Expert Systems With Applications 246 (2024) 123099

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Enhanced video clustering using multiple riemannian manifold-valued descriptors and audio-visual information

Wenbo Hu [a,b], Hongjian Zhan [a,b,c,*], Yinghong Tian [a], Yujie Xiong [d], Yue Lu [a,b]

[a] School of Communication and Electronic Engineering, East China Normal University, Shanghai 200062, China
[b] Shanghai Key Laboratory of Multidimensional Information Processing, Shanghai 200241, China
[c] Chongqing Institute of East China Normal University, Chongqing 401120, China
[d] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China

ARTICLE INFO

Keywords:
Riemannian manifolds
Audio-visual
Subspace clustering

ABSTRACT

Videos inherently blend multiple modalities in real-world scenarios, primarily visual and auditory cues. When synergized, these cues foster enhanced data representations. Standard clustering techniques, primarily designed for managing vectorial data in Euclidean spaces, struggle to handle multidimensional data with nonlinear manifold structures, such as video or image sets. While recent subspace clustering methods using Riemannian manifold representation tackle this issue, they often sideline auditory information, overlooking the potential harmony between visual and auditory modalities. This paper presents an innovative approach that crafts multiple Riemannian manifold-valued descriptors to bridge this gap, encapsulating multimodal video information in a unified structure. We architect a single-modality Riemannian subspace clustering for individual modal data and extend it to a multi-modality framework, leveraging the interplay of audio-visual data. Detailed optimization and convergence analysis are also provided. The proposed approach significantly outperforms the existing state-of-the-art methods, improving accuracy by 4%, 1%, and 2% on UCF-101, UCF-sport, and AVE datasets, respectively.

1. Introduction

Real-world scenarios often encompass multiple modalities such as audio and visual components that are highly interconnected and facilitate more accurate semantic information prediction (Kudithipudi, Aguilar-Simon, Babb, et al., 2022; Tan, Zhou, Tao, et al., 2021). The human brain's multisensory integration mechanisms have evolved to adeptly process and amalgamate information from various modalities. For example, nerve cells in the human brain's superior temporal sulcus respond concurrently to visual, audio, and tactile signals, highlighting the significance of auditory perception in decoding actions and dynamic occurrences in the visual realm (Scheliga, Kellermann, Lampert, et al., 2023). The synergistic relationship between visual and audio information enables interaction, leading to more precise perceptual signals.

With these considerations in the vanguard, our study seeks to address the following research question: How can the integration of cross-modal information, specifically audio and visual data, enhance the reliability and accuracy of video clustering?

In a variety of applications, particularly those encompassing tasks of video clustering with intricate manifold structures, conventional clustering methods often fall short due to their inability to efficiently navigate non-linear data spaces. Over the years, self-expressive clustering methodologies have garnered considerable interest due to their superior performance (Elhamifar & Vidal, 2013; Hu & Wu, 2020; Liu, Hu, Wang, et al., 2023; Liu et al., 2012; Wang, Wu, Ren, et al., 2023). These self-expressive techniques can be perceived as a specialized form of dictionary learning method, *i.e.*, utilizing the data itself as a dictionary. Notable algorithms include Sparse Subspace Clustering (SSC) (Elhamifar & Vidal, 2013) and Low Rank Representation (LRR) (Liu et al., 2012). For nonlinear manifold spaces where image sets or videos are thought to exist, these clustering methods using Euclidean distance may not deliver consistent performance.

Considering the limitations of existing clustering methods and the unique structure of videos, the concept of Riemannian manifold representation emerges as a promising avenue. Such a representation can capture the intrinsic geometries of data, making it particularly suitable

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

[*] Corresponding author at: School of Communication and Electronic Engineering, East China Normal University, Shanghai 200062, China.

E-mail addresses: wenbo@stu.ecnu.edu.cn (W. Hu), hjzhan@cee.ecnu.edu.cn (H. Zhan), yhtian@cee.ecnu.edu.cn (Y. Tian), xiong@sues.edu.cn (Y. Xiong), ylu@cee.ecnu.edu.cn (Y. Lu).

https://doi.org/10.1016/j.eswa.2023.123099
Received 4 July 2023; Received in revised form 28 December 2023; Accepted 28 December 2023
Available online 4 January 2024
0957-4174/© 2023 Elsevier Ltd. All rights reserved.

for video or image set clustering. Several researchers (Hu & Xu, 2022; Piao, Hu, Gao, et al., 2019; Shirazi, Harandi, Sanderson, et al., 2012; Wang, Hu, Gao, Sun, & Yin, 2014) have examined the method of Riemannian manifold representation as the basis for such tasks. These methods leverage the Riemannian manifold representation to process video or image set data, achieving superior clustering performance compared to traditional methods, such as SSC (Elhamifar & Vidal, 2013) and LRR (Liu et al., 2012). However, these schemes (Hu & Xu, 2022; Piao et al., 2019; Shirazi et al., 2012; Wang et al., 2014) often fail to account for the differences between consecutive video frames, neglecting to capture crucial video motion information. Moreover, they frequently disregard audio information, relying solely on video frames, leading to sub-optimal clustering performance.

