Video Classification Based On the Improved K-Means Clustering Algorithm

Taile Peng*, Zhen Zhang*, Ke Shenb, Tao Jiangc
School of Computer Science and Technology, Huaibei Normal University, Huaibei, 235000, China

*Corresponding author e-mail: tailep@163.com, 380869815@qq.com, hbshenke@163.com, 767906953@qq.com

Abstract. To solve the problem of low accuracy of video classification, this paper proposes an improved k-Means algorithm, which is used as a classifier to realize video classification. Firstly, video segmentation is carried out to extract the visual features of the video and form a group of visual features. The traditional k-Means clustering algorithm is improved to form initial clustering points with labeled video samples, optimize the objective function, and further optimize the clustering results of the video. Several experiments show that the proposed classification algorithm has high classification accuracy.

1. Introduction
Video is a rich content of information carrier, but also one of the main ways to express the data of various information sources. At present, the Internet is full of massive video data. With the explosive growth of massive video data, a large number of video applications have been produced. Video is a kind of multi-modal data with complex structure. Most of the videos in the Internet are labeled-free video data. The classification and annotation of video is one of the key steps to solve many video applications. From the perspective of semantics, there is always a certain correlation between videos in the same field. Domain experts use various attributes of video to establish video association, which is very important for video classification, indexing, video data filtering and video retrieval applications.

Video is a multimodal multimedia data. Video classification using visual features of video is a common method of video classification. For video classification of small data sets, many literatures use one or more visual features of video to achieve video classification. For example, in document [1], Meng Li and others realized video classification method based on histogram difference method. In document [2], Song Yating and others use tensor dynamic texture model to achieve Aurora video classification. Due to the diversity of video visual features, classification using a visual feature often fails to achieve better classification results. Many documents use a variety of visual and audio features to achieve video classification, and have achieved good classification results. In document [3] Yang classifies video from visual features, semantic features, audio features and additional text, and proposes that multi-modal classification is better than single-modal classification. There are also a lot of literature from the improvement of classifier, but also improve the classification effect. For example, YUAN [4] and others use decision tree to realize video classification.
For the video classification problem of large data sets or massive data sets, deep learning is used to extract video features to achieve better classification results. For example, document [5] proposes to use Convolutional Neural Network (CNN) to learn temporal and spatial features of video on video clips. Wang et al. [6] proposed Temporal Segment Networks (TSN) to capture long-term dynamic information, and achieved good classification accuracy.

For small data sets, video classification using deep learning often fails to achieve better classification results. Video classification accuracy can be effectively improved by combining the visual and audio features of the video, which are based on the multi-modal characteristics of the video presented by the changes of scene and angle and motion speed. On the basis of extracting color features, SURF features and texture features of video, this paper classifies video from small data sets by clustering algorithm. For a given video, first extract its color features and SURF features; then improve the classifier, and finally achieve video classification.

2. General flow of video classification based on Multi-Modality

![General framework for video classification](image)

**Figure 1.** General framework for video classification

### 2.1. Algorithmic Principle of K-Means Clustering

In video classification methods for small data sets, the common video classifiers include supervised learning machines, such as SVM and various improved algorithms, semi-supervised learning machines, and unsupervised learning clustering algorithms, such as K-Means clustering algorithm. K-Means clustering algorithm is a commonly used video classification method. Its algorithm idea is as follows:

1. Firstly, k clustering results are set and initial clustering values such as $v_1^{(1)}$, $v_2^{(1)}$, $v_k^{(1)}$ are selected.

2. During the fifth iteration, each pixel is assigned to one of $k$ classes according to the distance function.

   \[
   \text{If} \| f(x,y) - \mu_j^{(i)} \| < \| f(x,y) - \mu_j^{(i)} \| \text{ each pixel to the nearest class with the mean value is assigned.}
   \]

3. For each class, update the mean of that class.

   \[
   \mu_j^{(i+1)} = \frac{1}{N_j} \sum_{(m,n) \in O_j^{(i)}} f(x,y)
   \]

   Where $N_j$ is the number of pixels in the updated class $O_j^{(i+1)}$.

   If $\mu_j^{(i+1)} = \mu_j^{(i)}$ holds for all classes, the algorithm converges and ends; otherwise, it moves to the next iteration.

   From the k-Means clustering algorithm, we can see that the initial point of clustering has a great influence on the clustering results, and in extreme cases, the wrong clustering results may be obtained.

3. Video clustering algorithm based on improved k-Means

### 3.1. Shot Segmentation of Video

Video can be divided into stories, scenes, shots, key frames and other semantic units from the perspective of semantics. Shots are an important semantic unit of video. Videos with similar categories always contain similar shots. In this paper, shot edge detection and segmentation of video are carried out by using the shot edge detection algorithm provided in document [7]. Among them, the values of
parameters are: $\delta_1 = 0.3$, $\varepsilon_1 = 0.65$, $\varepsilon_2 = 0.4$, $r = 20$. The key frame selection is realized by using the method of reference [8]

### 3.2. Video visual feature extraction

#### 3.2.1. Color feature extraction

In color space, RGB color space is the most commonly used color space model, but HSV color feature has the best resolution. HSV is the closest color model to human subjective perception. HSV can better reflect the color distribution of the image. Color histograms and color moments are both ways of expressing color features, but they cannot express the spatial position of image color. In this paper, the color aggregation vectors proposed by PASS in document [9] are used to represent color features in HSV space. The color aggregation vector of the lens is obtained by taking the lens as a unit.

