Exploration of Deep Semantic Analysis and Application of Video Images in Visual Communication Design Based on Multimodal Feature Fusion Algorithm

Yanlin Chen, Xiwen Chen*
College of Visual Arts, Hunan Mass Media Vocational and Technical College, Changsha 410000, China

***Abstract*—Fully utilizing image and video semantic processing techniques can play a more effective role in visual communication design. In order to further explore the application of multimodal feature fusion algorithm (MFF) in video image feature analysis in visual communication design, with the aim of enhancing the depth and breadth of design creation. This article focuses on the application of video semantic understanding technology by combining image and video semantic processing techniques, in order to achieve a comprehensive, three-dimensional, and open expansion of design thinking. The MFF algorithm was proposed and implemented, which innovatively integrates multimodal information such as visual and audio in videos, deeply explores action semantics, and shows significant performance improvements compared to traditional algorithms. Specifically, compared to the other two mainstream algorithms, its performance has improved by 24.33% and 14.58%, respectively. This discovery not only validates the superiority of MFF algorithm in the field of video semantic understanding, but also reveals the profound impact of video semantic understanding technology on visual communication design practice, providing new perspectives and tools for the design industry and promoting innovation and development of design thinking. The novelty of this study lies in its interdisciplinary methodology, which applies advanced algorithm techniques to the field of art and design, and the significant improvement of the proposed MFF algorithm in enhancing design efficiency and creativity.**

***Keywords*—Video semantics; understanding; visual communication; design**

I. INTRODUCTION

Under the background of new media, visual communication design has become a comprehensive discipline, which promotes the emergence of new design art forms, and also changes the traditional design concept and thinking mode, which has been widely used in all aspects of life [1]. The video semantic understanding technology is updating day by day, which plays a driving role in the development of visual communication design, and promotes the information design to have different meanings and effects. In varying degrees, it has an impact on visual effects in various fields, becoming a role of visual culture construction, and also a designer of the new era [2]. In the process of visual design, there is an inseparable relationship between the state of graphic design and its aesthetic feeling. From the graphic chronicle in ancient times to the evolution of graphic symbols and character symbols today, the rule of graphic language has realized a good interaction between simple and complex, and has actually become one of the indispensable decorative languages in people's lives [3]. In a design work, almost all images can appear in the form of points, surfaces, lines, and the comprehensive application of points, lines and surfaces, which is flexible [4]. Moreover, the design should follow such a principle that visual graphics must be clearly identifiable, on the contrary, when a graphic structure is fuzzy, the meaning it conveys must be fuzzy [5]. Graphical symbols of visual design, as well as visual representations such as pattern modeling, pictures, signs, etc. aimed at the design theme, are called graphic information through the content, feelings and visual impact conveyed by these graphics [6]. In the visual design language, the modeling elements are points, lines, surfaces, tones and materials. Any kind of graphic information, such as trademarks, patterns, etc., can be understood as a specific modeling element [7]. Therefore, in visual design, a Fig. 1, a pattern or a style of symbols should be used accurately. Only in this way can they convey the necessary information [8].

With the rapid development of digital media technology, visual communication design is no longer limited to traditional static images and text, but is gradually evolving towards dynamic, multidimensional, and interactive directions. Video, as one of the most expressive and infectious forms of media, plays an increasingly important role in visual communication design. However, current video semantic understanding technology cannot fully meet the needs of visual communication design for accurate and efficient information extraction and expression, especially when dealing with complex scenes and multimodal data, there is a significant research gap. This study aims to fill this research gap by exploring the application of video semantic understanding technology in visual communication design and proposing a new method based on multimodal feature fusion algorithm (MFF). This algorithm aims to integrate multi domain scene information, especially action semantic information, in videos to achieve more accurate and comprehensive video content understanding and analysis. Through this method, we hope to enrich the thinking and expression methods of visual communication design, making design thinking more diverse and extensive, while improving the efficiency of information communication and aesthetic experience of design works. In

*Corresponding Author.

the process of using visual design language to communicate, designers mostly pay attention to and study how to convey "explicit" visual information [9]. In fact, these explicit visual information communications are established on the basis of "known information", and these "known information" are often ignored by designers, let alone effectively used. These "preset information" are just the basis and prerequisite for importing the visual information to be conveyed. Only by effectively combining this implicit known information with the explicit visual information to be conveyed can an effective visual communication process be generated [10]. If the traditional way of communication is adopted, the relevant staff need to fuse images, words, etc. to realize the transmission of visual information [11]. Through dynamic visual transmission, the transmission speed can be improved. At the same time, it allows people to obtain accurate information resources in a short time and observe the way of information transmission, while the dynamic way of transmission requires the assistance of video semantic understanding technology. Under the video semantic understanding technology, through dynamic transmission, the abstract visual works are more vividly presented in the eyes of the audience, so that the audience has a more profound memory [12]. Compared with traditional static images, the dynamic design of visual communication has more advantages to avoid visual fatigue of the audience, increase the sense of interest for the audience, and make the information communication effect good.

Semantic understanding of video is the highest level of research in machine vision. Firstly, video is processed and analyzed, and then the main content of video is described. Different video meanings and tasks require different working situations, and different objects appear. Of course, different visual comprehension algorithms are needed. In a certain period of time, multimedia data such as video frames, audio signals, and transcribed texts may not appear at the same time, and there is an unsynchronized phenomenon, but they all share a semantic meaning and are coupled with each other within the semantic duration [13]. For example, different shots that express the same meaning may look very different visually. Swimming and football, which also represent "sports", are mainly composed of blue swimming pool water and green football field grass. However, text features may express more similarities, thus making up for the weak correlation of other modes [14].