Analogous to human perception, a more refined model could be envisaged by harnessing complementary information from disparate modalities. In recent explorations, a plethora of works have ventured into audio-visual joint learning, Hu et al. (2020), Pham et al. (2022), Wang, Mesaros, Heittola, and Virtanen (2021) establish the feature representation of audio-visual input signals, followed by classification. In Kazakos, Nagrani, Zisserman, and Damen (2019), Morgado, Vasconcelos, and Misra (2021), Owens and Efros (2018), the recognition performance of the audio-visual models is augmented in a self-supervised manner employing multi-modal information. Further, video sequences usually contain ample motion information, showcasing temporal correlation between diverse events and object movements. The optical flow feature can be considered visual information that records the temporal correlation of the video (Kazakos et al., 2019; Song, Sun, & Li, 2022).

To address the complexities of video data, and as suggested by Hu et al. (2020), Morgado et al. (2021), Wang, Mesaros, et al. (2021) that leveraging multi-modal information can enhance performance, we initially constructe multiple Riemannian manifold-valued descriptors. These descriptors encapsulate key video components: the visual data provides a direct representation of scenes, optical flow captures motion dynamics, and audio signals enrich the context with auditory cues. Subsequently, we propose a single-modality Riemannian subspace clustering method, which can choose any single modality from the multiple Riemannian manifold-valued descriptors to execute the video's clustering task. Lastly, we introduce a novel multi-modality Riemannian subspace clustering method, built on the assumption of cross-modal information fusion, that can concurrently amalgamate information from multiple modalities. A series of comprehensive experiments on several datasets validate the superior performance of our proposed methods over other state-of-the-art methods for handling video clustering tasks.

The primary contributions of this work are encapsulated as follows:

(1) The introduction of multiple Riemannian manifold-valued descriptors capable of representing multimodal information in videos within a unified framework on Riemannian manifolds.

(2) Proposal of a single-modality Riemannian subspace clustering method to undertake the video clustering task using single-modal information extracted from the video, accompanied by corresponding optimization schemes and performance assessments.

(3) The suggestion of a novel multi-modality Riemannian subspace clustering method that harnesses audio-visual information to bolster clustering performance. This method aims to study the complementary attributes of different Riemannian manifold-valued descriptors to derive a more accurate and reliable representation, with associated optimization methods and convergence proofs furnished.

(4) Extensive experiments on benchmark datasets attest to the effectiveness of our methodologies. An ablation study and parameter analysis confirm the robustness and efficacy of the proposed methods.

The rest of this paper is organized as follows. The preliminaries and related works are presented in Section 2. The specifics of the proposed methods, including the optimized algorithms, are introduced in Section 3. In Section 4, we relay the experiment results and conduct experimental and statistical analysis. Conclusions are drawn in Section 5.

2. Related works

2.1. Subspace clustering

Low Rank Representation (LRR). Let us start by briefly introducing the related subspace clustering methods in the Euclidean space. Given a data set $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{d\times n}$, where d symbolizes the dimension of feature space and n signifies the number of data samples, the LRR model (Liu et al., 2012) can be formulated as follows:

$$\min_{C} \lambda \|C\|_* + \|X - XC\|_F^2, \tag{1}$$

where $C = [c_1, c_2, \dots, c_n] \in \mathbb{R}^{n\times n}$ is a coefficient matrix, with column c_i corresponding to the low rank representation of x_i. Moreover, the matrix C conveys the subspace structure information about the original data. After obtaining the coefficient matrix C, a similarity graph $W = (|C| + |C|^T)/2$ can be constructed. Then, the spectral clustering algorithm (such as Normalized Cut) (Shi & Malik, 2000; Song, Yao, Nie, et al., 2021) is used to divide the similarity graph W to obtain the final data segmentation.

