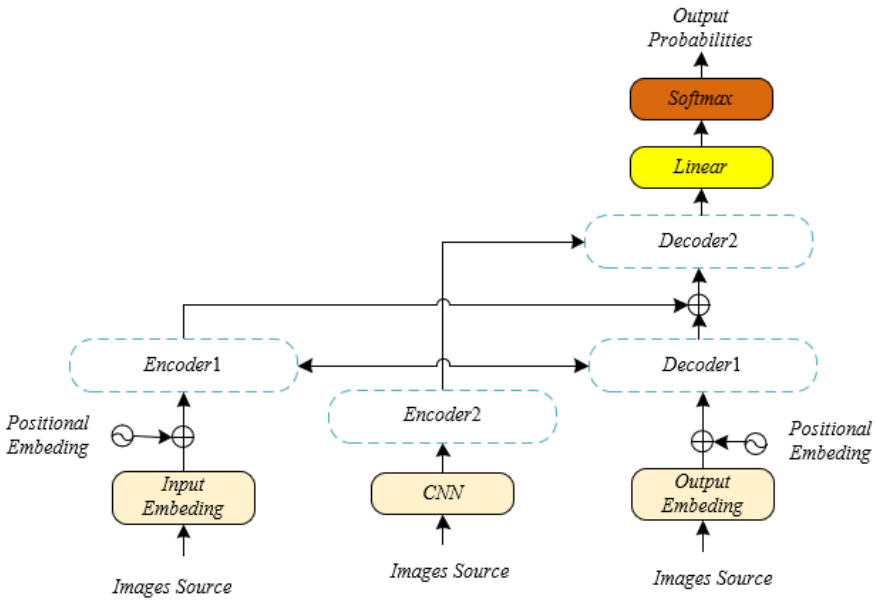

Figure 1. Structure of the multimodal alignment model for remote teaching interaction images

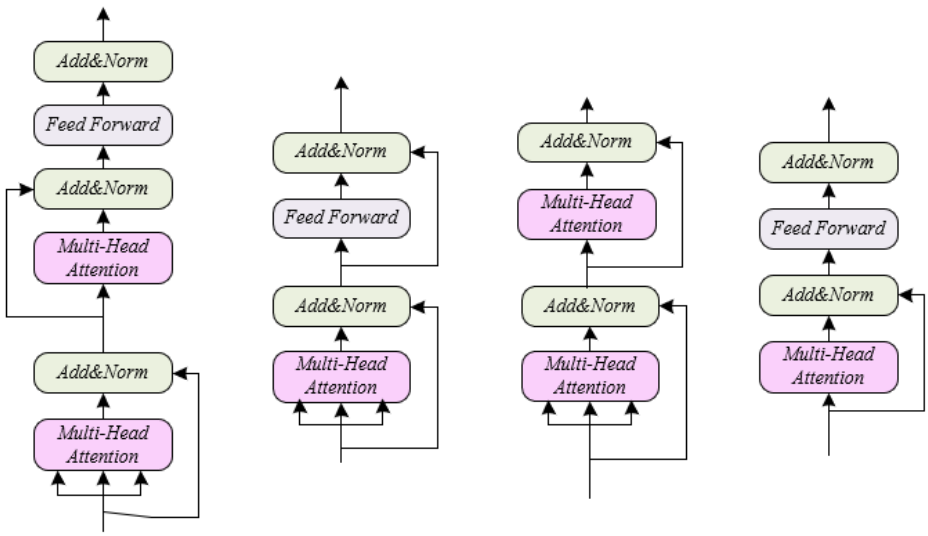

Figure 2. Encoder and decoder structure corresponding to Figure 1

The encoder-decoder model proposed in this paper is based on the *Transformer* architecture, specifically designed with source visual encoders and target visual encoders to process and integrate image information related to teaching content, as shown in Figure 1. The specific structure of the encoder and decoder corresponding to Figure 1 is detailed in Figure 2. The source visual encoder is responsible for extracting features of course-related images, while the target visual encoder extracts features of visual content in the target language that learners may need to generate. These two encoders, combined with traditional text encoders, capture the intrinsic connection between text and vision through the self-attention mechanism, achieving deep integration of source language text and related visual information. This paper also introduces a visual consistency decoder, which uses the text-visual joint representation obtained in the encoding phase, along with the output of the source and target visual encoders, to generate target language sentences consistent with the semantics of the source images. This module further optimizes text-visual interaction in the decoder, adjusting the model's attention distribution to produce more accurate and natural language expressions.

2.1 Encoder

The encoder-decoder model proposed in this paper processes multimodal inputs through the source visual encoder and the target visual encoder. The basic working principle of the source visual encoder is to extract features from teaching images and convert these visual features into high-dimensional representations corresponding to textual information. The multi-head attention mechanism in the encoder is core to its function, enabling the model to process multiple streams of information in parallel and capture complex text-visual associations. Each "head" has different weights in the attention mechanism, allowing the model to capture features in different subspaces and learn richer and complementary information. For example, some "heads" may focus on the correspondence between local details of images and text descriptions, while others may capture the relationship between the global structure of images and teaching objectives. In this way, the model can capture key information in images and combine it with textual data, providing rich context for the subsequent decoding process. The target visual encoder focuses on aligning these visual features with the target language generation process, ensuring that the textual information maintains context and semantic information consistent with the original images during translation or content generation. This working principle allows the model not only to focus on the textual content itself but also on the relationship between textual content and visual information, thereby achieving more accurate and enriched teaching content interaction.

