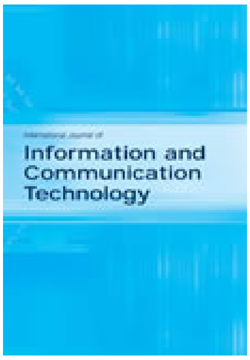


**International Journal of Information and Communication Technology**

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

# Application of distributed artificial intelligence technology in key frame extraction of film and television video

Feng Cheng

**DOI:** 10.1504/IJICT.2025.10074511

**Article History:**

| | |
|---|---|
| Received: | 01 August 2025 |
| Last revised: | 17 September 2025 |
| Accepted: | 26 September 2025 |
| Published online: | 20 November 2025 |

# Application of distributed artificial intelligence technology in key frame extraction of film and television video

Feng Cheng

School of Arts,
Xi'an International Studies University,
Xi'an, 710100, China
Email: chengf1985@outlook.com

**Abstract:** Traditional video analysis relies on video frames, which often contain redundant data, making key frame extraction essential. However, existing methods frequently suffer from missing or redundant frames. To address this, this paper proposes a video key frame extraction method based on distributed artificial intelligence. First, mutual information between video frames is calculated. Then, SIFT feature points are extracted and transformed into polar coordinates, with each frame divided into sector regions to count feature points and compute inter-frame distances. To enhance precision, the CaffeNet model is adopted as a deep neural network to extract deep features using three training techniques. This approach significantly improves the accuracy of key frame extraction. Experimental results show that the proposed method achieves higher fidelity and compression rates than traditional techniques, and the extracted key frames align closely with reference standards without frame omission, demonstrating its effectiveness and robustness in real-world applications.

**Keywords:** distributed artificial intelligence technology; film and television video; key frame extraction; SIFT feature points.

**Reference** to this paper should be made as follows: Cheng, F. (2025) 'Application of distributed artificial intelligence technology in key frame extraction of film and television video', *Int. J. Information and Communication Technology*, Vol. 26, No. 41, pp.43–58.

**Biographical notes:** Feng Cheng obtained his Master's in Broadcast Television Arts (2013) from Shaanxi Normal University, Xi'an. Presently, he is working as a Lecturer in the School of Arts, Xi'an International Studies University, Xi'an. He was invited as a Performance Director Consultant to give various lectures on teaching and practical activities. He has mentored numerous students in theatrical creation, with the plays they wrote and directed receiving various national and provincial honours. He has published articles in various core Chinese journals and conference proceedings. His research interests include theatrical performance, drama writing, drama criticism, film and television culture studies, and theatre education.

## 1 Introduction

With the rapid development of computer and multimedia technology, multimedia data carried by images, audio, and video has exploded. Among them, video data has gradually become the main form of information interaction due to its large amount of information and strong expressiveness (Yuan et al., 2021; Elahi and Yang, 2021). In tasks such as video content management, retrieval, and summary generation, accurate key frame extraction is a core step (Singh and Kaur, 2020). Key frames can effectively summarise the main content of a video with a small amount of data, thereby reducing the complexity of indexing and retrieval and improving users' ability to quickly understand the video within a limited time (Bommisetty et al., 2019; Prathiba and Kumari, 2021). Traditional key frame extraction methods mainly rely on low-level descriptions such as colour histograms, motion vectors, and texture features (Lingam and Reddy, 2020). However, such methods often suffer from the problem of redundant or missing frames. For example, key frame extraction methods based on global motion statistics are prone to failure in complex shot switching (Zhong et al., 2020); although the semantic correlation (SC) method improves the accuracy of frame selection through hierarchical clustering and histogram comparison, the computational cost is too high in large-scale video data processing (Wang and Lu, 2021); the feature fusion method combines deep features extracted by convolutional neural networks (CNN) with manual features, which improves the accuracy of key frame discrimination to a certain extent, but still lacks robustness and generalisation in different types of videos (Zhang and Zhang, 2019; Li et al., 2021). Therefore, how to further enhance robustness and generalisation while ensuring high fidelity and high compression ratio remains an important problem that needs to be solved in the field of key frame extraction. To address the above challenges, this paper proposes a key frame extraction method based on distributed artificial intelligence (DAI). Mutual information (MI) calculations are used to discriminate the correlation between adjacent frames, and the scale-invariant feature transform (SIFT) and polar sectorisation are combined to refine local feature differences. Finally, the CaffeNet network structure is improved for deep feature learning, and unsupervised, semi-supervised, and supervised retraining strategies are combined to improve the method's robustness and adaptability in complex environments.

The main contributions and innovations of this paper are as follows:

1 proposing a method that combines MI with SIFT under polar coordinate partitioning, and introducing a DAI framework to improve the accuracy and efficiency of key frame discrimination

2 in the deep convolutional neural network CaffeNet, improvements are made to the feature extraction strategy by addressing the limitations of convolutional layer (CL) and fully connected layer (FCL), thereby enhancing the modelling capability of global video semantics

3 designing unsupervised, semi-supervised, and supervised retraining methods to fully leverage label information and user feedback, improving the robustness and generalisation ability of the model in complex scenarios

4 the proposed method achieves fidelity and compression ratio results consistent with the reference standards, demonstrating strong practical value and application prospects.

