International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 1079-1086

https://doi.org/10.70517/ijhsa464100

Spatio-temporal Feature Extraction and Knowledge Graph Construction Method for Foreign Language Teaching Video Content Understanding

Zuqiao Wei[1,*]
[1] Guilin University of Technology, Guilin, Guangxi, 541004, China
Corresponding authors: (e-mail: gxweizuqiao@sina.com).

Abstract The development of online education has continuously promoted the level of intelligent analysis of foreign language teaching videos. This paper proposes a content comprehension analysis method for foreign language teaching videos that integrates spatio-temporal residual attention network (STRAN) and improved text rank algorithm (TextRank). The spatio-temporal features in the video are extracted by three-dimensional residual network (3DResNet), and the attention mechanism is combined to optimize the recognition accuracy of students' classroom expressions. Improved TextRank algorithm is utilized to filter the keywords of teaching focus and integrate multi-source data and storage technology to construct a knowledge graph of foreign language teaching. The results show that the recognition rate of all 6 kinds of student expressions of STRAN is more than 0.930. The 3 keyword extraction performance indexes of Improved TextRank are all greater than 90.00% in different numbers of keyword extraction. Comprehensively applying the method of this paper to assist teaching, the students' concentration and excitement expression scores are more than 90 points, and the scores of confused, distracted, nervous, and sleepy expressions are around 70-85 points.

Index Terms STRAN, TextRank algorithm, spatio-temporal features, expression recognition, knowledge mapping for foreign language teaching

I. Introduction

In recent years, with the rapid development and wide application of digital video media, how to efficiently and accurately understand and analyze video content has become a research hotspot in the field of computer vision and multimedia. Automatic comprehension and analysis of video media content is not only valuable in applications such as intelligent surveillance, video retrieval, and automatic driving, but also has far-reaching significance in promoting the progress of educational technology [1]-[4]. Video content feature extraction and integration technology puts higher requirements on new methods of teaching and learning, which determines that the rapid development of foreign language teaching videos has become a necessity [5], [6]. Platforms such as TeachView.com and TeachersListening.com appeared in time to become the basis for the development of teaching videos, recording and uploading quality classroom teaching content to the website for reference and learning by primary and secondary school teachers, students and the community in need, and preliminary results have been achieved [7]-[10].

However, traditional methods for video understanding and analysis mainly rely on manual feature extraction and shallow learning models, which often show limitations when facing complex and changing scenarios, making it difficult to adequately capture high-level semantic information in videos [11]-[13]. Research on feature extraction techniques for automatic comprehension and analysis of video media content is promoted to construct a knowledge graph for foreign language teaching, in order to improve the effectiveness and robustness of existing methods when dealing with complex foreign language teaching video scenarios [14]-[16].

In this paper, spatio-temporal residual attention network (STRAN) is constructed based on three-dimensional residual network (3DResNet). Optimize gradient propagation by residual connection and expansion convolution to solve the long-time dependency problem in foreign language teaching video data and improve the accuracy of students' classroom expression recognition. Design an improved TextRank algorithm, combined with lexical filtering and dynamic window strategy, to accurately extract the keywords of the knowledge points explained by the teacher in the video. Utilizing Jaccard coefficient and cosine similarity to fuse multi-source knowledge to enhance the knowledge breadth and depth of foreign language teaching knowledge graph. The storage and visualization of foreign language teaching knowledge is realized through Neo4j graph database, and a complete and applicable foreign language teaching knowledge graph is constructed.

II. Technical realization of spatio-temporal feature extraction and knowledge map construction

II. A.A model for recognizing students' classroom expressions based on spatio-temporal residual attention network

The backbone of the spatio-temporal residual network consists of 3DResnet, a three-dimensional convolutional architecture based on the residual network, ResNet, one of the most successful architectures in image classification, which provides shortcut connections that allow the signal to bypass one layer and move sequentially to the next. Figure 1 illustrates the basic residual model structure.