The introduction of this article first emphasizes that in the context of new media, visual communication design has become a comprehensive discipline, which not only promotes the emergence of new design art forms, but also changes traditional design concepts and ways of thinking, and is widely applied in various aspects of life. This clarifies the important position and influence of visual communication design in contemporary society. Furthermore, the introduction points out that the increasingly updated video semantic understanding technology has played an important role in promoting the development of visual communication design, and has facilitated information design to have different meanings and effects. This further emphasizes the positive impact of technological progress on the field of design. The contribution points of research innovation are as follows:

- This study proposes a multimodal feature fusion algorithm (MFF), which demonstrates unique advantages in video semantic understanding. By integrating semantic information of multi domain scene actions in videos, MFF algorithm can comprehensively understand video content, which is an important innovation for traditional video processing algorithms.
- The research has improved the application effect of video semantic understanding technology in visual communication design. The application of MFF algorithm in the field of visual communication design has achieved an organic combination of video semantic understanding technology and design art.
- Compared to the other two mainstream algorithms, the MFF algorithm has improved performance by 24.33% and 14.58%, respectively. This significant improvement not only validates the effectiveness of the MFF algorithm.
- Research the application of video semantic understanding technology in visual communication design from the perspective of semantic analysis. The update of this design concept helps designers better grasp design trends and create more innovative and artistic works.

II. Related Work

Ren believes that visual thinking is based on the information of the objective image itself, and uses visual thinking to capture the characteristics of the shape as a means to connect the abstract shape with the semantics of the concrete objective image [15]. Min, Wang, Rushan emphasized that in visual communication, the meaning of shape is generated, communicated, fed back in the "person shape person" system, and then generated, communicated, and fed back again, and so on [16]. Therefore, in the video semantics, Chang and others believe that visual communication design is not only a simple combination of words, sounds and pictures, but also a combination effect of three or more communication elements. The diversified design helps both the transmission and reception of video news achieve ideal effects [17]. Xue has shown in his research that graphics are one of the important elements in visual communication design, and each designer has different performance for the use of graphic design in visual communication. Because people live in different geographical environments, their cultural cognition is also different. However, through the interpretation of graphics and visual transformation, barriers and barriers in communication can be effectively solved, so that visitors can more accurately understand the design connotation [18]. Ge H regards multi-level semantic concepts as the hidden state of Markov chain, and regards multiple basic visual semantics with time semantic context constraints as the observable symbols corresponding to the hidden semantic state. Under this condition, using hierarchical hidden Markov model, the analysis and extraction process of multi granularity visual semantics can be transformed into the process of analyzing the most likely hidden state of known observable symbols [19]. In

the research of video semantic understanding, features of different modes cannot be directly analyzed and compared in the semantic analysis process. How to mine the complementary association between different modes and extract and construct data information consistent with the semantic content has become a problem to be solved. Bae J proposes an extended direct push support tensor machine, which is used as a semantic classifier to detect the semantic concept of video shots [20]. Xin thinks that video data contains a large amount of valuable information. However, due to semantic barriers, human beings have been unable to make full use of this information. However, the birth of data mining technology can help people make full use of this information [21]. Xiao et al. proposed a feature extraction method (Dense Trajectory, DT) based on dense trajectories. This method samples each frame of video intensively, then tracks the optical flow information of these sampling points, and obtains more accurate feature expression through screening and optimization [22]. Subsequently, Natalia et al. proposed an improved Dense Trajectory (iDT) algorithm, which mainly optimizes optical flow images, and improves the regularization and feature encoding methods at the same time, further improving the algorithm performance [23]. Wu et al. proposed a video humanoid parsing algorithm based on semantic transfer. This method uses the complementarity of image semantics and video semantics to propose a module that can integrate two semantics at the same time, and applies it in the convolutional neural network structure [24].

III. Theories Related to Video Semantic Understanding Technology

A. Video Structure

Faced with a variety of complex information, traditional information management methods have been stretched to the limit, and human beings must develop new technologies to help themselves deal with information [25,26]. Traditional video management is basically realized by the underlying features. However, many years of practice and practical experience tell us that video management must analyze the semantics of video. Video language, as a camera technology applied in video, can give people a strong impact on vision. Video image language not only has a passionate expression, but also has a calm way of thinking, bringing more visual feelings to people. Using video language in visual communication design can make advertisements and posters more personalized and aesthetic [27,28]. By combining video language with visual communication design, designers can make better use of film and television effects when designing advertisements, bringing more visual impact to people. A video is composed of continuous image changes. When the image changes more than 24 frames per second, the human eye can no longer distinguish whether the received picture information is still a static image. It looks like a series of fluent pictures. This is called video (Fig. 1). In general, video is composed of continuous image sequences, and the semantic association between each image constitutes the semantic information of the whole video. Moreover, the video data has a much larger capacity than the text data (Table I).

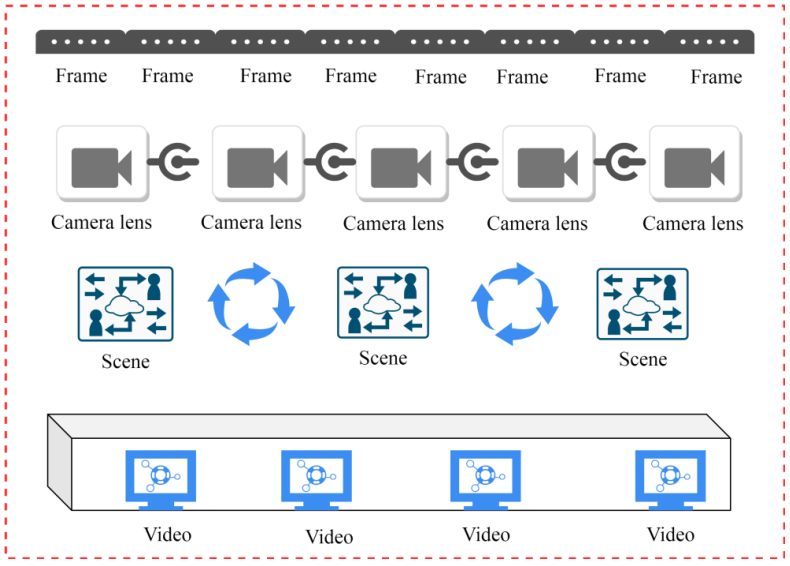

Fig. 1. The composition of each level of the video structure.

Video is mainly divided into structure and composition of each level. First, the adjacent video frames in the time dimension are divided into multiple different shot units without intersection. Key frames of shot units are extracted. The number of key frames extracted for each shot depends on the situation of the shot itself. Observe the video key frames and determine the combination of shots to get the video scene.