#### 3.2.2. SURF feature extraction

SURF is a local feature descriptor, which is insensitive to brightness change and rotation and has strong robustness. In this paper, the SURF features of frame images are extracted based on the method of reference [11].

SURF is an optimization acceleration of SIFT. Its process can be summarized as the following steps:

- **Step1:** Construct Hessian matrix and generate all interest points for feature extraction.
- **Step2:** Constructing scale space;
- **Step3:** Locate feature points accurately;
- **Step4:** Determine the main direction of feature points;
- **Step5:** Generate 64-dimensional feature point descriptors;
- **Step6:** Feature point matching

In SURF algorithm, 4*4 rectangular blocks around feature points are extracted, and the direction of rectangular area is the main direction of feature points. Each sub-region counts the Haar wavelet features of 25 pixels in horizontal and vertical directions. The Harr wavelet is characterized by the sum of the values in the horizontal direction, the sum of the values in the vertical direction, the sum of the absolute values in the horizontal direction and the sum of the absolute values in the vertical direction. With these four values as the feature vectors of each sub-block region, a 64-dimensional feature descriptor is formed. The SURF descriptor subset $S$, $S = \{S_1, S_2... S_n\}$.

#### 3.2.3. Texture feature extraction

The texture of an image is used to describe the spatial color distribution and light intensity distribution of an image or a local block area of an image. Gray level co-occurrence matrix (GLCM) is a feature extraction algorithm proposed by Harklick [10] et al in 1979. It is one of the algorithms used to describe global texture features of images.

Of the 14 texture features of GLCM, only 4 features (angular second moment, contrast, inverse moment, correlation) are irrelevant. Four features, correlation, contrast, energy and inverse gap, are taken as texture feature vectors of frame images.

### 3.3. Step of video clustering algorithm based on improved K-Means

For a given video No. $i$, the color feature $C_i$, SURF feature $S_i$ and texture feature $W_i$ are extracted. Data set $Data = \{x_j\}, i = 1, 2, ... m$, $x_j \in R^{C_i+S_i+W_i}$, initial clustering number $k$, label sample set $S = S_1 \cup S_2 \cdots \cup S_k$, $S_i \in R^{C_i+S_i+W_i}$ can be obtained

Result Data Set: N clustering results $S^{(c)} = \{x_1^{(c)} \cup x_2^{(c)} \cdots \cup x_N^{(c)}\}$ of Data set optimize the objective function.
(1) Initialization. The labeled samples are "seed points" of known clustering attributes, and many sub-points can form clustering initial points more accurately. For labeled samples (seed points), k initial clustering centers such as \( \{ \mu_h^{(0)} \} , h = 1, \cdots, k \) can be obtained by min-max algorithm.

(2) Repeat the following steps until final convergence

a) Distribution Clustering: Distribution of data points \( x_i \) to any class \( h^* \), so that \( h^* \) satisfies the optimal and sub-optimal, and obtains candidate classes.

\[
\hat{h}^* = \arg_{h} \min \| x_j - \mu^{(t)}_h \|^2 
\]  

(1)

B) For the candidate class \( h^{**} \), the following conditions are satisfied

\[
h^{**} = \arg_{h} \min \left( \| x_j - \mu^{(t)}_h \|^2 \times \sum_{j=1}^{n} \| x_j - x_j \|^2 / m \right)
\]  

(2)

In formula (2), the Euclidean distance of data points adjacent to \( x_i \) in class \( h^{**} \) is satisfied.

\[
\| x_i - x_j \|^2 \leq \arg_{h} \min \| x_i - \mu^{(t)}_h \|^2
\]  

(3)

3) Recalculate the central point of each cluster

\[
\mu^{(t+1)}_h = \frac{1}{\mid S^{(t+1)}_h \mid} \sum_{x \in S^{(t+1)}_h} X
\]  

(4)

4) Modify the iteration conditions:

\[
t \leftarrow t + 1
\]  

(5)

(3) A preliminary N-class partition of data set X is obtained.

(4) According to the result of N-class partition, if the number of cluster points \( \sum X \) in Class I is less than \( \varepsilon \), \( \varepsilon \) is empirical, and cluster I is invalid, and the data points in Class I are noise or outliers, delete them; otherwise, retain Class I, and get the N’ class partition of Data Set X.