## 2 Video key frame extraction

### *2.1 Video frame MI calculation*

$X$ is defined as the set of events that may occur in a random event, that is, $X = \{x_1, x_2, \ldots, x_n\}$, $p$ is the probability distribution of event $X$, $p \geq 0$, and $\sum_{x \in X} p_x(x) = 1$, then the entropy of random variable $X$ is:

$$H(X) = -\sum_{x \in X} p_x(x) * \log p_x(x) \tag{1}$$

The joint entropy of $X$ and $Y$ is:

$$H(X, Y) = -\sum_{x, y \in X, Y} p_{xy}(x, y) * \log p_{xy}(x, y) \tag{2}$$

In the formula, $p_{xy}(x, y)$ is the joint density function of $X$ and $Y$.

The mutual information (MI) between $X$ and $Y$ is:

$$I(X, Y) = -\sum_{x, y \in X, Y} p_{xy}(x, y) * \log \frac{p_{xy}(x, y)}{p_x(x) \times p_y(y)} \tag{3}$$

If $X$ and $Y$ are independent random variables, the MI has the following properties:

1 $I(X, Y) \geq 0$

2 if both $H(X)$ and $H(Y)$ are zero, then $I(X, Y) = 0$

3 $I(X, Y) = H(X) + H(Y) - H(X, Y)$.

Film and television video frames (VFs) can be regarded as a two-dimensional random variable. The entropy $H(f_t)$ of VF represents the average amount of information contained in VF. In film and television video, the amount of MI can represent the degree of correlation between adjacent frames, that is, the smaller the amount of MI $I(f_t, f_{t+1})$, the more uncorrelated the adjacent frames $f_t$ and $f_{t+1}$ (Singh and Kaur, 2020). In practical application, we can choose the colour, shape, texture and other information of the VF to calculate the entropy and MI of the VF. Taking colour as an example, the colour can be divided into three independent hue, saturation, value (HSV) colour spaces of chroma, saturation and brightness, and the colour of VF can be expressed as $C = aH + bS + V$, in which the weight is determined according to experience.

Thus, the entropy of $f_t$ is:

$$H(f_t) = aH_t^H + bH_t^S + H_t^V \tag{4}$$

The MI of adjacent frames $f_t$ and $f_{t+1}$ is:

$$I(f_t, f_{t+1}) = aI_{t,t+1}^H + bI_{t,t+1}^S + I_{t,t+1}^V \tag{5}$$

## 2.2 SIFT feature point distance calculation

Although MI can effectively characterise the overall correlation between adjacent frames, it is difficult to fully reflect the differences in local details by relying solely on global statistical features. Therefore, this paper further introduces SIFT to enhance the description of inter-frame differences. Calculate the SIFT feature points (SIFTFP) of film and television VFs, then divide each frame into multiple sector areas, convert the SIFTFPs to polar coordinates, count the number of SIFTFPs in each sector and calculate the inter frame distance (Gu et al., 2020; Zhao et al., 2019; Jian et al., 2019; Abulizi, 2019). The calculation process of SIFTFP distance is as follows.

The VF is separated into many sector zones to make the distribution of SIFTFPs in the VF more understandable. Since any part of the VF might show up in the key frame, the VF is separated into many parts based on various angles and radii. Since key frames can appear anywhere in the image, it is necessary to segment the frame based on radius and angle. In this experiment, the radius step size is set to 20 pixels, dividing the image into five concentric rings; the angle step size is set to 15°, forming 24 sectors. $h$ represents the height of the VF to be partitioned, and $w$ is the width of the frame. The center of VF in rectangular coordinate system is $O(x_o, y_o)$, its abscissa $x_o = w/2$ and ordinate $y_o/2$. Taking the centre point of the VF as the origin, the frame is divided into multiple sector regions according to different radii and angles. After the frame partition is completed through the above processing, the distribution quantity of SIFTFPs in each area in the frame is calculated. Count the quantity of feature points (FPs) in each area and convert all FPs from rectangular coordinate system to polar coordinates with $O$ point in the centre of VF as the polar centre (Li et al., 2020). $(x_i, y_i)$ represents the rectangular coordinates of the $i^{\text{th}}$ SIFTFP, and the polar coordinates of the FP $(x_i, y_i)$ are expressed as $(r_i, \theta_i)$.

Through formula (6), calculate the coordinate value $(x_i', y_i')$ of the FP with coordinate $(x_i, y_i)$ under the rectangular coordinate system with the frame centre as the origin:

$$\begin{cases} x_i' = x_i - w/2 \\ y_i' = y_i - h/2 \end{cases} \tag{6}$$

Convert each FP into a polar coordinate system in which the centre point $O(w/2, h/2)$ of the VF is the polar centre, and calculate the polar coordinate $(r_i, \theta_i)$ corresponding to sift point $(x_i, y_i)$ in the polar coordinate system according to formula (7):

$$\theta_i = \arctan\left(\frac{y_i' - y_o^i}{x_i' - x_o^i}\right), r_i = \sqrt{\left(x_i' - x_o^i\right)^2 + \left(y_i' - y_o^i\right)^2} \tag{7}$$

Repeat the above steps to obtain the polar coordinates of all FPs, and count the distribution quantity of FPs in each sector. To calculate the SIFT distribution distance (DD), first count the quantity of FPs in each region and count them into $count[i][j]$. Where $i$ and $j$ display the radius $r$ range and angle $\theta$ range of SIFTFPs respectively. When $r = 0$, it means that the pixel is within the smallest circle ($r_1 = w/6$); when $r = 2$, it means that the pixel is within or outside the largest circle ($r_3 = w/2$); when $0 < \theta < \pi/4$, the array $j$ value corresponding to the angle of the FP is 0.