TABLE I. Comparison of Text Data and Video Data

Features	Text data	Video data
Symbol set	Limited	Infinite
Resolving power	Low	High
Explanation ambiguity	Low	High
Interpretative function	Low	High
Data capacity	Small	Big

Semantic differences make it impossible for humans to directly apply the information of video data. There is a gap between the low-level features of images that can only be obtained by computers and the high-level semantics. When watching video, human can judge the events in the video by integrating their own knowledge. The computer has limited ability to obtain image features such as image color and texture, but has no ability to judge. Generally speaking, video semantics can be divided into three levels: low-level semantics, middle level semantics and high-level semantics. The semantics of the bottom layer includes the semantics of all the features of the video bottom layer, including the description of color, texture, edges, dynamics and other features. The middle level includes the description of video target action, density, traffic statistics, etc. The high-level semantics is the description of video event attributes. According to this division, we get the video semantic architecture of Fig. 2. In the multimodal analysis method of video semantic understanding, fusion is an indispensable step. Early integration and late integration are the two main integration methods at present. Most existing semantic concept detection methods are based on these two schemes. Pre-fusion refers to combining the features of single modality before learning and training, while post-fusion refers to learning the features of single modality separately, and then fusing and analyzing the results of multiple modalities.

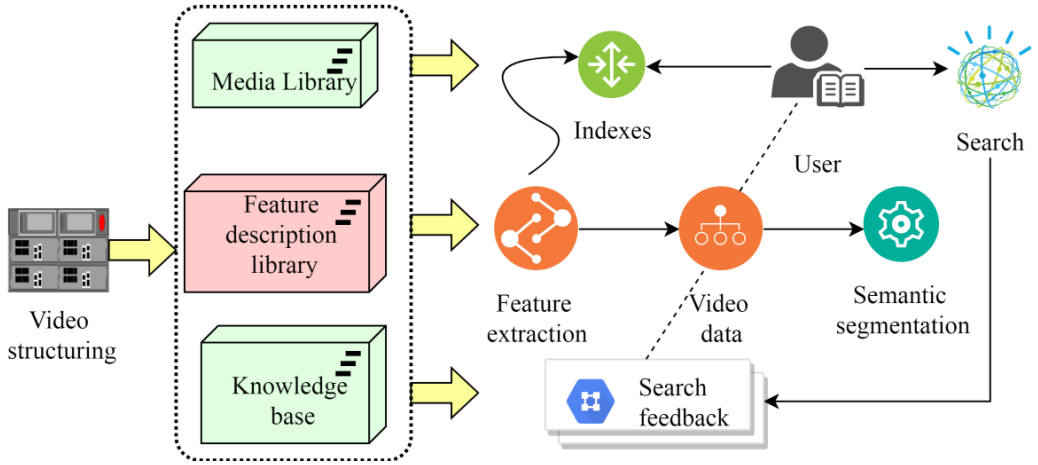

Fig. 2. Multi modal fusion steps for video semantic understanding.

In general, it is natural for an event model to have two or more state variables from a semantic perspective, forming a state chain over time. Factorization of state space into polymorphic chains is another way to simplify the event model. These multiple chains can correspond to sub-events that interact with one compound event or multiple objects at the same time.

B. Application in Visual Communication Design

In the process of visual communication design activities, designers can greatly enhance the user's emotional experience, realize the interaction of multiple senses, and let the audience experience in an all-round way through the application of video semantic understanding technology. Under the background of new media, designers can avoid aesthetic fatigue by reasonably designing graphics, words and videos, and communicate them in the eyes of people by animation, which has a good sense of experience. The central meaning of visual design is "the desire of artistic Fig. materialization". The whole design should be expressed around people's wishes, and it must have human visual visibility and readability. To make the semantics of visual form accurate, we must first understand the potential language of each design factor, understand the design principles, text semantics, graphic semantics, color semantics and psychological implications, and so on, so as to comprehensively convey information. Visual communication cannot be conveyed through people's mouth like people use language, but depends on the media: "shape" is conveyed in the system of people, shape and people. Therefore, it is necessary for people to strictly control and study the rules of "the meaning of form". The semantic meaning of the form with visual communication as the main content (such as signs, advertisements, posters, publicity cards, etc.) mainly conveys the idea of the producer. The semantic meaning of the shape with the use function as the main content (such as cars, rain gear, etc.) makes the user clear and easy to use through the information displayed by the shape, color, etc.

Visual communication design is a complete process of information design and dissemination. Under the influence of digital media technology, great changes have taken place in the mode of information dissemination and reception in people's lives. In the process of the integration of visual communication technology with computer and digital media, the communication mode has changed from dynamic to static, from passive to active, which leads to certain communication characteristics of visual communication design.

IV. Introduction to Algorithms

A. Semantic Concept Modeling

We obtain the contribution score of each action sample to its action semantic concept by preprocessing the classifier, and according to the score, integrate the sample subsets into a fusion vector encoding multi-domain invariant action information. This process can be modeled as:

$$G(x,y)=\sum_{i=1}^{t} r_{ij}(x_i y_j) \tag{1}$$

Where, x_i is the embedded matrix corresponding to the i field, and r_{ij} is the weight of action sample $x_i y_j$.

In order to further explore the potential association between multiple action semantic concepts, we use constraints to limit the number of non-zero rows in matrix W to control the number of sparse features in the model. Constraints are modeled as:

$$cons(R)=\sum_{i=0} I(\|w_i\|\rangle 0)\le u \tag{2}$$

$$P_r(B_r|x)=\frac{p_r(x|w_1)}{\sum_{j=1}^{r} p_j(w_1)} \tag{3}$$

When the Mercer condition is satisfied, replacing the dot product in the optimal classification surface with the inner product S is equivalent to transforming the original input space into a new feature space, and the corresponding high-level multimodal fusion function is:

$$f(x)=\mathrm{sgn}\left\lfloor\sum_{i=1}^{n} a_i k(s_i,s)+b\right\rfloor \tag{4}$$

Where, $a_i\rangle 0$ is the lagrange coefficient and b is the domain value of classification.

Because of the potential correlation between semantics, multi semantic modeling is essentially a process of establishing mathematical models for interrelated tasks. Traditional machine learning methods often adopt single task learning mode, and learn a model independently for each task. This single task learning method ignores the association between multiple tasks (semantic concepts), and loses hidden information existing between data or model parameters.