Formally, let the textual information be represented by $t^{TE}=(t^{TE}_1,\ldots,t^{TE}_V)$, where V words of a text sentence are represented by t^{TE}_V, and the attention weight matrix calculated by the softmax function is represented by $\beta_{u,k}$. The output calculation formula of the attention mechanism integrating source-end visual interaction information is given by the following equation:

$$c_u = \sum_{k=1}^{V} \beta_{u,k} \left(t_k^{TE} Q_{EN}^{N} \right) \tag{1}$$

Assuming the image information is represented by a^{IM}, the superimposition of $a^{TE}{}_u$ and a^{IM} is represented by $a^{TE}{}_u \oplus a^{IM}$, and the calculation formula for $\beta_{u,k}$ is as follows:

$$\beta_{u,k} = \mathrm{softmax}\left(\frac{\left(t_u^{TE} \oplus a^{IM} \right) Q_{EN}^{W} \left(t_k^{TE} Q_{EN}^{J} \right)^{T}}{\sqrt{f_j}} \right) \tag{2}$$

Let a target sentence of L words be represented by $s^{TE}=(t^{TE}_1,\ldots,t^{TE}_L)$. The calculation for obtaining the text-visual interaction information on the target end is as follows:

$$z_u = \sum_{k=1}^{L} \beta_{u,k} \left(s_k^{TE} Q_{DE}^{N} \right) \tag{3}$$

$$\beta_{u,k} = \mathrm{softmax}\left(\frac{\left(t_u^{TE} \oplus a^{IM} \right) Q_{DE}^{W} \left(t_u^{TE} Q_{DE}^{J} \right)^{T}}{f_j} \right) \tag{4}$$

2.2 Decoder

In the training process of the encoder-decoder model, the probability of decoding a source sentence into a target sentence through the *Transformer* can be calculated as follows:

$$o_{t \to s}\left(s \mid t \right) = O\left(b \mid s^{TE}, t^{TE} \right) \tag{5}$$

In the multimodal image recognition environment of remote teaching interaction platforms, integrating solely visual information into the encoder-decoder may cause ambiguity in visual alignment between the source and target languages. This ambiguity arises because, although source and target words may be semantically identical, they might be associated with different parts of an image, leading to information loss or confusion during encoding and decoding. To enhance the synchrony and accuracy of image and language information in remote teaching interaction platforms, it is necessary to design a mechanism that considers the visual information processing of both languages simultaneously. Therefore, this paper introduces a bilingual visual protocol decoder into the model to enhance the model's ability to handle bilingual visual interaction information, ensuring the consistency of visual elements of teaching content with both the source and target languages, thereby optimizing the overall remote teaching interaction experience. The basic working principle of this module is to use a hierarchical attention mechanism during the decoding process to deeply integrate visual and textual features. Firstly, the model calculates the context vectors for each image, capturing the key visual information of the images. Then, these visual context vectors are projected along with textual context vectors into a common feature space. In this common space, the model calculates the attention distribution of the projected vectors, obtaining a weighted context representation that contains the integrated features of both visual and textual information. Assuming the context vector for each image is represented by $z^d{}_u (d=IM)$, the feedforward network is represented by φ, the u-th source and target encoder hidden state is represented by t_u, the self-attention module for merging image and text vectors is represented by s_u, and the weight matrix is represented by Q^e. The calculation formulas,

when z_u is obtained from both image and text features, are as follows:

$$r_u^d = \phi\left(t_u, s_u, z_u^d\right) \tag{6}$$

$$\alpha_u^d = \frac{\exp\left(r_u^d\right)}{\sum_{e\in\{IM,TE\}} \exp\left(r_u^e\right)} \tag{7}$$

$$z_u = \sum_{e\in(IM,TE)} \alpha_u^e Q^e z_u^e \tag{8}$$

Assuming, the standard neural network translation predicts different values of b, represented by t, the final calculation formula for b is as follows:

$$o_{t\to s}\left(s \mid t\right) = O\left(b \mid s_{TE}, t_{TE}, z\right) \tag{9}$$

3. MULTIMODAL IMAGE ANNOTATION AND RECOGNITION FOR REMOTE TEACHING INTERACTIONS

As the use of image resources in remote teaching becomes more frequent, and given that images contain rich information not limited to the objects displayed but also including their contextual relationships and attributes, traditional image recognition methods may overlook the deep semantic association between visually prominent regions within images and textual content. This oversight can prevent teaching resources from fully utilizing image information to enhance the precision and interactivity of teaching. Therefore, researching an algorithm that considers both the visual prominence of images and the semantic salience of text is of significant importance for optimizing remote teaching interaction platforms. To meet the optimization needs of remote teaching interaction platforms, this paper proposes a multimodal image recognition and annotation algorithm based on semantic and visual salience. The proposed algorithm first uses advanced visual feature extraction techniques to identify and extract visually prominent regions within images, which are key to capturing students' attention and represent significant information within the image. Combined with a visual bag-of-words model, the algorithm converts these prominent regions into two-layer salience visual bag-of-words features, a process akin to translating images into visual "vocabulary" that machines can understand. Subsequently, the algorithm processes textual content, extracting semantic salience text corresponding to the visually prominent regions of images, ensuring semantic consistency between text and images. Finally, a multi-kernel Support Vector Machine (SVM) is used to learn features that integrate both visual and semantic salience, identifying the semantic categories corresponding to the images, and determining the final textual labels for the images through a nearest neighbor voting strategy. Figure 3 provides an example of the multimodal image annotation and recognition process.

In the research scenario of this paper, visual salience is redefined as the key features in the content of images on remote teaching interaction platforms that attract learners' attention. It includes not only traditional visual elements such as color, contrast, and edges but also extends to educational information density within images, such as shapes, symbols, formulas, etc. These areas act as "focal points" of information transmission during the teaching process. Semantic salience refers to the importance and relevance of text content corresponding to visually salient areas, emphasizing the role of text in conveying specific teaching purposes, ensuring that text can accurately describe key visual information in images, and aiding in understanding and memory. This paper combines these two concepts, aiming to automatically identify and annotate the features of visual and semantic salience in images through the algorithm.