4. Experiments and analysis

4.1. Selection and evaluation criteria of data sets in this paper

In order to validate the effectiveness of this algorithm, we select the Haberman sub-library, German Credit Data sub-library, Heart sub-library and data selected from the UCI (UC Irvine Machine Learning Repository) database of the University of California as data sources to verify the clustering effect of this algorithm. The attributes of the data set are shown in the table. 1 shows
Table 1. Selected Data Set Attribute Table

|      | Haberman      | German Credit Data | Heart |
|------|---------------|--------------------|-------|
| Instances | 306           | 1500               | 2500  |
| Attributes | 3            | 20                 | 13    |

This paper verifies the effectiveness of the algorithm from three aspects: classification accuracy (AC), classification accuracy (PE), recall rate (RE).

\[
AC = \frac{\sum_{i=1}^{k} p_i}{n} \tag{6}
\]

\[
PE = \frac{\sum_{i=1}^{k} p_i - q_i}{k} \tag{7}
\]

\[
RE = \frac{\sum_{i=1}^{k} p_i - r_i}{k} \tag{8}
\]

In formula (6), \( n \) denotes the number of samples in the data set, \( P_i \) denotes the number of samples correctly assigned to class \( i \), \( q_i \) denotes the number of samples mistakenly assigned to class \( i \), \( r_i \) denotes the number of samples that should be assigned to class \( i \) but not, and \( k \) denotes the number of clusters.

4.2. Comparisons between the proposed algorithm and related algorithms

In order to verify the effectiveness of the clustering algorithm in this paper, classical clustering algorithms \textit{k-Means}, \textit{Seeded k-Means} and \textit{COP-k-Means} are selected and compared under the same data set (UCI data set). The experimental results are shown in Table 2-Table 4.

Table 2. Experimental results of correlation algorithms on Haberman sets

![Graph showing experimental results of correlation algorithms on Haberman sets]
From Table 2- Table 4, it can be seen that the proposed algorithm is superior to k-means algorithm, Seeded k-Means algorithm and COP-k-Means algorithm in classification accuracy (AC), classification accuracy (PE) and recall rate (RE). K-means algorithm is a kind of unsupervised learning clustering, which cannot reasonably use labeled samples to guide the generation of clustering centers, and the selection of initial clustering points is random. Compared with the other four algorithms, the clustering effect of K-Means algorithm and COP-K-Means algorithm is the worst. Seeded K-Means algorithm and COP-K-Means algorithm use a small number of bands. Label samples are used to guide the initial clustering centers, and the uniform distribution of label samples is required to achieve better clustering effect. This algorithm also obtains the initial clustering points from the label samples. This algorithm does not need the uniform distribution of label samples. The clustering of data points not only considers the data points and the clustering centers. Distance, taking into account the density of local data distribution, makes the clustering results better reflect the characteristics of low coupling between classes and high aggregation within classes, and can also eliminate noise points. Compared with the other four algorithms, the proposed algorithm achieves better clustering results.

5. Conclusion
To solve the problem of video classification, an improved K-Means clustering video classification method is proposed. In this paper, a shot is represented by multi-visual features of video, and a video is further represented by multi-visual features. Compared with identifying video with single visual features, multi-visual features can identify video semantics more accurately. Using label samples to form the initial clustering center of K-Means algorithm can help to form a better clustering effect. In the future research work, we will further study the learning machine to achieve the purpose of accurate video classification.
Acknowledgments
This paper is jointly supported by Key Natural Science Research Projects in in Colleges of Anhui China (grant numbers KJ2016A630 and KJ2017A843).

References
[1] MENG Li, XU Fa-sheng, LI Jin-ping. Video Classification Method Based on Histogram Difference. JOURNAL OF UNIVERSITY OF JINAN (Sci.&Tech.), 2007, 21(2):100-103.
[2] Song Yating Han Bing Gao Xinbo. Tensor based dynamic textures model for aurora sequences classification. JOURNAL OF NANO. JING UNIVERSITY (NATURAL SCIENCES), 2016, 52(1):184-193.
[3] YANG L, LIU J, YANG X, et al. Multi-modality Web video categorization [C] // Proceedings of the 2007 International Workshop on Multimedia Information Retrieval. New York: ACM, 2007, pp. 265-274.
[4] Yuan Y, Song Q B, Shen J Y. Automatic Video Classification Using Decision Tree Method [C]. In Proceedings of ICMLC, 2002, pp.1153-1157.
[5] JI Shuiwang, XU Wei, YANG Ming, et al. 3D convolutional neural networks for human action recognition [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(1): 221–231.
[6] WANG Limin, XIONG Yuanjun, WANG Zhe, et al. Temporal segment networks: Towards good practices for deep action recognition [J]. ACM Transactions on Information Systems, 2016, 22(1): 20–36.
[7] Peng Taile 1, 2, Zhang Wenjun3, Wang Youbao 3, Huang Dongjin . Video shot boundary detection algorithm based on multi-features. Chinese Journal of Scientific Instrument, 2016, 22(1): 20–36.
[8] Hu Zhengping, Tu Xiaolei. Scene Classification with Multi-direction Context Features and Spatial Pyramid Model. SIGNAL PROCESSING, 2011, 27(10):1536-1542.
[9] Pass, Greg, Ramin Zabih, and Justin Miller. Comparing Images Using Color Coherence Vectors. Paper presented at the Proceedings of the fourth ACM international conference on Multimedia, 1997.
[10] Haralick R M. Statistical and structural approaches to texture [J]. Proceedings of the IEEE, 1979, 67(5):786-804.