B. Recurrent Neural Networks

Recurrent Neural Networks (RNN) are often used to process sequence data, such as time dependent voice data, video data, and natural language data with semantic coherence. Like convolutional neural network, cyclic neural network also uses the idea of weight sharing, so similar to convolutional neural network, it can process images of different sizes, and cyclic neural network can be used to process sequence data of any length. The basic principle of recurrent neural network can be expressed as:

$$h_t = f_w(h_t-1)\cdot x_t \tag{5}$$

Where, h_t is the hidden state information representation vector of time t, encoding the sequence information input at the first t times.

Fig. 3 shows a multi-layer recurrent neural network. Single layer refers to the number of layers of the recursive function itself, not the number of moments of the expanded graph. Multilayer recurrent neural network, that is, recursive function, has the form of multilayer network. Multilayer networks have more nonlinear structures and are often used to model sequences with complex structures.

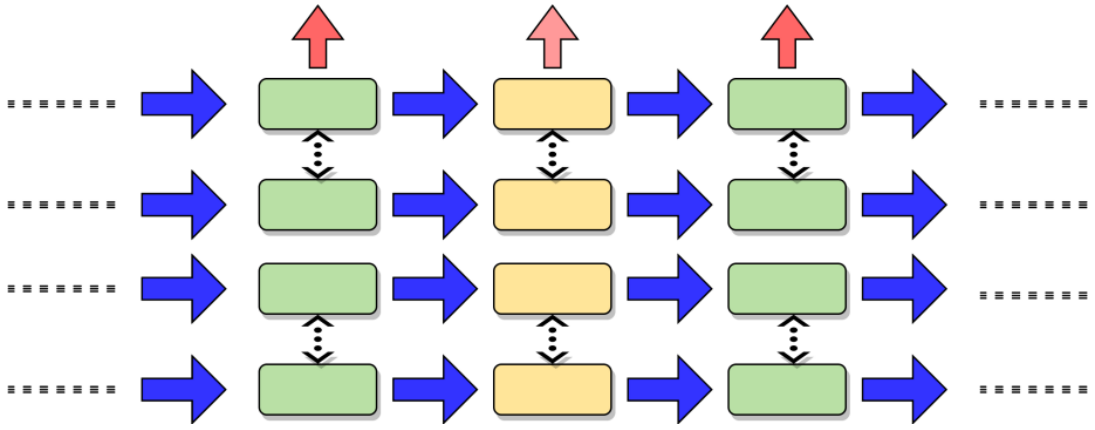

Fig. 3. Multilayer recurrent neural network.

C. Attention Model

When people observe an image, they focus not on all visual content, but on each pixel area of the image selectively. Firstly, RNN feature $V=\{v_1,v_2,\cdots,v_n\}$ is extracted from the video sequence, and then the feature sequence is weighted using the attention model according to the following formula:

$$a^t=\sum_{i=1}^{n} a_i v_i^t \tag{6}$$

By adjusting the attention weight a_i^t at any time, the model strengthens the correlation between the output words and the video content, thereby improving the quality of the output sentences.

Channel attention first performs global average pooling on features to generate vector Z and its t-th element:

$$Z_k=\frac{1}{h\cdot w}\sum_{i=1}^{h}\sum_{i=1}^{w}\theta_t(i,j) \tag{7}$$

D. Feature Extraction

Let the cumulative color histograms of two frames be $H_i(p)$ and $H_j(p)$ respectively, and p is a color statistic in the histogram. If $s_{ij}(p)$ is a local similarity measure around p, their global similarity is:

$$s_{ij} = \sum s_i(p) q_j \tag{8}$$

It can be calculated according to normalized cross-correlation:

$$a_i(p) = \sum H_i(p-q) W_t(q) \tag{9}$$

$$b_j(p) = \sum \left[H_i(p-q) - a_i(p-q) \right]^2 \tag{10}$$

where W_t is a window function of length t.

As long as the cross-correlation is added up, a global similarity can be obtained, and a global similarity can be obtained. Generally, the very dissimilar frames are removed and then judged. For a shot, the first frame is compared with the last frame of the first five or six shots. If a match is found, all shots between them can be considered as belonging to the same scene.

Due to the different observation of video events and the uncertainty of translation, a probabilistic event model is proposed. Bayesian model is a probabilistic event model, which uses the main semantic knowledge and is very powerful in factorizing the state space into variables. See formula (11) for the hierarchical characteristics of the simulated video events of this natural model.

$$\Delta k = \min(\alpha\Delta_2, \alpha\Delta_2, \cdots, \alpha\Delta_n) \prod_{t-1}^{n} \Delta\alpha^{\frac{1}{t}} \tag{11}$$

This is because the probability output at the top node of the sub event network can be easily integrated into an event model "observation" node at a higher level. Although these networks are large, effective inference can be realized according to the existing structure. Hierarchical model semantics adopt hierarchical Yebes network layer. Each layer corresponds to a higher-level semantic unit.

$$p(y|x) = \frac{yx^T w}{1 + yx^T w} \tag{12}$$

The next position of eye movement shall be based on the maximum distribution of each image slice:

$$H = 0.5 \sum_{i=1}^{n} \log(1 + \lambda_i / \mu) + n \log(2\pi) \tag{13}$$

Among them, λ_i is the eigenbit of the pixel covariance matrix in the middle part of each image slice, and μ is the noise level that is different from the pixel grayscale quantization value.

In order to detect the changing image position, the contrast is used to calculate, that is, the normalized result of the standard deviation of the local pixel gray level inside the image to the average gray level of the entire image, which helps to select the area with relatively high contrast as the eye movement pre-attention center.

$$c = G_N \sum_k (I_{ij} - I_k)^2 \tag{14}$$

When the problem is linear and indivisible, a nonlinear mapping algorithm is needed for transformation. The properties of high-dimensional feature space change when the low dimensional input linearly indivisible examples are converted to high-dimensional feature space, so that people can use it. The principle must be from linearly separable to linearly indivisible in more complex cases, and even nonlinear functions are applicable. According to the statistical theory of limited samples, the inequality is established by probability:

$$R(Q) \le R_{emp}(Q^*) + \sqrt{\frac{c}{l}\left(\frac{p^2}{m^2}\log p - \log m\right)} \tag{15}$$

In the formula: p is the spherical radius of the entire sample space, M is the space edge, and l is the number of samples in the sample space.

