Proceedings of CONF-SEML 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA32096

The Application and Prospect of Multi-modal Information Fusion in Human Action Recognition

Jingyi Hong

School of Software, East China University of Technology, Fuzhou, China
2022212898@ecut.edu.cn

Abstract. The recognition of unimodal behavioral postures is easily influenced by the environment, such as lighting, occlusion, etc. Therefore, most of the existing action recognition focuses on the multimodal field. Therefore, this article selects two typical methods at present for comparison, analysis and summary. Specifically, two methods, namely the dual-stream cross-modal fusion transformer and the model-based multimodal neural network, were selected for research. Among them, the core contribution of the dual-stream cross-modal fusion transformer is to achieve the fusion of RGB and depth information through modal enhancement and modal interaction. Model-based multimodal neural networks integrate OpenPose skeleton data with RGB frames and employ graph convolution and spatio-temporal ROI fusion. These two methods present two completely different approaches to action recognition in multimodal situations. In addition, this article also presents new research directions for the future， Such as lightweighting, etc. And a specific application analysis was provided.

Keywords: Action recognition, Deep learning, Multimodal fusion

1. Introduction

An important technology in the process of human-computer interaction is behavior recognition, However, machines' understanding of human behavior is complex and diverse. In addition, the application fields of behavior recognition include video surveillance, etc. And the demand for effective monitoring of public places such as airports, railway stations, shopping centers, crowded sports fields, and military facilities is increasing day by day, or it can be used in intelligent medical facilities, such as daily activity monitoring and fall detection in elderly homes. Behavior recognition in these fields is of great significance [1-3].

Traditional behavior recognition focuses on a single piece of information. It is easily affected by environmental factors such as light. For this reason, this paper reviews and summarizes behavior recognition in the context of multimodal information. The first method among them is a dual-stream cross-modal fusion transformer. The second method is model-based multimodal neural networks [4,5].

The chapters of this article are as follows. First, two action recognition technologies were introduced in detail and compared. Then analyze the application potential and existing deficiencies of multimodal fusion technology in action recognition. Then several typical application scenarios

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.BA32096

were introduced. Finally, the conclusion and the future research methods were given. In conclusion, this paper provides a scientific basis and theoretical support for action recognition under multimodal information fusion.

2. Action recognition technology

2.1. A dual-stream cross-modal fusion transformer for red green blue depth (RGB-D) action recognition

Because traditional convolutional neural networks, when extracting features. Good results have been achieved in single-modal feature extraction. However, feature extraction in the multimodal fusion mode remains a huge challenge. Whether it is the good capture of information or the better integration and complementarity of the characteristics of various modalities of information. In order to better extract and integrate the action recognition features of different patterns. The team proposed a dual-stream cross-modal fusion transformer (DSCMT) for the detection of action recognition. It mainly consists of a dual-stream feature extractor (ConvNet) and a cross-modal fusion transformer. Among them, the dual-stream feature extractor simultaneously extracts features of different modalities by two independent basic convolutional networks.

The basic layer of the cross-modal fusion transformer is composed of the modal enhancement module (MEM) and the mining information modeling (MIM).

2.1.1. Supply chain management (dell server configuration and management, DSCM) transformer

Dual-stream feature extractor: Unlike traditional convolutional networks, the dual-stream feature extractor extracts a pair of dynamic images from the RGB-D sequence. Virtual desktop infrastructure (VDI), the directional deviation index (DDI) is input into ConvNet. One for extracting spatiotemporal features from Red, Green, Blue, RGB images. Another flow is used for feature learning in deep dynamic images to obtain feature maps. The calculation as follows:

$$(fr, fd) \in RC \times H \times W \tag{1}$$

Cross-modal fusion transformer: The output of the upper-layer dual-stream feature extractor is used as the input of this transformer. Use 1×1 convolution to reduce $f_r \in R^{C\times H\times W}$ and $f_d \in R^{C\times H\times W}$ to a lower dimension. Obtain the arranged feature mappings $f_{ro} \in R^{H\times W\times D}$ and $f_{r1} \in R^{H\times W\times D}$, MEM can adaptively enhance the semantic information in each stream. MIM can take the information in two streams as input to complement the information gap between multiple modalities.