After obtaining the number of SIFT points in each sector area in the VF, the region is determined according to the polar coordinate ($r$, $\theta$) of the FP $S_i$. According to formula (7), the $(i, j)^{\text{th}}$ sector of the sector to which the FP ($r$, $\theta$) belongs can be calculated, and this FP is included in *count*[*i*][*j*].

$$i = \frac{1}{(w/6)}, \; j = \begin{Bmatrix} \dfrac{\theta}{(\pi/4)} x' > 0, y' > 0 \\ \dfrac{\theta + \pi}{(\pi / 4)} x' < 0 \\ \dfrac{\theta + 2\pi}{(\pi / 4)} x' > 0, y' < 0 \end{Bmatrix} \tag{8}$$

Judge the area of SIFTFPs one by one, and obtain a two-dimensional array for counting the number of fan-shaped FPs, which is used to describe the distribution of target FPs.

According to the extracted SIFTFPs, set the two consecutive needles in the film and television video as $f_k$ and $f_{k+1}$ respectively, and the corresponding SIFTFP distribution arrays are $count_k[i][j]$ and $count_{k+1}[i][j]$, respectively. Calculate the DD of FPs of two VFs in film and television video as:

$$SiftCountDiff\left(f_k, f_{k+1}\right) = \sqrt{\sum_{i=0}^{a} \sum_{j=0}^{b} \left( \left( count_k[i, j] - count_{k+1}[i, j] \right)^2 \right)} \tag{9}$$

Because the background may be falsely detected as the target in the video and video, the difference of *SiftCountDiff* increases sharply. Using the average value as the threshold to extract key frames can easily lead to missed detection of key frames where the changes are fast. Therefore, the degree of feature change is measured by the ratio of the DD of the FPs between neighbouring frames, as shown in equation (10)

$$MotionChange(k, k+1) = SiftCountDiff(k+1) / SiftCountDiff(k) \tag{10}$$

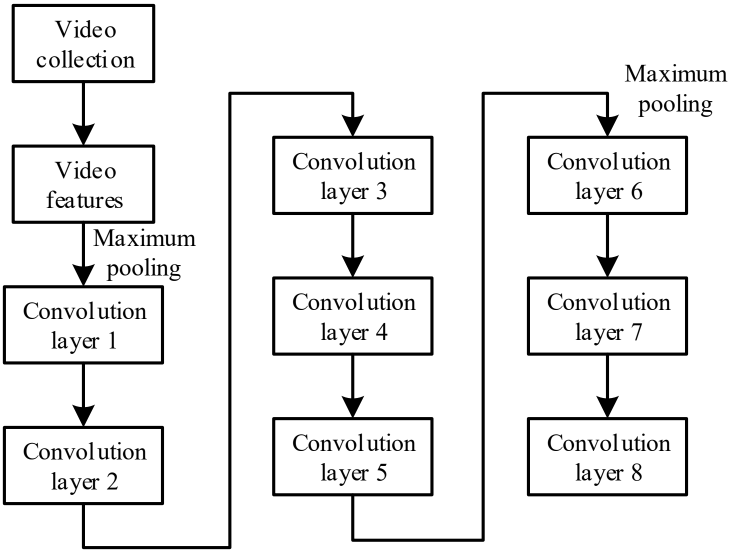
## 2.3 Video key frame extraction

While the SIFTFP distance provides a measure of inter-frame differences, the proposed DAI framework further integrates a convolutional neural network (CNN) model, namely CaffeNet, to extract the key frames of film and television videos. This distributed architecture enables parallel processing of video data, thereby improving scalability and efficiency. To fully leverage the previous structural calculations, the CNN is further enhanced to extract a portion of the KF information from the external memory (Yasin et al., 2020).

The CaffeNet model, the outcome of the AlexNet model trained on the ImageNet large-scale dataset, is used as the DNN model in this study. Eight trained NN layers make up the model: the first five are CLs, followed by FCLs and max pooling layers in the first, second, and fifth CLs. The third FCL produces a distributed result with 1,000 ImageNet classes, whereas the previous two FCLs employ the ReLU nonlinear activation function (AF).

Figure 1 depicts the general architecture of the CaffeNet using softmax loss for model training.

**Figure 1** The overall architecture of the CaffeNet DNN



This article has modified CaffeNet for the following reasons:

1 the FCL lacks spatial information because it is linked to every input neuron, but the CL does because the AF is a geographically structured structure

2 the FCL requires far more model parameters than the CL

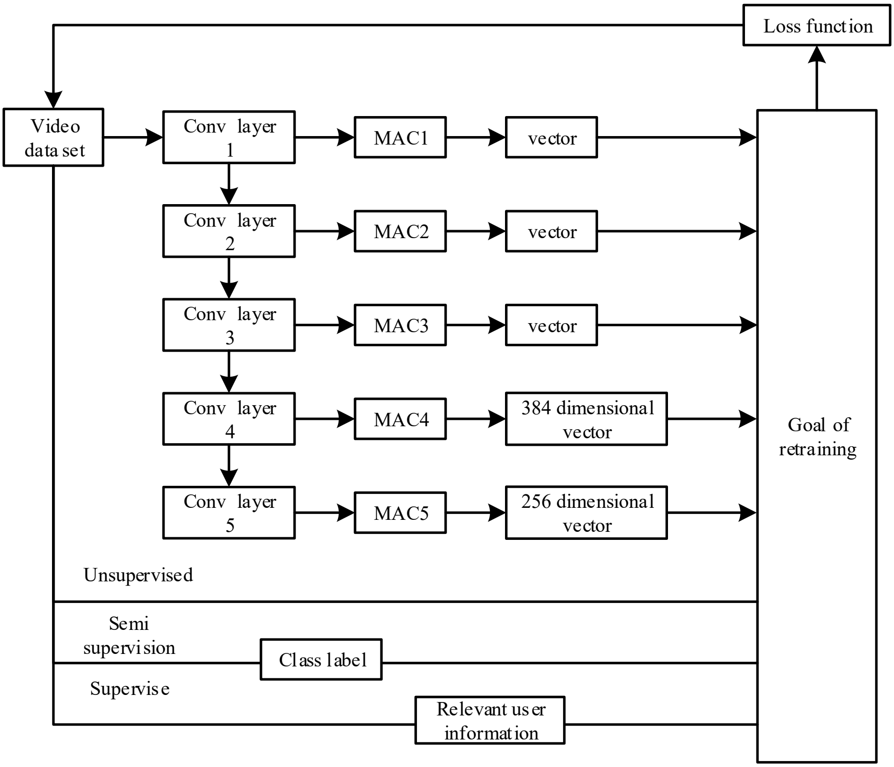
3 while the CL is exempt from this requirement, the FCL must forecast the input image's size.