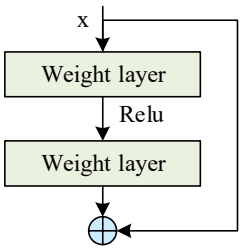

Figure 1: Basic residual model

Since these connections flow through the gradient of the network from later to earlier layers, they can facilitate the training of very deep networks. In turn, 3D convolution is able to extract spatio-temporal features directly from the original video. So fusing the two can effectively solve the problems of gradient explosion, slow convergence, and inability to extract spatio-temporal features that may result as the depth of the network increases when foreign language classroom teaching video data is used as the input to the network, and the basic principle of the residual block is shown in Equation (1).

$$H(x) = F(x) + x \tag{1}$$

where $H(x)$ represents the deep output and $F(x)$ represents the transformation of the two representations sandwiched between the shallow and deep layers, the residual block can be represented as (1).

And the basic form in each residual block can be represented as equation (2), where $h(x)$ is the constant mapping, F is the variation in the network, $f(x)$ is the transformation for the value after superposition, and in the original residual block is the Relu, where the network reduces the loss value by learning the parameters of F in it. And for a deeper layer L, its relation to the l layer can be expressed as (3), where the L layer is denoted as the sum of the residual parts between any l layer shallower than it and it.

$$\begin{cases} y_l = h(x_l) + F(x_l, W_l) \\ x_{l+1} = f_{(y_l)} \end{cases} \tag{2}$$

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i) \tag{3}$$

According to the chain rule of derivatives used in multilayer feedforward neural networks following the error backpropagation algorithm, the gradient of the loss function ε with respect to x_l can be expressed as (4), and it can be concluded that throughout the entire training process of the network, the gradient of $\sum_{i=l}^{L-1} F(x_i, W_i)$ cannot be -1.0 all the time, so the problem of vanishing gradients does not occur among residual networks.

$$\frac{\partial \varepsilon}{\partial x_l} = \frac{\partial \varepsilon}{\partial x_L} + \frac{\partial \varepsilon}{\partial x_L} \frac{\partial}{\partial x_l} \sum_{i=l}^{L-1} F(x_i, W_i) \tag{4}$$

In STRAN, the inflated ResnNet is used as the backbone network for 3D, and the native ResnNet, which consists of two convolutional layers, each followed by a BN and ReLU, is used as the backbone network for 2D. Then, the output features of the 3D convolutional operations are used as inputs to the spatial module, which are spatially compressed by different multiples compared to the input frames, respectively.

II. B.Knowledge map construction

II. B. 1) Classroom Knowledge Extraction

In this paper, we adopt the method of constructing foreign language teaching content knowledge map on the basis of knowledge base, and the first point of constructing knowledge map is that the entity extraction corresponds to the classroom knowledge points in the teaching content.

At present, there are many mature technologies and models that can accurately extract the entities in the text, but the existing natural language processing tools are limited to the accurate recognition of the names of people, places and institutions, and are not applicable to the extraction of classroom knowledge points. The TextRank algorithm is a classic keyword extraction algorithm in natural language processing, which has a better effect on the extraction of Chinese keywords.TextRank algorithm is derived from PageRank, which constructs a network through the neighbor relationship between words, then iteratively calculates the rank value of each node with PageRank, and then the rank value is obtained. The algorithm of TextRank for keyword extraction is as follows.

Step 1: Split the given text T according to the complete sentence, i.e.

$$T = [S_1, S_2, \ldots S_m] \tag{5}$$

Step 2: For each sentence S_i belonging to T, perform the word splitting and lexical labeling process, and filter out the deactivated words and keep only the words with specified lexical properties, such as nouns, verbs, and adjectives, i.e.

$$S_i = [t_{i,1}, t_{i,2}, \ldots t_{i,m}] \tag{6}$$

where $t_{i,j}$ is the candidate keyword after retention.

Step 3: Construct the candidate keyword graph $G = (V, E)$, where V is the set of nodes, which consists of the candidate keywords generated in the second step, and then construct the edges between any two points by using co-occurrence relation, and the existence of edges between two nodes is only possible when their corresponding words are co-occurring in the window of length K, where K denotes the size of the window, i.e., at most K words are co-occurring.

Step 4: Based on the above formula, iteratively propagate the weights of each node until convergence.