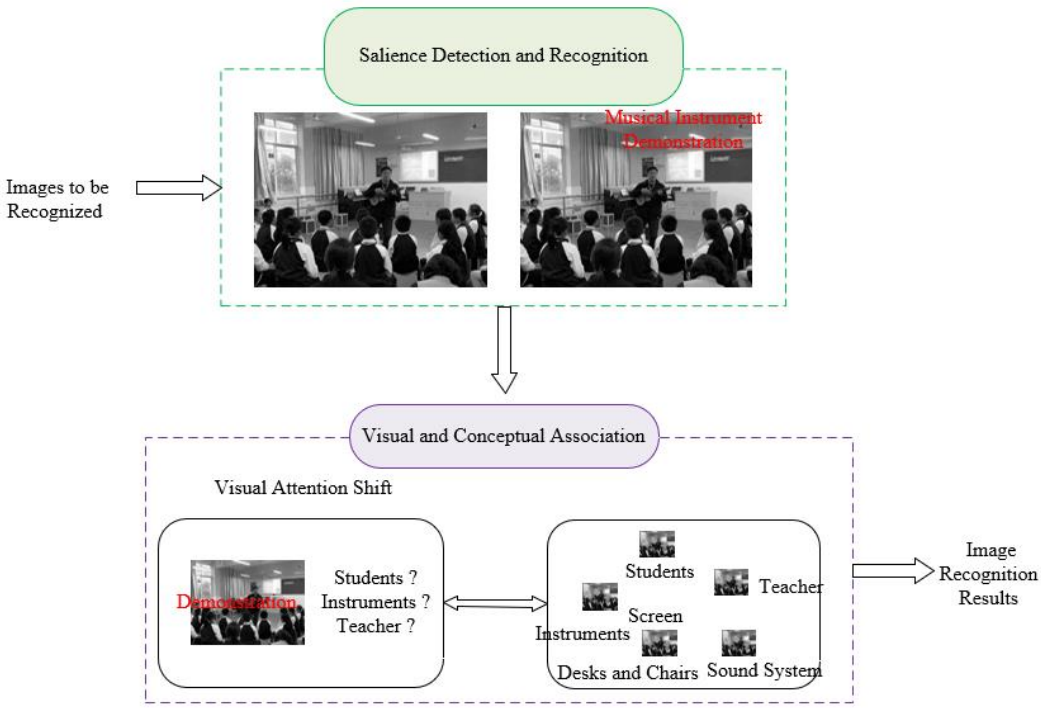

Figure 3. Example of multimodal image annotation and recognition process

3.1 Offline learning

In remote teaching interaction platforms, the organization of teaching content often relies on rich conceptual relationships, which are expressed through images and textual materials. The first step is to use given annotated training samples to build a concept map model that can map and reflect the associations between different concepts. These associations help determine the core and auxiliary concepts within teaching materials, thereby providing a structural background for subsequent semantic analysis and visual annotation. In practical teaching, this concept map can guide teachers and students to quickly understand the knowledge structure, optimizing the arrangement and presentation of teaching content. Figure 4 shows the algorithm's offline learning process.

Assume the constructed concept map model is represented by $H=\{N,R\}$, where its vertex set is represented by N, and the concept (label) set by $Z=\{z_1,z_2,\ldots,z_l\}$, with the weight of the directed edge r_{uk} represented by q_{uk}. If an image in the training set is annotated with both z_u and z_k, then the concepts z_k and z_u are connected by r_{uk}. Further, assume the conditional probability of annotating z_k given z_u is represented by $O(z_k|z_u)$, the total number of images in the training set with the label z_u by $V(z_u)$, and the number of images in the training set labeled with both z_u and z_k by $V(z_k,z_u)$. q_{uk} can be calculated through the following formula:

$$q_{z_u,z_k}=O(z_k\mid z_u)=\frac{V(z_u\mid z_k)}{V(z_u)} \tag{10}$$

Next, a community detection algorithm is used to analyze the set of textual vocabularies, aiming to detect semantic communities within the text and their potential semantic salience. This step identifies tightly associated groups of vocabularies within textual materials, representing specific knowledge points or conceptual domains, with each label being assigned to a specific semantic community. The identified semantic communities help understand and organize teaching content, allowing related knowledge points to be presented collectively, providing semantic focal points and a basis for deeper understanding for students. Assume a semantic label is represented by z_u, a community by ST_j, and the number of concepts in ST_j by V_{STj}, the paper specifically defines their association as follows:

$$CO(z_u,ST_j)=\frac{1}{V_{ST_j}}\sum_{z_k\in ST_j}q_{z_k,z_u}=\frac{1}{V_{ST_j}}\sum_{z_k\in ST_j}\frac{V(z_u,z_k)}{V(z_u)} \tag{11}$$

Assuming the number of semantic communities identified by the community detection algorithm is represented by V_{ST}, and the semantic salience of the label z_u in the semantic community ST_j can be calculated through the following formula:

$$SA(z_u)=\frac{CO(z_u,ST_j)}{\sum_{l=1}^{V_{ST}}CO(z_u,ST_l)} \tag{12}$$

Clearly, the larger $SA(z_u)$ indicates a stronger association between label w and ST_j, while if z_u is shared by multiple semantic communities, $SA(z_u)$ will not be very high. The algorithm further detects visually salient regions within images in the community and generates community salient bag-of-words models based on these regions. This step tightly integrates visual information with semantic content by identifying and extracting key visual features, transforming them into a visual bag-of-words usable for machine learning. On remote teaching platforms, these visually salient regions can highlight key image content, helping students to quickly locate and understand the important parts of teaching materials visually, enhancing the intuitiveness and effectiveness of learning.