Consequently, feature extraction is the responsibility of the CL rather than the FCL in the suggested strategy.

Considering that CL can preserve spatial position information, while FCL typically loses this information, hindering subsequent keyframe identification, previous research has shown that high-level convolutional features can better express global semantic information while requiring far fewer parameters than FCL, which helps improve efficiency and generalisation (Lan et al., 2023; Shamsipour et al., 2024). Therefore, using CaffeNet's high-level convolutional features for keyframe extraction is a reasonable and efficient design choice. This study's changes make use of either a fourth CL, Conv4, or a fifth CL, Conv5. Conv5 has dimensions of 13 × 13 × 384 features. Consequently, each VF receives a 256-dimensional feature representation, or a 384-dimensional feature representation, from the MAC layer.

The process of the retraining scheme in this paper is shown in Figure 2.

**Figure 2** Flowchart of deep learning-based retraining scheme



### *2.3.1 Unsupervised retraining*

Unsupervised retraining maximises the cosine similarity (CS) between each television and movie video key frame and its closest '*n*-KF' by retraining the previously learned CNN model using the provided dataset.

Suppose $I = \{I_1, I_2, \ldots, I_M\}$ is the set of $M$ key frames, $X = \{x_1, x_2, \ldots, x_M\}$ is the corresponding features of $M$ key frames, $\vartheta_i$ is the average vector (AV) of the nearest '*n*-key frame' of $x_i$, and the calculation formula of $\vartheta_i$ is:

$$\vartheta_i = \frac{1}{n}\sum_{i=1}^{n} x_l^i \tag{11}$$

The new goal of $I$ can be determined by solving the following optimisation problems:

$$\max_{x_i \in X} O = \max_{x_i \in X} \sum_{i=1}^{M} \frac{x_i^T \vartheta_i}{\|x_i\| \|\vartheta_i\|} \tag{12}$$

The gradient descent method is used to solve formula (12), and the first step of objective function $O$ is:

$$\frac{\partial O}{\partial x_i} = \frac{\vartheta_i}{\|x_i\|\|\vartheta_i\|} - \frac{x_i^T \vartheta_i}{\|x_i\|^3 \|\vartheta_i\|} x_i \tag{13}$$

The update rule of the $v^{\text{th}}$ iteration of each VF is:

$$x_{i,v+1} = \eta \left( \frac{\mu^i \vartheta_i}{\|x_{i,v}\|\|\vartheta_i\|} - \frac{x_{i,v}^T \vartheta_i}{\|x_{i,v}\|^3 \|\vartheta_i\|} x_{i,v} \right) + x_{i,v} \tag{14}$$

The update rules are normalised, which is described as the following formula, to better regulate the NN's learning rate:

$$x_{i,v+1} = \eta \|\vartheta_i\|\|x_{i,v}\| \left( \frac{\vartheta_i}{\|x_{i,v}\|\|\vartheta_i\|} - \frac{x_{i,v}^T \vartheta_i}{\|x_{i,v}\|^3 \|\vartheta_i\|} x_{i,v} \right) + x_{i,v} \tag{15}$$

Create a neural networks (NN) regression job, configure the CaffeNet weight, then train the NN using the backward propagation (BP) method using the aforementioned characteristics as the target in the interest layer. The regression task is trained using the Euclidean distance loss training function.

### 2.3.2 Semi ST

Semi supervised retraining (ST) uses the relevant information (RI) derived from available class tags to improve the performance of depth descriptors. Suppose $(x_i, y_i)$ represents a descriptor with a label, where $x_i$ is the key frame descriptor and $y_i$ is the label corresponding to the key frame. In this paper, the convolutional neural layer of CNN model is adjusted to maximise the CS between $x_i$ and the nearest '$m$-correlation descriptor', while minimising the CS between $x_i$ and the nearest '$l$-correlation descriptor'. Related frames are key frames that belong to the same class, while irrelevant frames are key frames that belong to distinct classes (Wang and Zhu, 2020; Zhang et al., 2022).