Step 5: Sort the node weights in reverse order to get the most important T words as candidate keywords.

Step 6: The most important T words obtained from Step 5 are tagged in the original text and combined into multiple keywords if they form neighboring phrases.

The TextRank algorithm is used to extract the keywords in the teacher's words of foreign language teaching videos. Extracting keywords is set to sort only nouns by TextRank, only nouns may be related to teaching knowledge points, TextRank algorithm calculates importance scores for nouns, combined with the keyword word frequency, it is found that when the TextRank score is lower than 0.10 nouns, the word frequency is almost 1.0, so the nouns with TextRank scores greater than 0.10 are used as foreign language keywords in the classroom teaching videos, which are the classroom knowledge points.

II. B. 2) Knowledge integration

There are two main sources of data in this paper: one from foreign language teaching videos and the other from Baidu Encyclopedia website data. When the two data sources are combined together, problems such as data inconsistency, data quality differences, data duplication and omission may occur. For example, the knowledge entity "British Culture" obtained in the foreign language teaching video is obtained, and the knowledge entity obtained in Baidu Encyclopedia is "English Country Culture", both of which point to "English Country Culture". Therefore, knowledge fusion of data is needed so as to integrate information from multiple data sources and improve the comprehensiveness and consistency of knowledge.

According to the relevant theories of knowledge fusion, this paper adopts the entity alignment approach to knowledge fusion of entity data in the field of foreign language teaching. Figure 2 shows the knowledge fusion design method.

In entity alignment, this paper adopts the Jaccard coefficient to measure the similarity between entities.Jaccard coefficient is a metric used to measure the similarity between two sets, which is determined by calculating the proportion of characters with the same text in the two sets to the whole sentence. The formula is shown below:

$$sim_t(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A + B + A \cap B|} \tag{7}$$

Where $sim_t(A, B)$ represents the similarity between the sets of entities A and B, $|A \cap B|$ denotes the number of elements in the intersection of the set A and the set B, and $|A \cup B|$ denotes the number of elements in the concatenation of the set A and the set B.

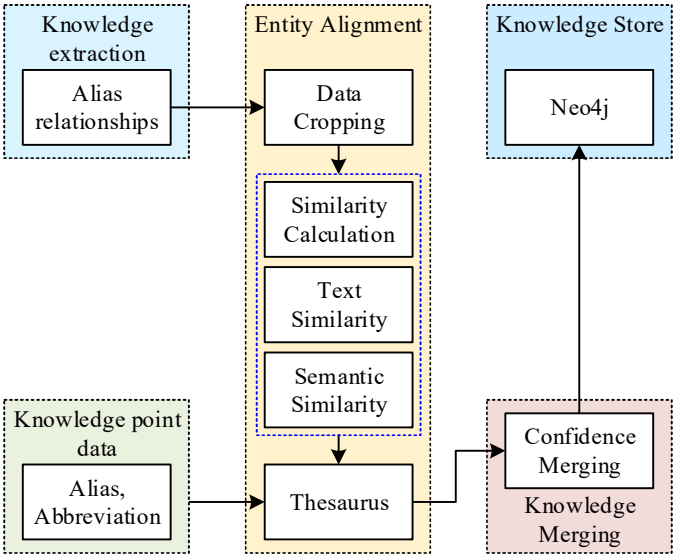

Figure 2: Knowledge Fusion Design Method

In text similarity computation, the text is usually converted to a vector representation and then the Jaccard coefficients are used to compare these vectors. In this paper, the cosine similarity is used to perform the calculation, and the formula is shown below:

$$sim_s(A,B)=\cos(\theta)=\frac{A\cdot B}{|A|\times|B|}=\frac{\sum_i^n A_i\times B_i}{\sqrt{\sum_i^n (A_i)^2}\times\sqrt{\sum_i^n (B_i)^2}} \tag{8}$$

where $sim_s(A,B)$ represents the semantic similarity between the sets of entities A and B, and $\cos(\theta)$ represents the cosine similarity, where the smaller the angle θ between the vectors, the more similar the two entities are. A_i represents the word vectors of the set A and B_i represents the word vectors of B.