V. Experiments and Results

In order to evaluate the effectiveness of the proposed method, a large number of experiments have been carried out on the NYUDv dataset and the SUN-RGBD dataset. First, we introduce the experimental setup, including experimental details, data sets and evaluation indicators, and then conduct ablation experiments to determine the segmentation performance of the newly proposed network. Finally, the proposed method is compared with the existing method in the above two datasets. General data enhancement strategies are adopted, such as random scaling, random horizontal flipping and random clipping of images within [0.1, 0.5] scales. For the above data set, adjust the size of the input image to 480 × 640 pixels (Table II). In the test, the prediction results for semantic segmentation are obtained only from the main branch of the decoder.

TABLE II. Comparison Results in Each Category with Other RGBD Image Semantic Segmentation Methods On The NYUDv Dataset

Method	Image A	Image B	Image C	Image D	Image E
Deep Lab	4	12.2	8.3	6	5.9
Literature [18]	6.2	3.6	6.4	13.3	7.1
CANet	4.7	9.3	1	9.3	13.4
Literature [22]	1.4	5.8	14.5	7.8	9
This paper	7.1	8.1	5.1	1.7	6.7

Here we combine the above two strategies. For the unlabeled sample video, we first calculate the score through temporal dependency and semantic relevance strategies, and then linearly fuse them to get the final score. Fig. 4 shows that the application of sample 1 in the two strategies further

improves the classification performance than that of sample 2.

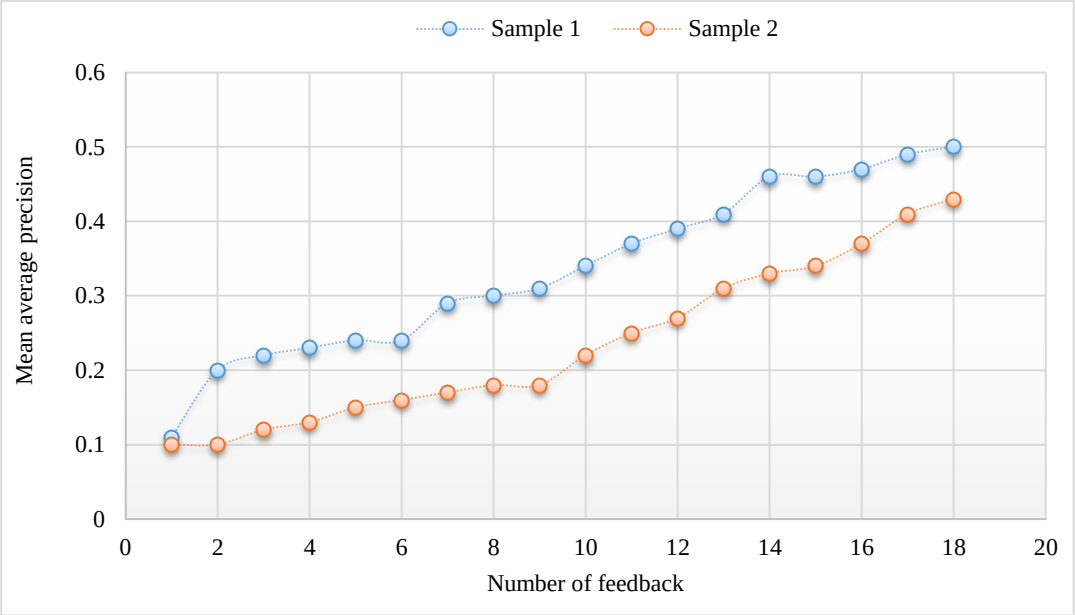

Fig. 4. The results of 18 feedbacks of the two sample combinations.

In the process of video data information mining, the low-level features at the bottom of the above video image, including texture, color, motion vector, detection edge, histogram and other information, are first extracted. Such information is beyond the scope of people's understanding ability, which provides a basic work for the mining of meta semantic information. People then use object detection [31], tracking and extracting semantic information from image feature comparison. Noise in natural images can affect the accuracy of super pixel similarity to some extent, and further hinder the propagation of associated attributes. In general, the interference of noise on natural images is reflected in their statistical characteristics, and mLab can be regarded as a special statistical feature of images. Compare the characteristics of the three segments to obtain Fig. 5 and 6.

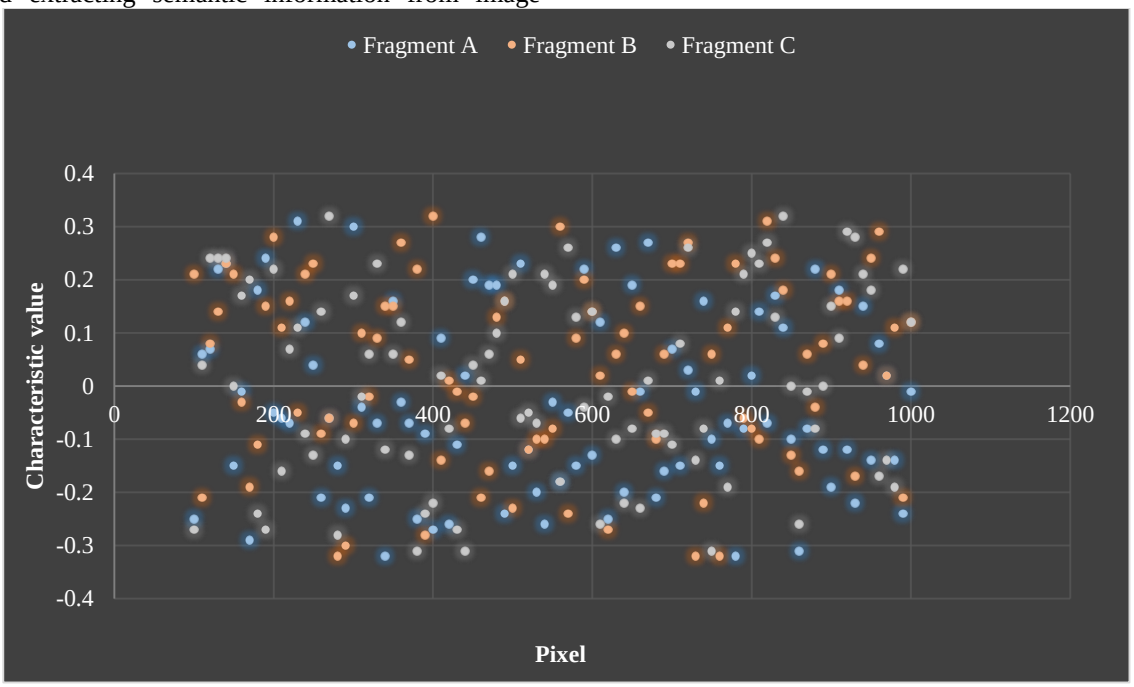

Fig. 5. Original image.