Suppose $I = \{I_1, I_2, \ldots, I_M\}$ is the set of $M$ image frames with RI, $x = F_L(I)$ is the $L$-layer output results of the training CNN model for the input VF set $I$, $X = \{x_1, x_2, \ldots, x_M\}$ is the characteristics of the corresponding VF, and $R_i = \{r_1, r_2, \ldots, r_k\}$ is the set of $K^i$-related descriptors of the $i^{\text{th}}$ key frame. The new descriptor of key frame set $I$ is determined by solving the following two optimisation problems:

$$\max_{x_i \in X} O^+ = \max_{x_i \in X} \sum_{i=1}^{M} \frac{x_i^T \vartheta_i^+}{\|x_i\|\|\vartheta_i^+\|} \tag{16}$$

$$\min_{x_i \in X} O^- = \min_{x_i \in X} \sum_{i=1}^{M} \frac{x_i^T \vartheta_i^-}{\|x_i\|\|\vartheta_i^-\|} \tag{17}$$

In the formula, $\vartheta_i^-$ is the AV of the nearest '$l$-descriptor' of $x_i$, and $\vartheta_i^+$ is the AV of the nearest '$m$-descriptor' of $x_i$.

Iteration $v$'s normalisation update rule is modified to use the following formula:

$$x_{i,v+1} = \psi_1 \|\vartheta_i^+\|\|x_{i,v}\| \left( \frac{\vartheta_i^+}{\|x_{i,v}\|\|\vartheta_i^+\|} - \frac{x_{i,v}^T \vartheta_i^+}{\|x_{i,v}\|^3 \|\vartheta_i^+\|} x_{i,v} \right) + x_{i,v} \tag{18}$$

$$x_{i,v+1} = \beta_1 \left\|\vartheta_i^-\right\| \left\|x_{i,v}\right\| \left( \frac{\vartheta_i^-}{\left\|x_{i,v}\right\| \left\|\vartheta_i^-\right\|} - \frac{x_{i,v}^T \vartheta_i^-}{\left\|x_{i,v}\right\|^3 \left\|\vartheta_i^-\right\|} x_{i,v} \right) + x_{i,v} \tag{19}$$

Fuse the normalisation update rules, that is, add formulas (18) and (19):

$$\begin{aligned} x_{i,v+1} = {} & \psi_1 \left\|\vartheta_i^+\right\| \left\|x_{i,v}\right\| \left( \frac{\vartheta_i^+}{\left\|x_{i,v}\right\| \left\|\vartheta_i^+\right\|} - \frac{x_{i,v}^T \vartheta_i^+}{\left\|x_{i,v}\right\|^3 \left\|\vartheta_i^+\right\|} x_{i,v} \right) \\ & - \beta_1 \left\|\vartheta_i^-\right\| \left\|x_{i,v}\right\| \left( \frac{\vartheta_i^-}{\left\|x_{i,v}\right\| \left\|\vartheta_i^-\right\|} - \frac{x_{i,v}^T \vartheta_i^-}{\left\|x_{i,v}\right\|^3 \left\|\vartheta_i^-\right\|} x_{i,v} \right) + x_{i,v} \end{aligned} \tag{20}$$

BP technology can retrain the NN using the target description mentioned above.

### *2.3.3 Supervise retraining*

The principle of relevant feedback is where the concept of supervised training originated. ST takes into account user feedback information, which is made up of video queries. In order to maximise the CS between the given query and its connected key frames and decrease the CS between the specified query and its unrelated key frames, this technique aims to adjust the model parameters.

Suppose $Q = \{Q_1, Q_2, \ldots, Q_K\}$ is the extraction set, $I_+^k = \{I_1, I_2, \ldots, I_Z\}$ represents the relevant key frame of the specified query, $I_-^k = \{I_1, I_2, \ldots, I_J\}$ represents the irrelevant key frame of the specified query, $x = F_L(I)$ represents the $L$-layer output of the key frame $I$ and the trained CNN model, and $r = F_L(Q)$ represents the $L$-layer output of the NN for the extracted key frame (Akilan et al., 2019).

The new targets of the query's relevant and irrelevant key frames can be found by resolving the following optimisation issues:

$$\max_{x_i \in X_+^k} O^+ = \max_{x_i \in X_+^k} \sum_{i=1}^{Z} \frac{x_i^T r_k}{\left\|x_i\right\| \left\|r_k\right\|} \tag{21}$$

$$\min_{x_i \in X_-^k} O^- = \min_{x_i \in X_-^k} \sum_{j=1}^{J} \frac{x_j^T r_k}{\left\|x_j\right\| \left\|r_k\right\|} \tag{22}$$

The following two formulae represent the normalised update rule of iteration $v$:

$$x_{i,v+1} = \tau \left\|r_k\right\| \left\|x_{i,v}\right\| \left( \frac{r_k}{\left\|x_{i,v}\right\| \left\|r_k\right\|} - \frac{x_{i,v}^T r_k}{\left\|x_{i,v}\right\|^3 \left\|r_k\right\|} x_{i,v} \right) + x_{i,v} \tag{23}$$

$$x_{j,v+1} = \tau \left\|r_k\right\| \left\|x_{j,v}\right\| \left( \frac{r_k}{\left\|x_{j,v}\right\| \left\|r_k\right\|} - \frac{x_{j,v}^T r_k}{\left\|x_{j,v}\right\|^3 \left\|r_k\right\|} x_{j,v} \right) + x_{j,v} \tag{24}$$

The DNN is retrained using the aforementioned key frame description as the goal of the layer of interest.

Through the above training, the key frames of film and television video are extracted.

## 3 Experimental verification

This paper uses Visual C++ to implement the key frame extraction method used in this study, and implements it in the environment of Intel i7, 2.4 GHz CPU, 4 GB memory, Windows 8 (64-bit) to verify the effectiveness of the method.

### 3.1 Experimental data

The length of the experimental video used in this paper ranges from hundreds of frames to several thousand frames, and the types are rich, including advertisements, animations, movies, and news. The sampling frequency of the VFs used in the experiment is 60 frames/sec.