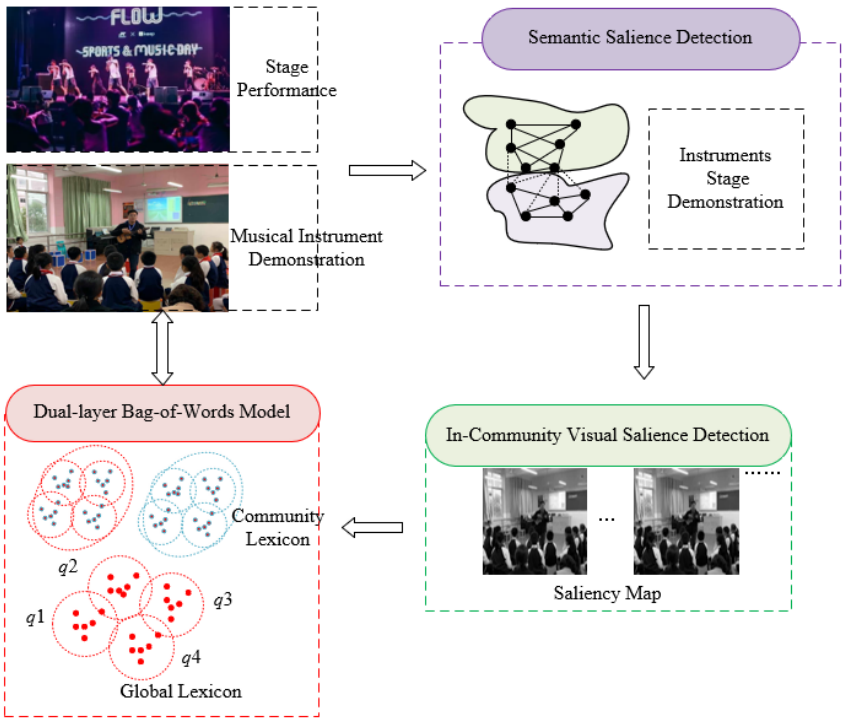

Figure 4. Algorithm offline learning process

Specifically, this paper reflects the local features of images through *SIFT* interest points. Assume the image of size $L*V$ in each semantic community corresponds to a salience map of $\{L_{j,l*v}\}$, where $l<=L$, $v<=V$. The set of local interest points extracted from each image is represented by $\{F_{j,k}\}$, with the number of interest points represented by v_j. Based on the saliency degree of interest point corresponding pixels $L_{j,u}$, the paper generates two codebooks: salient visual and non-salient visual vocabularies. This process can be considered a clustering process, with the distance calculation formula between salient region interest points $F_{j,u}$ and $F_{j,k}$ given as follows:

$$f_{u,k} = \left\| F_{j,u} - F_{j,k} \right\| \exp^{\frac{\left\| L_{j,u} - L_{j,k} \right\|}{\delta}} \tag{13}$$

After generating the "community-global" two-layer visual bag-of-words codebooks, a basic quantification strategy yields a two-layer bag-of-words model based on global salience, i.e., the salient and non-salient two-layer bag-of-words features for each image. Further, the two-layer bag-of-words model is used for training the community classifier. This step integrates local visual salience and global semantic salience, forming a powerful feature representation system for accurately classifying and annotating various communities within teaching content. In remote teaching interaction platforms, such classifier training allows the platform to better identify and annotate key knowledge points in images automatically, offering customized learning paths and materials, optimizing the student learning experience, and the efficiency of teaching resource utilization.

3.2 Online recognition and annotation

The algorithm further processes images to be annotated using the two-layer bag-of-words model, extracting salient features from the images. Through this step, the algorithm can identify key elements in the images, such as objects, people, and scenes, and understand their semantic associations in the teaching content, such as historical background, scientific principles, and cultural significance. In this way, the algorithm can capture the deeper information of image content, which is crucial for the effective classification and subsequent retrieval of teaching resources. For example, when processing images in history teaching, the algorithm can recognize significant cultural artifact features and link them to corresponding historical events or cultural backgrounds. Figure 5 shows the online recognition and annotation process of the algorithm. Specifically, assume the salient bag-of-words feature is represented by $\phi^T\theta_{SA}(U)$, the non-salient bag-of-words feature by $\lambda^T\theta_{UN}(U)$, and the global feature by $\alpha^T\mu(U)$. The community association index for each image U can be defined by the following formula:

$$D(U) = \Phi^T\Theta(U) = \varphi^T\theta_{SA}(U) + \lambda^T\theta_{UN}(U) + \alpha^T\mu(U) \tag{14}$$

Assuming the similarity judgment function is represented by $J_{SA}(U,U_j)$, and an image within a semantic community is represented by U_{ST}. The expression for $\phi^T\theta_{SA}(U)$ is:

$$\varphi^T\theta_{SA}(U) = \sum_{U_j \in U_{ST}} \varphi_j J_{SA}(U,U_j) \tag{15}$$

The expression for $\lambda^T\theta_{UN}(U)$ is:

$$\lambda^T\theta_{UN}(U) = \sum_{U_j \in U_{ST}} \lambda_j J_{UN}(U,U_j) \tag{16}$$

Given the limitations of saliency detection algorithms in processing images without clear salient regions or where foreground-background segmentation is poor, the inclusion of global features provides a solution to these problems. Global features can capture the overall layout, color distribution, texture, etc., of images, which are crucial for images without clear salient regions, helping the algorithm better understand the content of images and their semantic relations in teaching. For example, an image discussing the structure of the Earth might not have clear salient objects, but its global layered structure features are very helpful for identification and classification. The expression for the global feature $\alpha^T\mu(U)$ is:

$$\alpha^S\mu(U) = \sum_{U_j \in U_{ST}} \alpha_j J_{GL}(U,U_j) \tag{17}$$

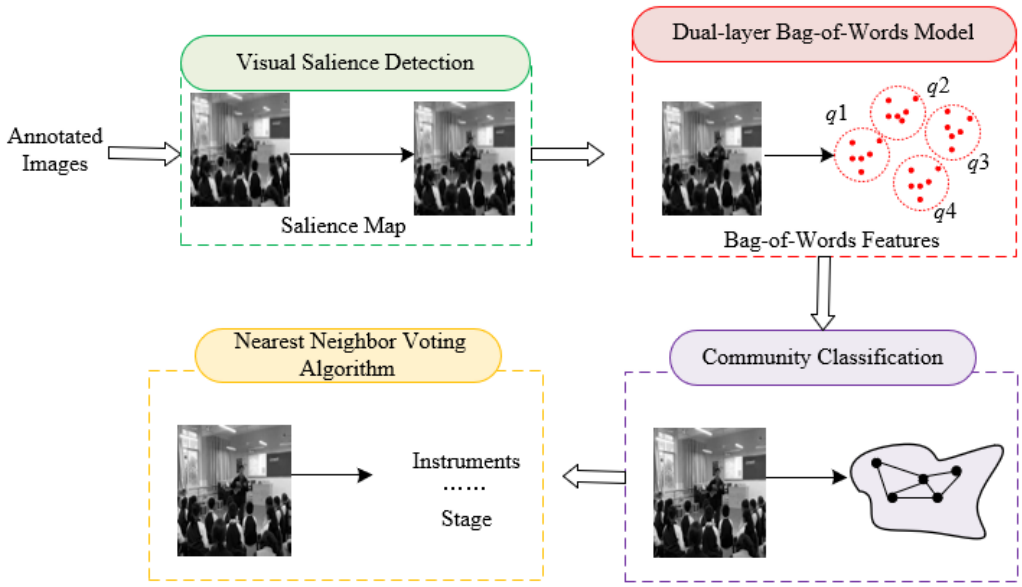

Figure 5. Algorithm online recognition and annotation process

Further, a multi-kernel support vector machine framework is used to solve for the weights of each term. Assuming the weighted result of multiple sub-kernels is represented by $J(\cdot)$. The sub-kernel corresponding to the l-th visual feature is represented by $J_l(\cdot)$, and the weight corresponding to this feature is represented by q_l. The final decision function definition is given as follows:

$$\begin{aligned} D(U) &= \sum_{U_j \in U_{ST}} \varphi_u J_{SA}(U,U_j) + \lambda_u J_{UN}(U,U_j) + \alpha_u J_{GL}(U,U_j) \\ &= \sum_{U_j \in U_{ST}} \beta_j \left\{ \frac{\varphi_j}{\beta_j} J_{SA}(U,U_j) + \frac{\lambda_j}{\alpha_j} J_{UN}(U,U_j) + \frac{\alpha_j}{\beta_j} J_{GL}(U,U_j) \right\} \\ &= \sum_{U_j \in U_{ST}} \beta_j \sum_l q_l J_l(U,U_j) = \sum_{U_j \in U_{ST}} \beta_j J(U,U_j) \end{aligned} \tag{18}$$

To obtain a sparse solution, the paper further adds an $L1$ norm constraint, with the final optimization problem expressed as:

$$\begin{aligned} & MIN \frac{1}{2} \| D \| + Z \sum_{U_j \in U_{ST}} \zeta_j \\ & s.t. D(U) = \sum_{U_j \in U_{ST}} \beta_j J(U,U_j) \\ & J(U,U_j) = \sum_l q_l J_l(U,U_j), q_l \geq 0, \sum_l q_l = 1 \\ & \zeta_j \geq 0, b_j D(U_j) \geq 1 - \zeta_j \end{aligned} \tag{19}$$

The community classifier, based on the features generated in the first step, determines the semantic community corresponding to the image to be annotated. When training the community classifier, the relationship between the content of images and their semantic labels is considered, to classify new images more accurately into existing semantic communities. In remote teaching scenarios, this means images can be automatically classified into the correct teaching modules or conceptual domains, such as natural sciences, literature, art, etc., thereby helping teachers and students quickly locate relevant teaching content and support the construction of modular and personalized learning paths.

At last, the final image label annotation is completed using a nearest neighbor voting strategy. In this step, the algorithm considers instances similar to the image to be annotated among already annotated images. By analyzing the most similar (i.e., "nearest neighbor") set of annotated images, the algorithm adopts a voting method to determine the most probable label. This process is similar to a recommendation system, optimizing annotation accuracy through accumulated similar instances. In remote teaching platforms, this method ensures that even within complex and diverse teaching resources, image annotations are consistent and accurate, allowing students to receive more precise and relevant results when searching or browsing teaching resources.

4. EXPERIMENTAL RESULTS AND ANALYSIS

From Table 1, the method proposed in this paper scores 41.25 in *BLEU* and 55.98 in *METEOR* on the training set, both of which are higher than the other listed multimodal alignment methods. This demonstrates a stronger capability of capturing the alignment between images and text during the training process. On the test set, the proposed method scores 29.6 in *BLEU* and 51.4 in *METEOR*, indicating that the model also has significant generalization capability on unseen data. Although the MBP method scores slightly higher in *METEOR* on the test set, reaching 52.6, the proposed method still maintains the lead in the *BLEU* metric (29.6 compared to 27.8). This highlights the advantage of the proposed method in considering semantic accuracy and linguistic fluency, especially in maintaining a high translation quality (*BLEU* score). The analysis concludes that the multimodal alignment method based on self-attention proposed in this paper has significant performance advantages in multimodal image recognition tasks. With specifically designed source visual encoders and target visual encoders, this method can better process image information related to teaching content and effectively integrate visual and textual cues. This is proven by the high scores on the training set and stable performance on the test set, especially in terms of leading *BLEU* scores, indicating the model's strong capability in translation quality and detail capture.

Table 1. Performance comparison of different multimodal image alignment methods

Methods	Training Set		Test Set	
	BLEU	***METEOR***	***BLEU***	***METEOR***
BITM	34.25	52.68	22.4	44.2
MTN	35.69	53.47	-	-
X-modality SA	37.54	54.12	-	-
DCAN	-	-	28.9	51.2
MBP	37.12	55.89	27.8	52.6
The Proposed Method	41.25	55.98	29.6	51.4

Table 2 provides the results of the ablation study to analyze the importance of different components in the proposed method. It is shown that after removing both the source visual consistency module and the target visual consistency module, the model's scores on the training set drop to 34.21 in *BLEU* and 52.69 in *METEOR*, and the scores on the test set also decrease to 23.1 and 44.1, indicating the significant role these two modules play in the model. When the target visual consistency module is removed alone, the scores on the test set decrease to 27.8 in *BLEU* and 52.3 in *METEOR*; removing only the source visual consistency module further decreases these scores to 26.9 and 47.8. This indicates that the source visual consistency module is especially important for maintaining model performance on the test set, likely playing a key role in helping the model capture and understand the content of source images. In contrast, while the target visual consistency module contributes to performance improvement, its impact seems not as significant as the source visual consistency module. Based on these experimental results, we can conclude that the source visual consistency module and the target visual consistency module in the proposed multimodal alignment method based on self-attention play crucial roles in enhancing model performance. These two modules work together to improve the accuracy of alignment between teaching content and image information, thereby enhancing the interactivity and teaching adaptability in remote teaching interaction platforms. Specifically, the source visual consistency module is more critical for performance improvement, likely because it directly affects the model's ability to understand the content of source images.