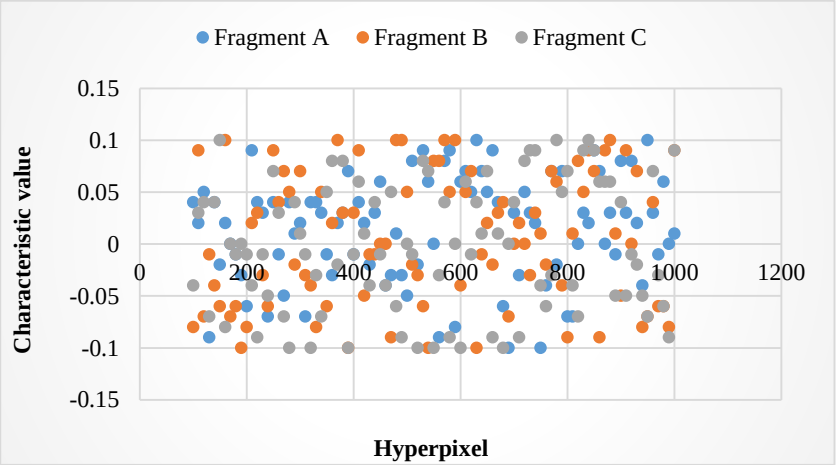

Fig. 6. Smoothed image.

From the above results, it can be seen that the subspace of multi-scale super pixels is always sparse, and the changes presented are also significantly smaller than the changes in area when the number of super pixels decreases. Therefore, the subspace maintenance can better reveal the membership relationship of super pixels. The smoothed image obtains some different image regions, thus creating sub regions of the image. On the basis of these sub regions, the basic semantic image slice of the image can be roughly obtained by determining the characteristics between the sub regions. And the visual features of the image are obtained on the basis of color and texture acquisition, which is helpful for the transition of image understanding from the underlying visual cognition to the connection between image semantics and the analysis of image structure composition.

TABLE III. COMPARISON OF TWO MODAL PERFORMANCES

RGB mode		Depth mode	
Side view	Foresight	Side view	Foresight
0.6	0.75	0.61	0.59
0.71	0.85	0.75	0.78
0.79	0.84	0.62	0.63
0.74	0.73	0.53	0.74
0.73	0.69	0.62	0.85
0.66	0.76	0.7	0.74

It can be seen from the comparison of the performance of the two modes in Table III that the cross-modal feature propagation improves the performance of semantic segmentation. MFF algorithm handles the details well, and accurately captures the differences between classes and the consistency within classes. In order to segment the super pixel image at different scales, it is necessary to construct a hybrid graph to describe the relationship between pixels and super pixels and between super pixels and super pixels. First, a multimodal fusion module based on attention mechanism is proposed, which aims to process multi-level pairing complementary information in the dual stream encoder, that is, to process color and depth images respectively in the same backbone network. In order to clearly test the effect of each link, we respectively evaluate in RGB and depth scenes of NYUDv database.

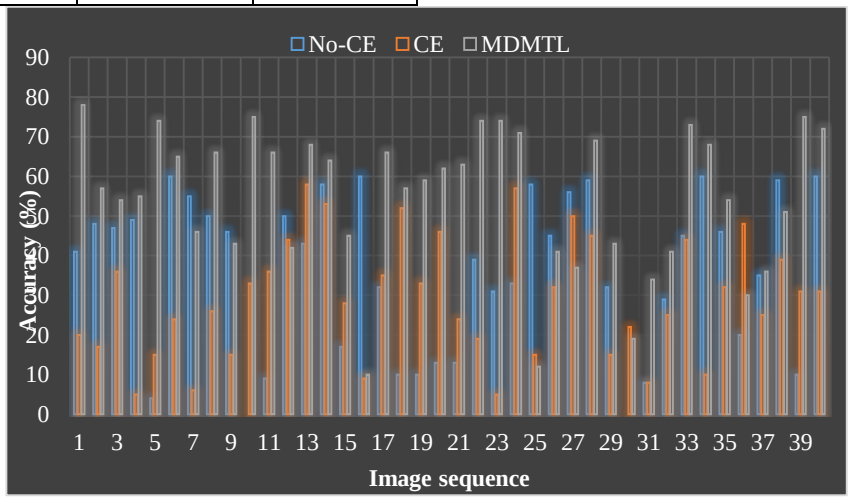

Fig. 7. Evaluation of RGB modal scene feature extraction methods.

It can be seen from the data in Fig. 7 that compared with No-CE sub-experiment, CE sub-experiment can achieve better classification performance, which is nearly 28.2% and 17.6% higher than No-CE in RGB and depth scenes, respectively, which proves the effectiveness of multi-domain embedded matrix.