Twenty shots were randomly selected from the above four types of film and television videos for comparative analysis of key frame extraction. The content changes of these shots, the number of frames included, and the size of frame images were different. Table 1 displays the essential data.

**Table 1** Experimental dataset

| *Video type* | *Number of shots* | *Total frames* | *Number of key frames* | *Resolution* | *Average duration (s)* | *Content complexity* |
|---|---|---|---|---|---|---|
| Advertisement | 20 | 5350 | 106 | 1,280 × 720 | 15–40 | Frequent scene changes, high motion |
| Animation | 20 | 4426 | 92 | 1,920 × 1,080 | 20–60 | Stylised content, moderate motion |
| Film | 20 | 8440 | 248 | 1,920 × 1080 | 60–180 | Complex scenes, high variability |
| Journalism | 20 | 6473 | 190 | 1,280 × 720 | 30–120 | Dialogue and scene alteration |

### 3.2 Analysis of experimental results

#### 3.2.1 Fidelity and compression ratio

The measuring standards used are fidelity and compression ratio in order to impartially assess the efficacy and rationale of the chosen key frames. A good key frame extraction technique should have a high compression rate and great fidelity.

For a single shot, fidelity is defined as the semi-Hausdorff distance between the key frame set and all frames in the shot, calculated as:

$$F\left(S_i, KF_i\right) = 1 - d\left(S_i, KF_i\right) \tag{25}$$

In the formula, $S_i$ represents the $i^{\text{th}}$ shot, $KF_i$ represents the key frame set selected from shot $S_i$, and $d(S_i, KF_i)$ represents the distance between shot $S_i$ and key frame set $KF_i$. For a video clip, the average fidelity can be used as the measurement index, and the calculation formula of the average fidelity is:

$$F_{avg} = \frac{1}{N_s} \sum_{i=1}^{N_s} F\left(S_i, KF_i\right) \tag{26}$$

In the formula, $F(S_i, KF_i)$ is the fidelity of the $i^{\text{th}}$ shot, and $N_s$ is the quantity of shots in a video clip.

In the process of calculating $d(S_i, KF_i)$, the histogram intersection method is used to calculate the distance between two images, which can be calculated by the following formula:

$$Diff\left(f_i, f_j\right) = 1 - \sum_{k=0}^{k} \min\left(H_i(k), H_j(k)\right) \tag{27}$$

In the formula, $H_i$ and $H_j$ are the normalised HSV colour histograms of image frames $f_i$ and $f_j$ respectively.

For a video sequence, the calculation formula of compression ratio is:

$$R = 1 - \frac{N_{KF}}{N_F} \tag{28}$$

In the formula, $N_{KF}$ is the quantity of key frames selected from the video sequence, and $N_F$ represents the total number of frames of the video sequence.

In the experimental process, the proposed method is compared with the extraction method (EM) based on SC and the EM based on fusion features. Tables 2 and 3 display the fidelity and compression rate findings of the three techniques used to extract key frames from the aforementioned experimental datasets.

**Table 2** Fidelity

| *Video type* | *Fidelity* | | |
|---|---|---|---|
| | *Paper method* | *EM based on SC* | *EM based on fusion feature* |
| Advertisement | 0.650011 | 0.595948 | 0.566267 |
| Animation | 0.783786 | 0.736046 | 0.698335 |
| Film | 0.808362 | 0.603223 | 0.503859 |
| Journalism | 0.782301 | 0.727695 | 0.715206 |

**Table 3** Compression ratio

| *Video type* | *Compression ratio* | | |
|---|---|---|---|
| | *Paper method* | *EM based on SC* | *EM based on fusion feature* |
| Advertisement | 0.985153 | 0.983779 | 0.983006 |
| Animation | 0.989381 | 0.986916 | 0.973271 |
| Film | 0.984375 | 0.978536 | 0.988251 |
| Journalism | 0.991653 | 0.990995 | 0.986493 |

In Table 2, compared to the EMs based on SC and fusion features, the fidelity of the approach used in this study is noticeably greater. In Table 3, the approach in this work can better acquire the global optimal solution of film and television video key frames because it has a greater compression rate level than the two comparative methods.

### 3.2.2 Lens precision and recall

During the experiment, the general evaluation criteria for lens detection are adopted: precision $P$ and recall $Q$.

$$P = \frac{N_c}{N_c + N_f} \times 100\% \tag{29}$$

$$Q = \frac{N_c}{N_c + N_m} \times 100\% \tag{30}$$

In the formula, $N_c$, $N_m$ and $N_f$ represent the correct detection number, missed detection number and false detection number of the lens respectively.

In the process of feature extraction and verification, the method based on semantic fusion is compared with the method based on semantic fusion. The comparison results of lens precision and recall of the three methods are shown in Table 4.

**Table 4** Lens precision and recall of different methods

| *Video type* | *Methods* | *Precision rate* | *Recall rate* |
|---|---|---|---|
| Advertisement | Paper method | 96.3 | 92.6 |
| | EM based on SC | 90.0 | 64.3 |
| | EM based on fusion feature | 59.3 | 57.1 |
| Animation | Paper method | 90.6 | 93.5 |
| | EM based on SC | 87.5 | 90.3 |
| | EM based on fusion feature | 55.3 | 83.9 |
| Film | Paper method | 90.5 | 90.5 |
| | EM based on SC | 88.2 | 71.4 |
| | EM based on fusion feature | 65.2 | 71.4 |
| Journalism | Paper method | 94.4 | 92.7 |
| | EM based on SC | 93.3 | 76.4 |
| | EM based on fusion feature | 72.6 | 81.8 |

To validate the contributions of different modules within the proposed method, ablation experiments were further designed. These involved progressively removing the MI, SIFT feature polar coordinate segmentation, and deep learning retraining strategy, respectively, to assess their impact on keyframe extraction performance. The specific results are presented in Table 5.