2.1.2. Bottleneck excitation feedforward block

Introduce a fully connected feedforward neural network (FNN) in the encoder and decoder to enhance the expressive power. Although the performance has been improved, it also led to an increase in the number of parameters. Therefore the research team drew on the ResNet model and introduced a bottleneck excitation feedforward block, this block receives a feature A $\in R^{N\times HW\times d}$ as input. Here, N represents the batch size, HW represents the sequence length，d is the embedding dimension. First, compress the sequence information through the global average pooling operation， generate the global context embedding z for the input features. Then, by using the

DOI: 10.54254/2755-2721/2026.BA32096

embeddings aggregated during the compression operation, the input features are recalibrated through the embedding dimension d. The specific formula is as follows:

$$\mathrm{s} = \mathrm{Sigmoid}(W_2 \mathrm{ReLU}(W_{1z})) \tag{2}$$

$W_1 \in R^{\frac{d}{r} \times d}$, and $W_2 \in R^{\frac{d}{r} \times d}$, the Sigmoid and ReLU functions are activated by linear parameters, and the final output of the block is obtained by recalibrating the input features through bottleneck activation.

2.1.3. Modal enhancement and module interaction

The research team combined two MEMs and two MIMs to form a basic fusion layer. The cross-modal fusion transformer is composed of multiple sets of fusion layers. The structures of MEM and MIM are shown in the Figure 3, MEM consists of a multi-head attention layer and a BEF. The MEM mechanism can be summarized as [6]:

$$X_{ME} = \mathrm{LN}\left(\widetilde{X_{ME}} + \mathrm{BEF}\left(\widetilde{X_{ME}}\right)\right) \tag{3}$$

$$\widetilde{X_{ME}} = LN\left(X + \mathrm{MultiHead}\left(\widetilde{X}, \widetilde{X}, \widetilde{X}\right)\right) \tag{4}$$

$$\widetilde{X} = X + P_X \tag{5}$$

In the formula， $\mathrm{X} \in R^{N \times HW \times d}$ is the input feature of a certain mode, $P_X \in R^{N \times HW \times d}$ is the spatial position code, $P_X \in R^{N \times HW \times d}$ is for MEM output. MEM can adaptively enhance the semantic information of features through its multi-head self-attention and bottleneck excitation.

MIM consists of a multi-head cross-attention layer, two normalization layers and a BEF. MIM takes the characteristics of the output branches of MEM as input, carry out multimodal fusion. The characteristics of different branches of MIM play different roles. The expression ability of the fused features was enhanced through their respective BEFs.

$$X_{\mathrm{fusion}} = X_{MI}^{r} + X_{MI}^{d} \tag{6}$$

$$\mathrm{X}_{MI} = \mathrm{LN}\left(\widetilde{X_{MI}} + \mathrm{BEF}\left(\widetilde{X_{MI}}\right)\right) \tag{7}$$

$$\widetilde{X_{MI}} = LN\left(X_q + \mathrm{MultiHead}\left(\widetilde{X_q}, \widetilde{X_{kv}}, \widetilde{X_{KV}}\right)\right) \tag{8}$$

$$\widetilde{X_q} = \mathrm{X}_q + \mathrm{P}_q \tag{9}$$

$$\widetilde{X_{kv}} = \mathrm{X}_{kv} + \mathrm{P}_{kv} \tag{10}$$

DOI: 10.54254/2755-2721/2026.BA32096

In the formula, deployment $X_q \in R^{N\times HW\times d}$ is the input of the branch where the module is located, $\mathrm{P}_q \in R^{N\times HW\times d}$ is the spatial position code corresponding to X_q. $\mathrm{X}_{kv} \in R^{N\times HW\times d}$ is the input of another branch, $\mathrm{P}_{kv} \in R^{N\times HW\times d}$ is the spatial encoding of X_{kv}. $\mathrm{X}_{MI} \in R^{N\times HW\times d}$ is the output of MIM. $X_{\text{fusion}} \in R^{N\times HW\times d}$ is obtained by superimposing RGB and depth stream MIM features.

2.2. MMNet model based on multimodal fusion

Due to the current development of sensors, the unimodal method has been substantially improved. And research on human action recognition has also taken note of this, such as methods based on skeleton and RGB video. However, this unimodal method also has obstacles, such as the input of the skeleton being limited by the lack of texture and appearance features. The input of RGB is also subject to the problem of inefficiency due to the lack of a three-dimensional structure. The introduction of multimodal methods among them can effectively improve the performance of HAR [7]. However, how to effectively integrate data and improve the accuracy of recognition has also become a problem. Therefore, the research team introduced the MMNet model in an attempt to address these issues, aiming to enhance recognition accuracy by leveraging complementary information among data from different models. The MMNet model takes the skeleton and RGD video as input and employs three independent neural networks. Input learning from three aspects: bone joints, bone structure, and RGD videos. Then, the ST-ROI constructed through RGB input is fused with the skeletal information learned by the graph convolutional neural network model for feature fusion on the model.