Finally, this paper combines the similarity of these two ways and sets different weights on them respectively. The effect of different weights is verified through experiments, and then the weights are set according to the experimental results. Setting the weight of $sim_t(A,B)$ to 0.40 and the weight of $sim_s(A,B)$ to 0.60 can get more similarity knowledge point entities. Finally, the similarity results are obtained and the entities with high data similarity are fused, and the calculation formula is shown below:

$$sim(A,B)=0.4\times sim_t(A,B)+0.6\times sim_s(A,B) \tag{9}$$

II. B. 3) Knowledge storage

The processed data will be stored.Neo4j is a kind of graph-based database, which employs nodes and relations to describe the relationship between data, and is therefore very suitable for storing knowledge graph data. In the process of constructing the knowledge graph in the field of foreign language teaching, the node-based storage method is used, specifically, the knowledge points in the field of foreign language teaching are stored as nodes in the Neo4j database.

The Neo4j database can be imported in various ways, including using the official Neo4j-admin-import tool, importing the CSV file with ternary data through the Cypher load csv statement, and writing the Cypher creater statement. Among them, writing Cypher creater statement for importing is too slow and not suitable for importing knowledge point data with large amount of data. The official Neo4j-admin-import tool not only needs to close the Neo4j database before importing, but also needs to initialize the database before importing. So this paper adopts the way of Cypher load csv statement to import, and at the same time combines the import method of writing Cypher creater statement to supplement the missing data.

III. Spatio-temporal feature extraction and knowledge graph application practice and analysis

III. A. Analysis of the effect of STRAN-based expression recognition

In order to test the performance of spatio-temporal residual attention network STRAN for student expression recognition in foreign language teaching videos, the foreign language teaching classroom videos of 200 freshmen majoring in Business English in a university are used as the research dataset, and STRAN is utilized for student

expression recognition. Figure 3 shows the confusion matrix of STRAN's recognition results for six common student classroom expressions. The recognition accuracies of the six common expressions of concentration, confusion, distraction, excitement, nervousness, and fatigue reach 0.942, 0.957, 0.936, 0.938, 0.939, and 0.945, which are all higher than 0.930 and the highest reaches 0.957. At the same time, the misrecognition rates of the six types of expressions are all lower than 0.010, and the possibility of misrecognition of the expressions for categorization is extremely small. This indicates that the spatio-temporal residual attention network STRAN in this paper can more accurately recognize the expressions appearing in students' foreign language teaching classroom videos, which lays a good foundation for accurately judging students' learning of specific knowledge points.

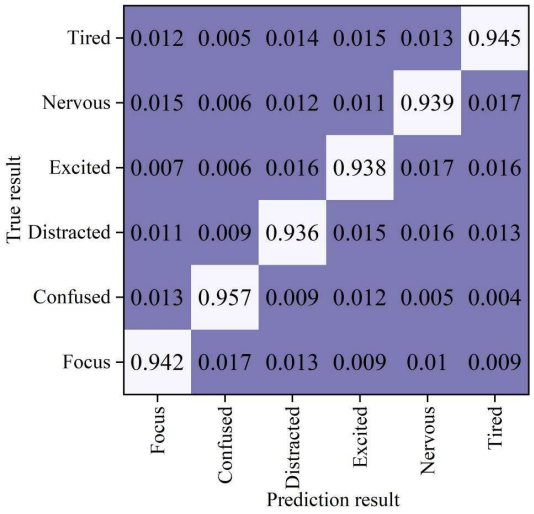

Figure 3: Confusion matrix of the recognition results of students' expressions

III. B. Comparison of the effect of knowledge point extraction based on TextRank algorithm

III. B. 1) Comparison of speech enhancement effects

In order to better use the TextRank algorithm to extract the keywords in the teacher's discourse of the foreign language teaching video, the teacher and student speech samples in the video are filtered and preprocessed to remove the classroom noise in the speech samples, so as to make the speech samples more pure. Figure 4 shows the comparison of speech enhancement effect before and after preprocessing. 0-15s of speech samples, the original speech waveform in the connection part of a large number of fine fluctuations, that is, there is more interference noise, for example, the position of the 14-15s by a large number of fine fluctuations, which represents that the keywords may be extracted inaccurately at that location. After the noise reduction process, the minor fluctuations in the connection part are basically removed, the interference noise is reduced, and comparing the fluctuations at the 14-15s, it can be seen that the waveform is basically smooth, and the keywords mentioned by the teacher can be extracted smoothly at that location.