**Table 5** Comparative results of ablation experiments

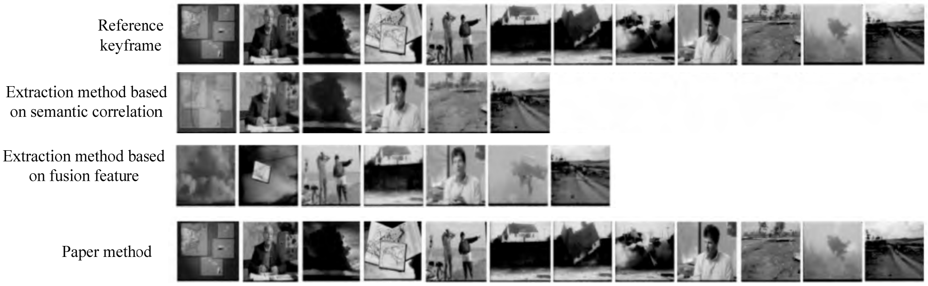
| *Method setup* | *Precision (%)* | *Recall (%)* | *Fidelity* |
|---|---|---|---|
| Removal of mutual information computation | 87.2 | 83.5 | 0.695 |
| Removal of SIFT segmentation | 88.6 | 84.7 | 0.702 |
| Removal of retraining strategy | 85.1 | 81.9 | 0.678 |
| Complete model | 92.9 | 90.1 | 0.783 |

As Table 5 demonstrates, removing either the MI or SIFT module results in varying degrees of performance degradation, indicating that both play a crucial role in complementing local and global features. However, the most pronounced decline occurs when the deep learning retraining strategy is omitted, with fidelity decreasing by approximately 13.4%. This underscores the retraining mechanism's vital contribution to enhancing the model's robustness and adaptability.

### *3.2.3 Key frame extraction effect*

In order to more intuitively express the extraction effectiveness of this method, a film and television video about coastal hurricanes is selected, and 12 key frames are extracted as the reference standard. The key frame extraction experiments are carried out by using the method in this study, the EM based on SC and the EM based on fusion features. The experimental results are shown in Figure 3.

**Figure 3** Comparison experiment of key frame extraction effect


Reference keyframe
Extraction method based on semantic correlation
Extraction method based on fusion feature
Paper method

In Figure 3, the key frame extraction results of the method in this paper are consistent with the reference standards, while more key frames are lost in the SC EM and fusion feature EM. Therefore, it shows that the key frame extraction effect of text method is better.

### *3.2.4 Robustness and generalisation validation*

To further validate the robustness and generalisation capability of the proposed method, this study expanded upon the original experimental dataset with three additional experiments. Firstly, Gaussian noise (σ = 15, 25) and salt-and-pepper noise (noise density 0.05, 0.1) were introduced into video frames to examine keyframe extraction performance under varying noise intensities. Secondly, film and sports event videos featuring rapid motion, abrupt lighting changes, and frequent camera cuts were selected to evaluate the method's stability in complex dynamic environments. Additionally, beyond the original four video categories (advertisements, animations, films, and news), sports event and surveillance footage were incorporated to validate the distributed artificial intelligence framework's applicability across diverse video types. Specific results are presented in Table 6.

**Table 6** Robustness and generalisation validation results

| *Experimental environment* | *Precision (%)* | *Recall (%)* | *Fidelity* |
|---|---|---|---|
| Raw data (baseline) | 92.9 | 90.1 | 0.783 |
| Gaussian noise σ = 15 | 91.2 | 88.5 | 0.764 |
| Gaussian noise σ = 25 | 89.8 | 86.9 | 0.742 |
| Salt-and-pepper noise density = 0.05 | 90.5 | 87.6 | 0.751 |
| Salt-and-pepper noise density = 0.1 | 88.7 | 85.4 | 0.733 |
| Complex scenes (rapid motion) | 90.1 | 87.9 | 0.758 |
| Complex scenes (sudden lighting changes) | 89.6 | 86.7 | 0.746 |
| Cross-category (sports events) | 91.8 | 89.3 | 0.772 |
| Cross-category (surveillance footage) | 92.1 | 89.7 | 0.775 |

As shown in Table 6, under noisy conditions, both the precision and recall of the proposed method exhibit a slight decline. However, the overall performance degradation remains within 5%, indicating robust resilience to noise interference. In complex scenarios, the proposed method continues to maintain stable keyframe extraction performance, with an average fidelity reduction not exceeding 4.7%. In cross-category generalisation experiments, the proposed method demonstrated performance comparable to baselines across sports events and surveillance footage, indicating the distributed artificial intelligence framework possesses robust cross-domain generalisation capabilities. In summary, the proposed method exhibits commendable stability and adaptability under conditions of noise interference, complex environments and cross-category video data.

## 4 Conclusions

Distributed artificial intelligence technology is used in video key frame extraction to increase the accuracy of the process. The method's effectiveness is confirmed by both theory and experiment. The key frame extraction effect is good, and the compression rate and fidelity of this approach are great. In particular, this method's fidelity and compression rate are far higher than those of the EM based on SC. Compared with the EM based on fusion features, the key frame extraction effect of this method is higher, there is no frame leakage, and the extraction results are consistent with the reference standards.