2.2.1. Construct ST-ROI

To reduce the input of a large amount of video data, the research team designed to construct an ST-ROI feature map to focus on the different appearance features of the entire RGB video frame. First, mark $\mathrm{V} = \left\{V^{(i)} i = 1, \ldots, N\right\}$ as the RGB mode. It contains N video samples, within the time interval [1,T]. The video sequence can be represented as $\mathrm{V}^{(i)} = \left(f_1^{(i)}, \ldots, f_t^{(i)}, \ldots, f_T^{(i)}\right)$, $f_t^{(i)}$ represents the frame at time t, to crop the spatial ROI from action videos. From this, the OpenPose tool was introduced to retrieve skeleton joints. At one video frame, a transformation function g is defined to define and construct the space of a joint. The formula is as follows:

$$R_{tj}^{(i)} = g\left(f_t^{(i)}, O_{tj}^{(i)}\right),\ j \in (m_1, \ldots, m_{Mo}),\ M_o \le M_o \tag{11}$$

In the formula, $O_{tj}^{(i)}$ represents the JTH joint of the OpenPose skeleton at time t, m_1 to m_{Mo} refer to the MIO of the skeletal joints of OpenPose in the video sequence. At the time interval t= $\{\text{interval} \times \mathrm{l} | \mathrm{l} = 1, \ldots, \mathrm{L}, \text{interval} = \mathrm{T}/\mathrm{L}\}$, connect each video frame into a square ST-ROI. Thereby extracting the appearance of the image and the movement information of the action [8,9].

2.2.2. Learn joint weights from skeletal patterns

In order to collect information from the skeletal modalities, the research team constructed a spatiotemporal graph to represent the skeletal framework, as shown in Figure 1.

DOI: 10.54254/2755-2721/2026.BA32096

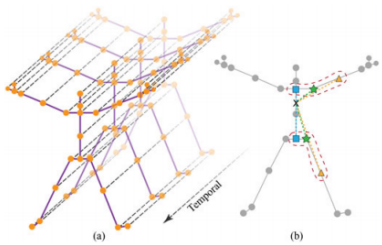

Figure 1. Spatiotemporal diagram skeletal framework diagram (picture credit : original)

First, set M joints in the skeleton frame observed at time t as $J_t^{(i)} = \left(J_{t1}^{(i)}, \ldots, J_{tj}^{(i)}, \ldots, J_{tM}^{(i)}\right)$, $J_{tj}^{(i)} \in R^C$. Among them, the bones and joints in a single skeleton are represented by orange circles. And connect them with purple straight lines, different adjacent skeletons are directly connected at the joints with black. The attributes of the graph vertices can be the three-dimensional coordinates corresponding to each joint. The skeleton diagram of $\mathrm{J}^{(i)}$ can be represented as $\mathrm{G} = (\mathrm{V}, \varepsilon)$. Among them, V and ε represent joints and bones respectively. Node set $\mathrm{V} = \left\{v_{ij} \middle| v_{ij} = J_{tj}^{(i)}, \mathrm{t} = 1, \ldots, \mathrm{T}, \mathrm{j} = 1, \ldots, \mathrm{M}\right\}$ contains all the joints of the input. The edge set $\varepsilon = \left\{\varepsilon_t \middle| \varepsilon_t = B_t^{(i)}, \mathrm{t} = 1, \ldots, \mathrm{T}, \mathrm{j}, \mathrm{k} = 1, \ldots, \mathrm{M}\right\}$ represents all the input bones. To represent the sampling area of the convolution operation, the domain set of A node v_{ti} is defined as $N(v_{ti}) = \{v_{ti} d(v_{ti}, v_{tj}) \leq D\}$, and D is the maximum path length of $d(v_{ti}, v_{tj})$.