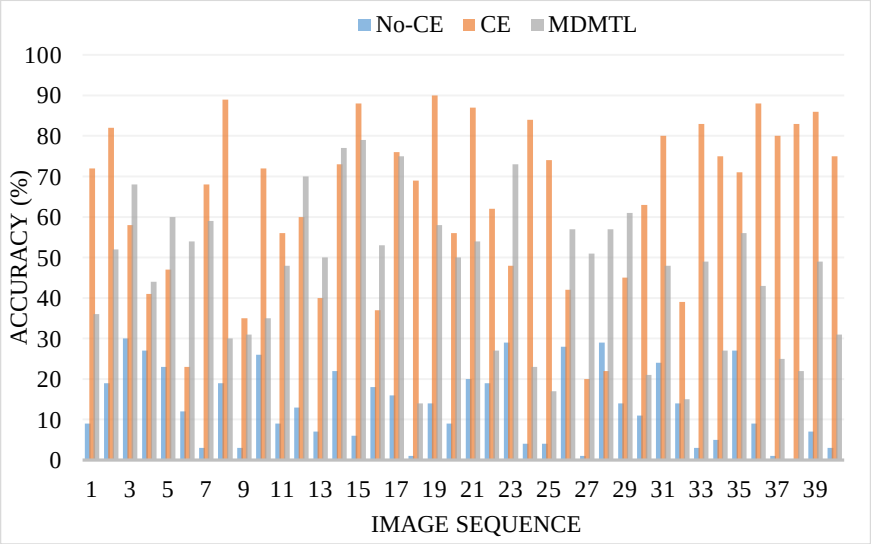

Fig. 8. Evaluation of deep modal scene feature extraction methods.

It can be seen from the data in Fig. 8 that by comparing the performance of CE sub experiment and multimodal feature fusion algorithm, the classification performance of MDMTL in depth scenes is 15.4% higher than that of CE, which proves the effectiveness of multi domain feature fusion.

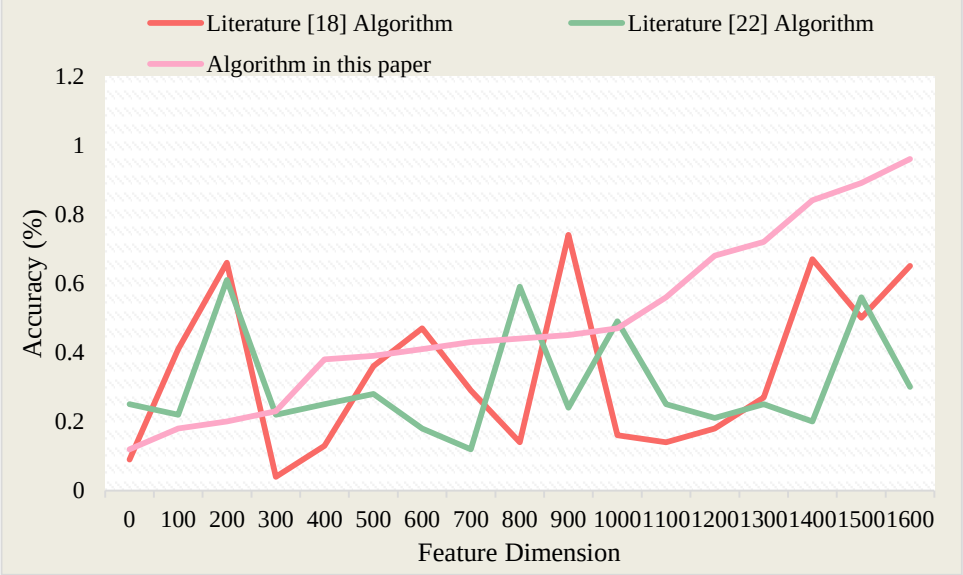

Fig. 9. Comparison of semantic information accuracy of three algorithms.

Fig. 9 shows that because the algorithm in this paper integrates the action semantic information in the multi domain scene, it has a certain anti-interference ability on the impact of feature noise. Therefore, its performance is nearly 24.33% higher than that of the Literature [18] algorithm, and nearly 14.58% higher than that of the Literature [22] algorithm. So, we conclude that with the help of video semantic understanding technology, visual designers can better grasp consumers' psychology with the support of technology. So as to better grasp the initiative in the whole design process and guide consumers to obtain information. This article emphasizes that the MFF algorithm achieves deeper semantic understanding by integrating action semantic information from multiple domain scenes. Compared with those researches that only rely on single modal information [29] (such as only vision or only audio), MFF algorithm can capture key information in video more comprehensively.

The comparison results in Fig. 9 show that the MFF algorithm is significantly better than the other two algorithms in terms of semantic information accuracy. This is mainly due to the MFF algorithm integrating action semantic information from multiple domain scenes and possessing certain

anti-interference capabilities. In practical applications, video data often contains various noise and interference factors, such as lighting changes, occlusion, motion blur, etc. The MFF algorithm effectively reduces the impact of these noise factors on segmentation results by introducing multimodal feature fusion and attention mechanism [30], thereby improving the accuracy and reliability of semantic information. With the help of video semantic understanding technology, visual designers can gain deeper insights into consumer psychology and more accurate scene understanding. This helps designers to better grasp the initiative throughout the entire design process, enhance consumer participation and satisfaction through precise information communication and guidance. Meanwhile, video semantic understanding technology also provides designers with rich data support and decision-making basis, making the design process more scientific and intelligent. Therefore, applying video semantic understanding technology to visual communication design not only helps improve the quality and effectiveness of design works, but also promotes innovation and development in the design industry.

VI. Conclusions

This study thoroughly explores the application of multimodal feature fusion algorithm (MFF) in video semantic understanding, successfully demonstrating its significant advantages in improving the efficiency and effectiveness of visual communication design. The MFF algorithm integrates action semantic information from multiple domain scenes to achieve comprehensive and accurate analysis of video image features, providing a richer and deeper semantic understanding foundation for visual communication design. The experimental results show that compared with existing algorithms, the MFF algorithm achieves performance improvements of 24.33% and 14.58%, respectively, which fully demonstrates its superiority in the field of video semantic understanding. This study not only enriches the theoretical system of video semantic understanding technology, but also brings new technical support and creative inspiration to the field of visual communication design. Through the application of MFF algorithm, designers can have a more comprehensive and in-depth understanding of video content, thereby better grasping consumers' psychology and needs in the design process, and achieving a comprehensive, three-dimensional, and open understanding and perception of design thinking. This technological innovation not only enhances the quality and attractiveness of design works, but also points out the direction for the future development of the visual communication design industry.