Although the proposed distributed artificial intelligence-based video keyframe extraction method demonstrates favourable performance across multiple datasets and complex environments, certain limitations remain. These include relatively constrained experimental data scale and sources, computational efficiency requiring optimisation for large-scale or real-time scenarios, and potential performance degradation in cross-modal video or high-noise environments.

Future research will extend to more complex and diverse video datasets, exploring lightweight neural network architectures and efficient parallel computing to enhance real-time performance. Concurrently, integrating multimodal information such as audio and text will strengthen semantic understanding, while incorporating adaptive feedback mechanisms will improve robustness and generalisation across varying environmental conditions. This will unlock greater potential for applications in video content analysis and intelligent retrieval.

## Declarations

The author declares that he has no conflicts of interest.

## References

Abulizi, W. (2019) 'Keyframe feature matching simulation of 3D stereoscopic image', *Computer Simulation*, Vol. 36, No. 12, pp.186–189.

Akilan, T., Wu, Q. and Zhang, W. (2019) 'Video foreground extraction using multi-view receptive field and encoder-decoder DCNN for traffic and surveillance applications', *IEEE Transactions on Vehicular Technology*, Vol. 48, No. 22, pp.214–220.

Bommisetty, R.M., Prakash, O. and Khare, A. (2019) 'Keyframe extraction using Pearson correlation coefficient and color moments', *Multimedia Systems*, Vol. 26, No. 1, pp.267–299.

Elahi, G. and Yang, Y.H. (2021) 'Online learnable keyframe extraction in videos and its application with semantic word vector in action recognition', *Pattern Recognition*, Vol. 122, No. 2, pp.108–116.

Gu, X., Lu, L., Qu, S. et al. (2020) 'Sentiment key frame extraction in user-generated micro-videos via low-rank and sparse representation', *Neurocomputing*, Vol. 410, No. 2, pp.441–453.

Jian, M., Zhang, S., Wu, L. et al. (2019) 'Deep key frame extraction for sport training', *Neurocomputing*, 7 February, Vol. 328, No. 2, pp.147–156.

Lan, X., Zhou, M., Xu, X., Wei, X., Liao, X., Pu, H. et al. (2023) 'Multilevel feature fusion for end-to-end blind image quality assessment', *IEEE Transactions on Broadcasting*, Vol. 69, No. 3, pp.801–811.

Li, M., Ji, G. and Zhao, B. (2021) 'Key frame extraction algorithm for video-based person re-identification based on walking cycle clustering', *Journal of Nanjing University of Aeronautics & Astronautics*, Vol. 53, No. 5, pp.780–788.

Li, Y., Kanemura, A., Asoh, H. and Kawanabe, M. (2020) 'Multi-sensor integration for key-frame extraction from first-person videos', *IEEE Access*, Vol. 34, No. 9, pp.141–151.

Lingam, K.M. and Reddy, V. (2020) 'Content relative thresholding technique for key frame extraction', *International Journal of Knowledge-Based and Intelligent Engineering Systems*, Vol. 23, No. 4, pp.249–258.

Prathiba, T. and Kumari, R. (2021) 'Eagle eye CBVR based on unique key frame extraction and deep belief neural network', *Wireless Personal Communications*, Vol. 116, No. 6, pp.411–441.

Shamsipour, G., Fekri-Ershad, S., Sharifi, M. and Alaei, A. (2024) 'Improve the efficiency of handcrafted features in image retrieval by adding selected feature generating layers of deep convolutional neural networks', *Signal, Image and Video Processing*, Vol. 18, No. 3, pp.2607–2620.

Singh, Y. and Kaur, L. (2020) 'Effective key-frame extraction approach using TSTBTC-BBA', *IET Image Processing*, Vol. 14, No. 4, pp.638–647.

Wang, J. and Lu, X. (2021) 'Video key frame extraction algorithm based on semantic correlation', *Computer Engineering and Applications*, Vol. 57, No. 4, pp.192–198.

Wang, Z. and Zhu, Y. (2020) 'Video key frame monitoring algorithm and virtual reality display based on motion vector', *IEEE Access*, Vol. 65, No. 19, pp.19–25.

Yasin, H., Hussain, M. and Weber, A. (2020) 'Keys for action: an efficient keyframe-based approach for 3D action recognition using a deep neural network', *Sensors*, Vol. 20, No. 8, pp.2226–2234.

Yuan, Y., Lu, Z., Yang, Z., Jian, M., Wu, L., Li, Z. and Liu, X. (2021) 'Key frame extraction based on global motion statistics for team-sport videos', *Multimedia Systems*, Vol. 28, No. 8, pp.387–401.

Zhang, F.D., Zhao, Z.Y., Sun, R. et al. (2022) 'Cable fire warning algorithm based on artificial intelligence and multi-sensor information fusion', *Electronic Design Engineering*, Vol. 30, No. 6, pp.86–90.

Zhang, X. and Zhang, Y. (2019) 'Video keyframe extraction method based on fusion feature', *Computer Systems & Applications*, Vol. 28, No. 11, pp.176–181.

Zhao, H., Wang, W.J., Wang, T. et al. (2019) 'Key-frame extraction based on HSV histogram and adaptive clustering', *Mathematical Problems in Engineering*, Vol. 2019, No. 12, pp.1–10.

Zhong, Q., Zhang, Y., Zhang, J., Shi, K. and Liu, C. (2020) 'Key frame extraction algorithm of motion video based on priori', *IEEE Access*, Vol. 8, No. 2, pp.174424–174436.