Table 2. Ablation study results

Methods	Training Set		Test Set	
	BLEU	***METEOR***	***BLEU***	***METEOR***
- Source Visual Consistency Module - Target Visual Consistency Module	3421	52.69	23.1	44.1
- Target Visual Consistency Module	37.85	55.41	27.8	52.3
- Source Visual Consistency Module	37.59	54.89	26.9	47.8
The Proposed Method	41.26	55.36	28.9	51.4

Table 3. Multimodal consistency verification of image multimodal alignment methods in different scenarios

Methods	Classroom Demonstration Scenario		Student Interaction Q&A Scenario		Personalized Learning Path Recommendation Scenario	
	BLEU	***METEOR***	***BLEU***	***METEOR***	***BLEU***	***METEOR***
- Source Visual Consistency Module- Target Visual Consistency Module	47.56	66.93	32.14	51.23	42.36	61.24
- Target Visual Consistency Module	48.96	67.89	31.26	53.47	43.18	61.24
- Source Visual Consistency Module	48.62	68.25	32.69	52.48	43.59	62.39
The Proposed Method	51.23	71.45	32.15	53.69	45.26	62.58

Table 4. Performance comparison of different multimodal image annotation and recognition methods

Methods	***Precision***	***Recall***	***F1-score***
TrIC	0.121	0.224	0.159
BUTD	0.121	0.312	0.159
ORT	0.189	0.178	0.186
M2 Transformer	0.178	0.179	0.183
CRF-Nets	0.123	0.169	0.151
GCN-MML	0.223	0.231	0.234
The Proposed Method	0.254	0.278	0.263

From the data provided in Table 3, it is evident that the multimodal alignment method proposed in this paper outperforms other comparative methods across different scenarios. Specifically, in the classroom demonstration scenario, the proposed method scores 51.23 in *BLEU* and 71.45 in *METEOR*, which are higher than the scores obtained by methods without the source visual consistency module or the target visual consistency module. This indicates that the proposed method is more precise in processing visual information closely related to teaching content. In the student interaction Q&A scenario, although the *BLEU* score of the proposed method is comparable to the method without the source visual consistency module (32.15 vs 32.69), it shows improvement in the *METEOR* score (53.69 vs 52.48), highlighting the proposed method's advantage in understanding student questions and providing relevant feedback. In the personalized learning path recommendation scenario, the proposed method also exhibits higher *BLEU* and *METEOR* scores (45.26 and 62.58, respectively), indicating that the proposed method can provide visual content that matches individual student learning needs more effectively. Overall, the experimental results demonstrate a clear advantage of the proposed method in multimodal consistency verification, offering more precise and efficient synchronous expression of images and teaching content across different remote teaching interaction scenarios. Comparative analysis leads to the conclusion that the multimodal alignment method based on self-attention proposed in this paper significantly enhances the accuracy of synchronous expression of images and teaching content in various remote teaching scenarios. This is not only reflected in the higher *BLEU* and *METEOR* scores compared to comparative methods but also in the model's finer semantic understanding and integration of visual information when processing multimodal data. By accurately aligning visual and semantic information, the method not only enhances the accessibility and interactivity of remote teaching content but also provides more personalized and adaptable support for teaching.

Table 4 shows the performance comparison of the proposed method with other multimodal image annotation and recognition methods in terms of Precision, Recall, and F1 score. It can be observed that the proposed method outperforms the others in all three evaluation metrics, with a Precision of 0.254, Recall of 0.278, and an F1 score of 0.263. In contrast, the best-performing method among the others is GCN-MML, with Precision, Recall, and F1 scores of 0.223, 0.231, and 0.234, respectively. This indicates that in multimodal image annotation and recognition tasks, the proposed method can more effectively identify key information in images and accurately annotate them with text. Synthesizing these results, we can conclude that the multimodal image annotation and recognition algorithm proposed in this paper, based on semantic and visual saliency, significantly improves the performance of image annotation tasks. This improvement is attributed to the algorithm's ability to accurately extract visually salient regions in images and effectively align these regions with corresponding textual content. Through such alignment, the algorithm not only captures key visual information in images but also ensures semantic consistency between this information and text labels. Additionally, the application of multi-kernel SVM further

enhances the model's learning capability for integrated features, while the nearest neighbor voting strategy effectively improves the accuracy of annotations.

According to the data provided in Figure 6, the performance of the proposed method under different semantic salience threshold settings exhibits certain fluctuations. When the threshold is set to 0, the F1 score is 0.23. As the threshold increases, the F1 score gradually rises, reaching its peak at a threshold of 0.4 with an F1 score of 0.273. However, when the threshold continues to increase to 0.5 and above, the F1 score begins to significantly decrease, dropping to 0.165 when the threshold is at 1. This indicates that the combination of semantic salience and visual salience is most effective when the threshold is between 0.3 and 0.4, allowing for more accurate identification and annotation of important information in images. This may be due to moderate semantic filtering being able to eliminate text content that does not match well with visually salient regions of the image, thereby improving overall recognition precision and annotation consistency. From these results, it can be concluded that the algorithm proposed in this paper achieves optimal performance under appropriate semantic salience threshold settings, emphasizing the importance of setting suitable semantic salience thresholds for multimodal image recognition and annotation methods. Proper threshold settings not only ensure that the textual content captured by the algorithm remains highly consistent with the visually salient regions of the image but also prevent the loss of important information due to too high thresholds or the introduction of irrelevant noise due to too low thresholds.

The experimental results from Figure 7 show a clear upward trend in the F1 score of the proposed method as the dimensionality of the salience bag-of-words increases. When the dimensionality increases from 100 to 900, the F1 score gradually rises from 0.18 to 0.29. This indicates that as the dimensionality of visual bag-of-words features increases, the model can capture the visual salience features of images more exhaustively, thereby achieving better performance in image recognition and annotation tasks. At lower dimensionalities, the model may not fully express the complexity and diversity of image content, leading to lower recognition and annotation accuracy. However, as the dimensionality increases, the richness of features can better depict key information in images, enhancing the algorithm's recognition capability and annotation accuracy. Comprehensive analysis of the results indicates that the multimodal image recognition and annotation algorithm proposed in this paper, by extracting and utilizing a combination of high-quality visual features and semantic information, can significantly improve annotation and recognition performance with the increase in the dimensionality of the salience bag-of-words. This outcome proves that the algorithm is more effective in processing finer-grained features, demonstrating excellent feature representation ability and classification performance.