Although significant progress has been made in the combination of video semantic understanding technology and visual communication design in this study, there are still some limitations. Firstly, this study mainly evaluates based on specific databases such as NYUDv, which may not fully cover all possible video scenes and design requirements. Therefore, the generalization ability of the algorithm needs further verification. Secondly, although multimodal feature fusion algorithms integrate action semantic information from multiple domain scenes, their performance may be limited when dealing with extremely complex or highly dynamic video content. In addition, this study did not delve into the specific needs and application scenarios of video semantic understanding technology in different design fields (such as advertising, animation, film, etc.), which limits the wide applicability of the research results.

References

[1] Xu, J., Huang, F., Zhang, X., Wang, S., Li, C., Li, Z., & He, Y. (2019). Sentiment analysis of social images via hierarchical deep fusion of content and links. Applied Soft Computing, 80, 387-399.

[2] Zhu, W., Wang, X., & Li, H. (2019). Multi-modal deep analysis for multimedia. IEEE Transactions on Circuits and Systems for Video Technology, 30(10), 3740-3764..

[3] Add C. A Classification Framework to Support the Design of Visual Languages - ScienceDirect. Journal of Visual Languages & Computing, 2020, 13(6):573-600.

[4] Liu Q. Visual Elements Mining in the Packaging Design of Children's Products based on OpenGL and SVM. Electronics and Sustainable Communication Systems, 2018, 17(1):118-140.

[5] Abdu, S. A., Yousef, A. H., & Salem, A. (2021). Multimodal video sentiment analysis using deep learning approaches, a survey. Information Fusion, 76, 204-226.

[6] Smith J R, Srinivasan S, Amir A. Integrating Features, Models, and Semantics for TREC Video Retrieval. National Institute of Standards and Technology, 2019, 7(5):409.

[7] Xu C, Cheng J, Zhang Y. Sports Video Analysis: Semantics Extraction, Editorial Content Creation and Adaptation. Journal of Multimedia, 2019, 4(2):69-79.

[8] Wei Y, Bhandarkar S M, Li K. Semantics-Based Video Indexing using a Stochastic Modeling Approach. Image Processing, 2019, 38(4):34.

[9] Kumar, A., Srinivasan, K., Cheng, W. H., & Zomaya, A. Y. (2020). Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. Information Processing & Management, 57(1), 102141.

[10] Guo, J., Song, B., Zhang, P., Ma, M., & Luo, W. (2019). Affective video content analysis based on multimodal data fusion in heterogeneous networks. Information Fusion, 51, 224-232.

[11] Cai Y, Milcent G, Marian L. Visual Digest Networks. Springer-Verlag, 2019, 23(2):89.

[12] Kompatsiaris Y, Hobson P. Introduction to Semantic Multimedia. Springer London, 2018, 2(14):135.

[13] Leggett M G. Mnemovie: visual mnemonics for creative interactive video. Plos One, 2018, 4(2): e4311.

[14] Park M C, Son J Y. Design of 3D Mobile Phones and Application for Visual Communication. Springer-Verlag, 2021, 1744(4):042130 (4pp).

[15] Ren R. Research on the Specific Application of Ethnic Elements in Dynamic Visual Communication Design Based on Intangible Cultural Heritage. Clausius Scientific Press, 2018, 30(4):9.

[16] Min, Wang, Rushan. Innovative Application of the Ceramic Pattern in the Modern Visual Communication Design. Applied Social Science, 2017, 8(11):381.

[17] Chang S K, Costagliola G, Orefice S. A methodology for iconic language design with application to augmentative communication. Workshop on Visual Languages, 2019, 89(132):1479.

[18] Xue H. Research on the Application of Visual Information Communication in Web Design in the Digital Age. Information Science and Education, 2020, 29(5):9.

[19] Ge H,Yu H. The application and design of neural computation in visual perception. Journal of Visual Communication and Image Representation, 2020, 1533(4):042035 (5pp).

[20] Bae J,Watson B. Toward a Better Understanding and Application of the Principles of Visual Communication. Springer New York, 2020, 1648(4):042029 (5pp).

[21] Xin C. A Research on the Creativity Design Method of Visual Communication Design. Design concept, 2017, 55(16):463-469.

[22] Xiao Y. The application of visual communication design in display design. Science herald, 2015, 15(021):321-321.

[23] Natalia D, Fathoni A. "Tigerheart" short animation visual communication design. Earth and Environmental Science, 2021, 729(1):012051 (7pp).

[24] Wu Y Y. A Creative Research of Decorative Features of Art Nouveau in Visual Design on Record Application. About visual communication, 2018, 30(7):718-724.

[25] Dimitri, G. M. (2022). A short survey on deep learning for multimodal integration: Applications, future perspectives and challenges. Computers, 11(11), 163.

[26] Kwek, C. L., Yeow, K. S., Zhang, L., Keoy, K. H., & Japos, G. (2022). The Determinants of Fake News Adaptation during COVID-19 Pandemic: A Social Psychology Approach. Recoletos Multidisciplinary Research Journal, 10(2), 19–39. https://doi.org/10.32871/rmrj2210.02.05

[27] Al-Azani, S., & El-Alfy, E. S. M. (2020). Enhanced video analytics for sentiment analysis based on fusing textual, auditory and visual information. IEEE Access, 8, 136843-136857.

[28] |Kay Hooi Keoy, Yung Jing Koh, Javid Iqbal, Shaik Shabana Anjum, Sook Fern Yeo, Aswani Kumar Cherukuri, Wai Yee Teoh and Dayang Aidah Awang Piu (2023), Streamlining Micro-Credentials Implementation in Higher Education Institutions: Considerations for Effective Implementation and Policy Development, Streamlining Micro-Credentials Implementation in Higher Education Institutions: Considerations for Effective Implementation and Policy Development. https://doi.org/10.1142/S02196492235

[29] Wang, Q., Tong, G., & Zhou, S. (2023). A study of dance movement capture and posture recognition method based on vision sensors. HighTech and Innovation Journal, 4(2), 283-293.

[30] Dibs, H., Ali, A. H., Al-Ansari, N., & Abed, S. A. (2023). Fusion Landsat-8 thermal TIRS and OLI datasets for superior monitoring and change detection using remote sensing. Emerging Science Journal, 7(2), 428-444.

[31] Kurdthongmee, W., Suwannarat, K., & Wattanapanich, C. (2023). A framework to estimate the key point within an object based on a deep learning object detection. HighTech and Innovation Journal, 4(1), 106-121.