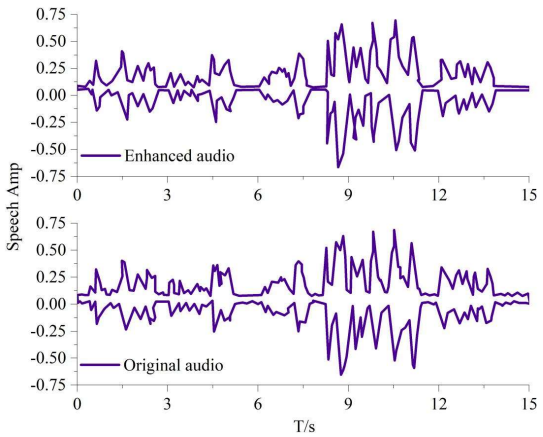

Figure 4: Comparison of voice enhancement effects

III. B. 2) Keyword extraction performance comparison

Based on the enhanced processed speech samples, this paper uses the TF-IDF keyword extraction algorithm based on word frequency statistics, the traditional TextRank keyword extraction algorithm, and the improved TextRank algorithm proposed in this paper to extract the first 4, 6, 8, 10, and 15 keywords mentioned by the teachers in the classroom, and adopts the three evaluation indexes, namely, accuracy, recall, and F-value, to assess its keyword extraction effect. Table 1 shows the test results. The accuracy rate of this paper's algorithm reaches 90.89%, 90.11%, 90.05%, 90.01%, 90.00% for different numbers of keyword extraction tests; the recall rate is 91.36%, 91.02%, 90.67%, 90.43%, 90.19%; and the F-value is 92.17%, 91.39%, 90.95%, respectively, 90.54%, 90.20%. It always stays above 90.00%. And in the case of the increasing number of keyword extraction, the performance of the three indicators decreases more slowly. It can be seen that the improved TextRank algorithm used in this paper has a better keyword extraction ability than the comparison algorithm, and can accurately extract the keywords of teachers in foreign language classroom teaching videos to construct a knowledge graph.

Table 1: Test results of keyword extraction performance

Evaluation index	Algorithm	Top 4	Top 6	Top 8	Top 10	Top 15
Accuracy rate (%)	TF-IDF	50.12	48.34	46.92	44.72	42.38
	TextRank	71.65	69.87	68.40	66.41	64.03
	Improve TextRank	90.89	90.11	90.05	90.01	90.00
Recall rate (%)	TF-IDF	47.92	46.14	44.74	42.67	40.33
	TextRank	63.78	62.25	60.85	58.83	56.49
	Improve TextRank	91.36	91.02	90.67	90.43	90.19
F value (%)	TF-IDF	52.04	50.26	48.86	46.76	44.42
	TextRank	67.20	65.42	64.02	61.65	59.31
	Improve TextRank	92.17	91.39	90.95	90.54	90.20

III. C. Analysis of the effect of teaching aids

III. C. 1) Learning outcomes for individual students

The constructed knowledge map for foreign language teaching is used to assist students in foreign language learning, while different expressions appearing when students use the knowledge map to assist their learning are recognized through the spatio-temporal residual attention network. The average score of students in learning specific knowledge points is calculated, thus determining the students' mastery of the knowledge point, which facilitates teachers to optimize the knowledge map subsequently. Take Freshman A, who majored in Business English in the experimental university, as an example to show how individual students use the method of this paper to assist their learning. Figure 5 shows the learning effect of student A. Due to the large number of subknowledge points covered by the four knowledge points of business writing, business communication, business etiquette, and business strategy, the six expression changes that occur when students learn these knowledge points are also more complex. For example, in business writing, the 6 expressions of concentration, confusion, distraction, excitement, nervousness, and fatigue appeared alternately with the deepening of learning, with scores of 90.42, 82.41, 88.35, 89.54, 86.56, and 92.10, respectively.The students went from concentrating on their learning at the beginning, to appearing confused and distracted, to experiencing a short period of excitement through the learning of the knowledge map, and after that, to into intense learning, to a more exhausted state after finally learning the knowledge point. This shows that there are more and more difficult knowledge points about business writing in the knowledge map, so in addition to using the knowledge map to assist learning, teachers should also explain the teaching appropriately to help students improve their interest in learning.