The experimental results from Figure 8 indicate a slow but steady increase in the F1 score of the proposed method as the dimensionality of the non-salient bag-of-words increases, from 0.255 to 0.28. This result suggests that even non-salient bag-of-words features contribute to enhancing the algorithm's annotation and recognition performance as the dimensionality increases from 100 to 2000. It can be inferred that while non-salient features may not contain information as directly relevant as salient features, at higher dimensions, they still provide useful contextual information for the model, assisting in a more comprehensive understanding of image content and improving the precision of semantic category recognition and annotation quality. These results emphasize that information from non-salient regions should not be entirely disregarded, as they may contain supplementary information beneficial to the overall recognition and annotation tasks. By integrating these non-salient features, the proposed algorithm demonstrates its effectiveness in enhancing multimodal image recognition and annotation performance. Moreover, the performance improvement with increasing bag-of-words dimensionality shows the algorithm's capability to handle large-scale features, effectively refining key information while ensuring comprehensive information processing, thus offering a robust methodological foundation for multimodal image processing.

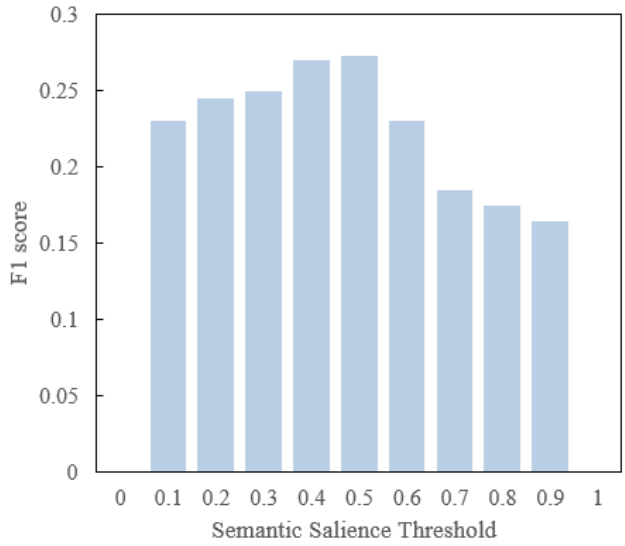

Figure 6. Impact of semantic salience threshold settings on the annotation and recognition performance of the proposed method

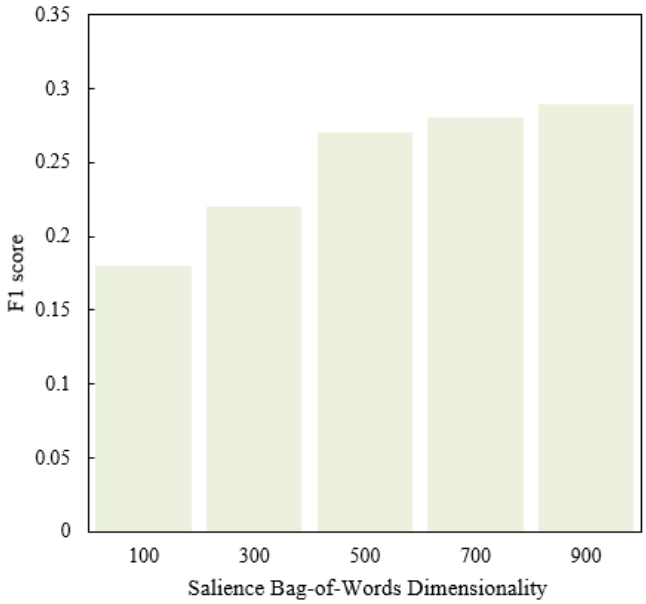

Figure 7. Impact of salience bag-of-words dimensionality on the annotation and recognition performance of the proposed method

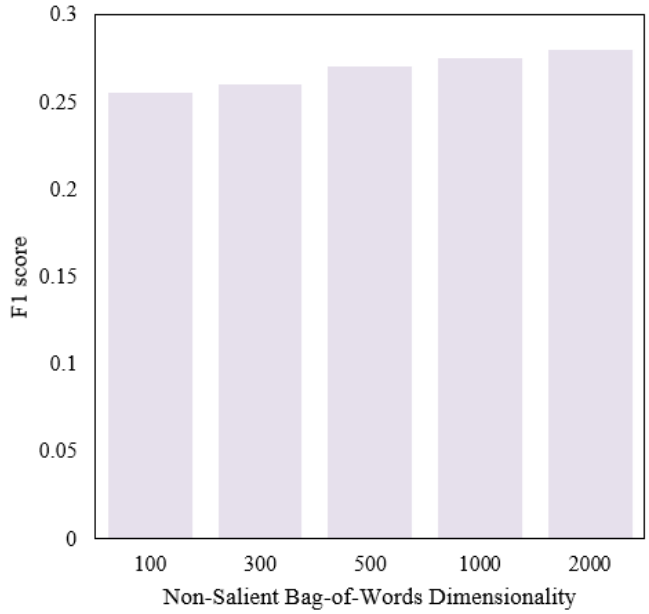

Figure 8. Impact of non-salient bag-of-words dimensionality on the annotation and recognition performance of the proposed method

5. CONCLUSION

This paper focuses on improving the understanding of image content in remote teaching and the practicality of interaction platforms through multimodal alignment and image annotation and recognition methods. By designing a novel multimodal alignment method based on self-attention, this study achieves highly synchronous expression of visual information with teaching content, which is crucial in remote teaching interactions as it directly impacts students' learning efficiency and interaction quality. Furthermore, the proposed multimodal image annotation and recognition algorithm, which combines semantic and visual saliency, significantly enhances the accuracy of understanding image content by meticulously distinguishing between salient and non-salient features.

Experimental results validate the effectiveness of the proposed methods through comparative analysis with other multimodal alignment and annotation recognition methods, especially in discussions regarding the impact of semantic salience threshold settings and the dimensionality of salient and non-salient bag-of-words on algorithm performance, further highlighting the superiority of the proposed approach. These experiments not only prove the efficacy of the method but also provide important reference data and methodologies for future multimodal image processing research.