3. Analysis of typical multimodal applications

Multimodal information fusion utilizes the complementary advantages of different data sources to overcome the inherent limitations of single-modal methods. It has demonstrated outstanding efficacy in human motion recognition in various practical application scenarios [10,11]. For instance, in the intelligent monitoring system deployed in nursing homes, RGB video streams and depth sensor data are integrated through a dual-stream cross-modal fusion converter (DSCMT). It can reliably detect slip incidents even under harsh lighting conditions and partial occlusion. Compared with the method that only uses RGB. The false alarm rate has decreased by more than 15% in interactive games and virtual reality environments. The skeletal joint trajectories extracted by the pose estimation algorithm are fused with the spatiotemporal ROI representations of RGB frames using the MMNet architecture. Achieved a real-time gesture recognition accuracy rate of over 98%, this brings about a more immersive and responsive user experience [4].

In the field of medical rehabilitation,Multimodal fusion of wearable inertial measurement unit (IMU) signals and video-based pose analysis has been used to accurately quantify patients' movement indicators (such as range of motion and movement smoothness), Thus achieving personalized treatment adjustments, compared with the assessment that solely relies on motion capture. The result of functional recovery has improved by 20%. Similarly, motion analysis applications benefit from combining high-speed camera lenses with gyroscope and accelerometer data from sensors worn by athletes. The combined processing of visual and inertial cues enables the precise identification of complex motion patterns (such as tennis serves or gymnastic vault). The error rate is less than 5%, and it can provide real-time biomechanical feedback to improve athletic performance. In autonomous driving and pedestrian safety systems, the integration of lidar point clouds, stereo vision depth maps and channel attention mechanisms in multimodal deep networks

DOI: 10.54254/2755-2721/2026.BA32096

can accurately predict the intentions of pedestrians (such as starting to cross the road or suddenly stopping). This enhances the ability to predict risks and reduces reaction delay by up to 30% [4].

Industrial robot technology also employs multimodal fusion, by jointly analyzing the tactile sensor array, force-torque measurement and visual monitoring. Realize safe human-machine collaboration on the assembly line, enable the robot to dynamically adjust its grasping force and trajectory based on contact feedback and visual prompts for hand movements.

4. Conclusion

This paper summarizes the action recognition methods under multimodal information fusion. The main focus is on comparing two methods: the dual-stream cross-modal fusion transformer and the model-based multimodal neural network. The first type of method achieves fusion through the modal enhancement module and the modal interaction module. The second type of method provides graph convolution and spatio-temporal ROI fusion mechanisms. The two methods provide new ideas for action recognition from the perspectives of feature interaction and integration. However, multimodal fusion has also brought about problems such as high computing power overhead and complex model deployment. The robustness of the existing work under the guarantee of real-time performance needs to be further addressed. Future research can focus on lightweighting and addition, as well as the expansion and optimization of application scenarios.

References

[1] Gandhi, A., Adhvaryu, K., Poria, S., et al. (2023). Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion, 91, 424–444.

[2] Cheng, Q. (2023). Research on human action recognition based on multimodal fusion (Master's thesis). Guilin University of Electronic Technology, Guilin, China.

[3] Yadav, S. K., Tiwari, K., Pandey, H. M., et al. (2021). A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. Knowledge-Based Systems, 223, 106970.

[4] Liu, Z., Cheng, J., & Liu, L. B. (2022). Dual-stream cross-modality fusion transformer for RGB-D action recognition. Knowledge-Based Systems, 255, 109741.

[5] Yu, B. X. B., Liu, Y., Zhang, X., et al. (2023). MMNet: A model-based multimodal network for human action recognition in RGB-D videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3), 3522–3538.

[6] Hu, S., Fan, J., Zou, G., et al. (2025). Multi-aspect self-guided deep information bottleneck for multi-modal clustering. In Proceedings of the AAAI Conference on Artificial Intelligence, 39(16), 17314–17322.

[7] Zhu, X., Feng, H., & Cambria, E. (2025). EMVAS: End-to-end multimodal emotion visualization analysis system. Complex & Intelligent Systems, 11(8), 1–15.

[8] Zhang, B., Han, Y., & Han, X. (2025). Research on multi-modal retrieval system of e-commerce platform based on pre-training model. Artificial Intelligence Technology Research, 2(9).

[9] Wang, R., Xu, D., Cascone, L., et al. (2025). RAFT: Robust adversarial fusion transformer for multimodal sentiment analysis. Array, 100445.

[10] Wang, J., Gao, M., & Zhai, W. (2025). Knowledge generation and distillation for road segmentation in intelligent transportation systems. IEEE Transactions on Intelligent Transportation Systems.

[11] Zhu, X., Wang, Y., & Cambria, E. (2025). RMER-DT: Robust multimodal emotion recognition in conversational contexts based on EMS, diffusion and transformers. Information Fusion, 103268.