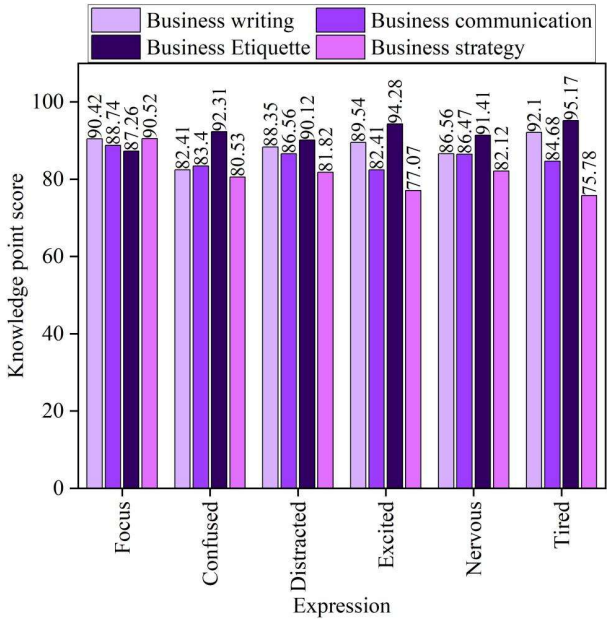

Figure 5: The learning effect of Student A

III. C. 2) Learning effects of grade level collectives

Figure 6 shows the knowledge point scores (averaged) of the first-year students after using the method of this paper to assist their learning. Among the six expressions that appeared in learning the four knowledge points, the experimental students scored 92.71, 93.15, 93.46, and 92.78 for the Focused expression, which were all greater than 90. Confused expression scores 79.23, 80.32, 78.15, 78.42, which are around 80. Walking away expression scores 70.89, 71.43, 70.45, 71.26, all less than 75. Excited expression scores 92.52, 92.57, 93.13, 93.47, all greater than 90. Nervous expression scores 70.34, 71.26, 70.25, 71.41, and sleepy expression scores 72.89, 75.17, 71.48, 72.37, all less than 80. By comparing the scores, it is judged that the students are using the constructed knowledge map for assisted learning, the positive expressions (concentration, excitement) have higher scores, and the negative expressions (confusion, distraction, nervousness, sleepiness) have lower scores, and the method of this paper can improve the students' foreign language learning effect.

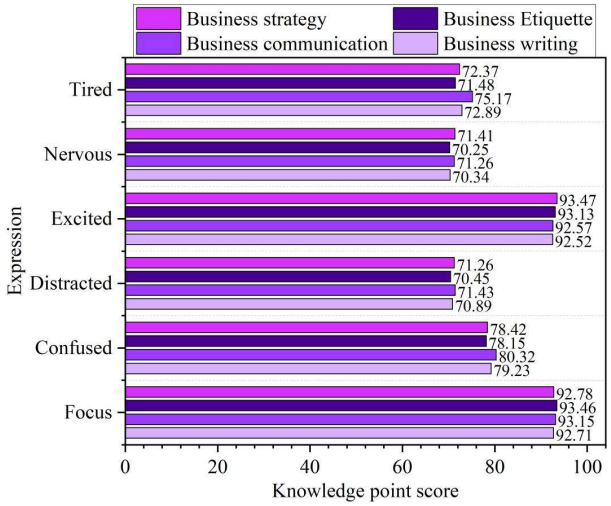

Figure 6: The scores of knowledge points of freshmen

IV. Conclusion

In this paper, we use spatio-temporal residual attention network to recognize students' expression changes when learning with knowledge map and improve their foreign language learning. The recognition rate of six kinds of student expressions of spatio-temporal residual attention network reaches 0.942, 0.957, 0.936, 0.938, 0.939, and 0.945, respectively, and the error of expression recognition is small. In the process of improving TextRank algorithm