Despite significant achievements, this study has some limitations. For example, the model may be overly optimized for specific datasets or types of images, while its generalizability to a broader range of real-world applications remains untested. Additionally, the high-dimensional feature processing and self-attention mechanism increase the model's computational complexity, potentially affecting the smoothness of real-time interactions. Future research directions could include further exploring the model's generalizability to accommodate a more diverse range of image content and remote teaching scenarios, optimizing model structures and algorithms to reduce computational resource consumption and enhance real-time processing capabilities, and investigating more data augmentation and transfer learning techniques to strengthen the model's performance in few-shot learning. Through these studies, future work hopes to advance the development of remote teaching interaction platforms, improving teaching quality and learning experiences while building on the current research outcomes.

FUNDINGS

The 2022 Xinjiang Uygur Autonomous Region Philosophy and Social Science Fund Project "Research on the Cultivation System of Digital Literacy for Farmers in Southern Xinjiang under the Background of Rural Revitalization" (Grant No.: 22BJYX076).

REFERENCES

[1] Wu, X., Sun, G. (2024). Interactive college drama teaching based on internet remote technology. International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 19(1): 1-11. https://doi.org/10.4018/IJWLTT.336837

[2] Dallal, A., Zaghloul, M.A., Hassan, A. (2021). New instructors perspectives on remote teaching methods. In 2021 ASEE Virtual Annual Conference Content Access. https://doi.org/10.18260/1-2--37525

[3] Marinelli, A., Papile, F., Sossini, L., Del Curto, B. (2023). Enhancing active learning in remote collaboration: An experience in teaching functional materials. International Journal of Mechanical Engineering Education, 51(1): 4-22. https://doi.org/10.1177/03064190221143312

[4] Cavanlit, K.L., Encabo, E.M., Vilbar, A. (2023). Using recorded lectures in teaching higher education in an online remote learning context. In Novel & Intelligent Digital Systems Conferences, pp. 187-194. https://doi.org/10.1007/978-3-031-44097-7_20

[5] Paea, S., Sharma, B., Bulivou, G., Katsanos, C. (2024). Emergency remote teaching in higher education institutes: A taxonomy of challenges faced by first-year mathematics students in the pacific region. IEEE Access, 12: 6339-6355. https://doi.org/10.1109/ACCESS.2024.3351098

[6] Yan, J., Wang, N., Wei, Y., Han, M. (2023). Personalized learning pathway generation for online education through image recognition. Traitement du Signal, 40(6): 2799-2808. https://doi.org/10.18280/ts.400640

[7] Zhang, X., Yuan, J., Li, L., Liu, J. (2023). Reducing the bias of visual objects in multimodal named entity recognition. In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, pp. 958-966. https://doi.org/10.1145/3539597.3570485

[8] Cai, J., Lin, Y., Ma, R. (2023). Multimodal emotion recognition based on long-distance modeling and multi-source data fusion. In 2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE), Ottawa, ON, Canada, pp. 497-503. https://doi.org/10.1109/CIPAE60493.2023.00100

[9] Zhang, P., Fu, M., Zhao, R., Wu, D., Zhang, H., Yang, Z., Wang, R. (2023). ECMER: Edge-cloud collaborative

personalized multimodal emotion recognition framework in the internet of vehicles. IEEE Network, 37(4): 192-199. https://doi.org/10.1109/MNET.003.2300012

[10] Li, J., Wang, X., Lv, G., Zeng, Z. (2023). GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation. Neurocomputing, 550: 126427. https://doi.org/10.1016/j.neucom.2023.126427

[11] Yang, L., Jia, Y. (2022). Leg posture correction system for physical education students based on multimodal information processing. In International Conference on E-Learning, E-Education, and Online Training, pp. 91-102. https://doi.org/10.1007/978-3-031-21161-4_8

[12] Yu, Z. (2021). Research on Multimodal Music Emotion Recognition Method Based on Image Sequence. Scientific Programming, 2021: 7087588. https://doi.org/10.1155/2021/7087588

[13] Zheng, J., Zhang, S., Wang, Z., Wang, X., Zeng, Z. (2022). Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition. IEEE Transactions on Multimedia., 25: 2213-2225. https://doi.org/10.1109/TMM.2022.3144885

[14] Chen, S., Rao, B. Y., Herrlinger, S., Losonczy, A., Paninski, L., Varol, E. (2023). Multimodal microscopy image alignment using spatial and shape information and a branch-and-bound algorithm. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1-5. https://doi.org/10.1109/ICASSP49357.2023.10096185

[15] Öfverstedt, J., Lindblad, J., Sladoje, N. (2022). Fast computation of mutual information in the frequency domain with applications to global multimodal image alignment. Pattern Recognition Letters, 159: 196-203. https://doi.org/10.1016/j.patrec.2022.05.022

[16] Wang, X., Shu, K., Kuang, H., Luo, S., Jin, R., Liu, J. (2021). The role of spatial alignment in multimodal medical image fusion using deep learning for diagnostic problems. In Proceedings of the 2021 International Conference on Intelligent Medicine and Health, pp. 40-46. https://doi.org/10.1145/3484377.3484384

[17] Landolsi, M.Y., Haj Mohamed, H., Ben Romdhane, L. (2021). Image annotation in social networks using graph and multimodal deep learning features. Multimedia Tools and Applications, 80: 12009-12034. https://doi.org/10.1007/s11042-020-09730-8

[18] Zhu, S., Li, X., Shen, S. (2015). Multimodal deep network learning-based image annotation. Electronics Letters, 51(12): 905-906. https://doi.org/10.1049/el.2015.0258

[19] Guerrero, R.E.D., Gupta, Y., Bocklitz, T., Oliveira, J. L. (2023). A multimodal image registration system for histology images. In 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS), L'Aquila, Italy, pp. 17-22. https://doi.org/10.1109/CBMS58004.2023.00185

[20] Miao, Z., Cheng, C. (2023). Construction of multimodal music automatic annotation model based on neural network algorithm. In Seventh International Conference on Mechatronics and Intelligent Robotics (ICMIR 2023), Kunming, China, pp. 482-488. https://doi.org/10.1117/12.2689482