to extract 4, 6, 8, 10, 15 keywords, the three performance index scores of accuracy, recall, and F-value are more than 90%, and the keyword extraction effect is better. Combining knowledge graph + spatio-temporal residual attention network to assist students' foreign language learning can enhance students' positive expression scores (average scores of concentration and excitement >90) and reduce students' negative expression scores (average scores of confusion, distraction, nervousness and sleepiness <85). In the future, we can further explore the possibility of applying real-time monitoring mechanism in spatio-temporal residual attention network to improve the real-time performance of students' expression recognition.

References

[1] Sun, G., Li, T., & Liang, R. (2022). SurVizor: visualizing and understanding the key content of surveillance videos. Journal of Visualization, 1-17.

[2] Qian, R., Dong, X., Zhang, P., Zang, Y., Ding, S., Lin, D., & Wang, J. (2024). Streaming long video understanding with large language models. Advances in Neural Information Processing Systems, 37, 119336-119360.

[3] Altun, M., & Celenk, M. (2017). Road scene content analysis for driver assistance and autonomous driving. IEEE transactions on intelligent transportation systems, 18(12), 3398-3407.

[4] Carmichael, M., Reid, A., & Karpicke, J. D. (2018). Assessing the impact of educational video on student engagement, critical thinking and learning. A SAGE white paper, 1-21.

[5] Kumar, B. S., & Seetharaman, K. (2022). Content based video retrieval using deep learning feature extraction by modified VGG_16. Journal of Ambient Intelligence and Humanized Computing, 13(9), 4235-4247.

[6] Latif, A., Rasheed, A., Sajid, U., Ahmed, J., Ali, N., Ratyal, N. I., ... & Khalil, T. (2019). Content-Based Image Retrieval and Feature Extraction: A Comprehensive Review. Mathematical problems in engineering, 2019(1), 9658350.

[7] Kamelia, K. (2019). Using video as media of teaching in English language classroom: expressing congratulation and hopes. Utamax: Journal of Ultimate Research and Trends in Education, 1(1), 34-38.

[8] Hairuddin, N. H., Abubakar, M., & Astri, Z. (2022). The utilization of video-based learning in teaching English for non-English major students. Seltics Journal: Scope of English Language Teaching Literature and Linguistics, 27-33.

[9] Masyitoh, N. F., Malihah, N., Risdianto, F., & Guritno, A. (2019, December). Video as educational multimedia to teach english speaking. In Journal of Physics: Conference Series (Vol. 1339, No. 1, p. 012118). IOP Publishing.

[10] Winaldo, M. D., & Oktaviani, L. (2022). Influence of video games on the acquisition of the English language. J. English Lang. Teach. Learn, 3(2), 21-26.

[11] Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., ... & Xu, C. (2025). Video understanding with large language models: A survey. IEEE Transactions on Circuits and Systems for Video Technology.

[12] Hayakawa, Y., Oonuma, T., Kobayashi, H., Takahashi, A., Chiba, S., & Fujiki, N. M. (2017). Feature extraction of video using artificial neural network. International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), 11(2), 25-40.

[13] Salim, F. A., Haider, F., Conlan, O., & Luz, S. (2018). An approach for exploring a video via multimodal feature extraction and user interactions. Journal on Multimodal User Interfaces, 12, 285-296.

[14] Saienko, N., & Shevchenko, M. (2020). Authentic videos in teaching English to engineering students at universities. International Journal of Learning, Teaching and Educational Research, 19(8), 350-370.

[15] Huang, J., Zhou, W., Li, H., & Li, W. (2018). Attention-based 3D-CNNs for large-vocabulary sign language recognition. IEEE Transactions on Circuits and Systems for Video Technology, 29(9), 2822-2832.

[16] Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T. L., Bansal, M., & Liu, J. (2021). Less is more: Clipbert for video-and-language learning via sparse sampling. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7331-7341).