Subspace Clustering on Grassmann Manifolds. To address the clustering problem of high-dimensional nonlinear Grassmann manifolds data, researchers have designed several subspace clustering methods on Grassmann manifolds. For a given data set $\mathcal{X} = [X_1, X_2, \dots, X_n]$ where $X_i \in \mathcal{G}(p,d)$ and n represents the number of samples, Wang et al. (2014) have formulated the LRR method on Grassmann manifolds and proposed a low rank representation model on Grassmann manifolds (G-LRR):

$$\min_{C} \lambda \|C\|_* + \sum_{i=1}^{n} \|X_i \ominus \biguplus_{j=1}^{n} X_j \odot c_{ij}\|_{\mathcal{G}}, \tag{2}$$

where $\|X_i \ominus \biguplus_{j=1}^{n} X_j \odot c_{ij}\|_{\mathcal{G}}$ represents the manifold measurement, and $(\biguplus_{j=1}^{n} X_j \odot c_{ij})$ denotes the“linear combination” of sample points on all Grassmann manifolds. The symbols $\ominus, \biguplus, \odot$ represent subtraction, summation, and multiplication operations on the Grassmann manifold, respectively.

To seek the sparse expression on the previously learned latent representation, Wang, Hu, Gao, et al. (2018) proposed a cascaded low-rank and sparse representation on Grassmann manifolds (G-CLRSR):

$$\min_{C,Z} \lambda_1 \|C\|_* + \lambda_2 \|Z\|_1 + \lambda_3 \|C - CZ\|_F^2 + \sum_{i=1}^{n} \|X_i \ominus \biguplus_{j=1}^{n} X_j \odot c_{ij}\|_{\mathcal{G}}, \tag{3}$$

where λ_1, λ_2 and λ_3 are three balancing parameters, and the coefficient $Z, C \in \mathbb{R}^{n\times n}$.

Similarly to G-LRR and G-CLRSR, Piao et al. (2019) formulated a double nuclear norm-based low-rank representation model on Grassmann manifolds (G-DNLR):

$$\begin{aligned} \min_{C,A,B} \lambda(\|A\|_* + \|B\|_*) + \sum_{i=1}^{n} \|X_i \ominus \biguplus_{j=1}^{n} X_j \odot c_{ij}\|_{\mathcal{G}}, \\ s.t. \quad C = AB, \end{aligned} \tag{4}$$

where $\|A\|_* + \|B\|_*$ is the double nuclear norm for C, $A \in \mathbb{R}^{n\times r}$, $B \in \mathbb{R}^{r\times n}$, and $r \le n$ is the expected rank of coefficient matrix C.

Recently, for a set of Grassmann samples $\mathcal{X} = [X_1, X_2, \dots, X_n]$, Hu and Xu (2022) proposed a one-step sparse clustering framework on Grassmann manifolds (G-OKSC):

$$\min_{C} \|\phi(\mathcal{X})\|_* + \lambda_1 \|C\|_1 + \frac{\lambda_2}{2} \sum_{i=1}^{n} \|\phi(X_i) - \sum_{j=1}^{n} \phi(X_j) c_{ij}\|_F^2, \tag{5}$$

where $\phi(\cdot)$ is expressed as the operation which can map data on the Grassmann manifold to Reproducing Kernel Hilbert Space (RKHS). Exploiting the kernel method can maintain global information and produce more discriminative data distribution.

These methods are predominantly utilized for image set clustering issues, yet they are also applicable to video clustering tasks. They exclusively capitalize on the visual information in videos, neglecting the rich audio information that is intrinsically present.

2.2. Audio-visual joint representation

In videos, two inherent modalities, visual and audio streams, are naturally presented. The audio modality, accompanying the visual, encapsulates complementary information that can significantly augment the representational capacity of video features (Tagliasacchi, Gfeller, de Chaumont Quitry, & Roblek, 2020). In recent times, the realm of audio-visual representation learning has garnered increasing attention in an endeavor to fully harness audio information (Akbari et al., 2021; Alwassel et al., 2020; Gao, Oh, Grauman, & Torresani, 2020). Particularly in dynamic audio-visual videos, the audio carries a plethora of complementary information to the RGB image sequences, which can ameliorate challenges associated with distinguishing video samples based on appearance alone, such as differentiating between playing a violin and a guitar.

Presently, a predominant focus has been directed toward employing both visual and auditory information through supervised learning approaches. For instance, studies denoted as Kazakos et al. (2019), Morgado et al. (2021), Owens and Efros (2018) have demonstrated an enhancement in the recognition performance of audio-visual models through a self-supervised manner, utilizing multi-modal information from each modality. Arandjelovic and Zisserman (2017) trained an audio-visual model to ascertain whether image and audio segments correspond with one another. Furthermore, the advantages of the audio modality have been validated in tasks such as video understanding (Morgado, Li, & Nvasconcelos, 2020; Senocak, Kim, Oh, et al., 2023). Contrary to these methods, this paper embarks on a venture of employing both visual and auditory information for video clustering tasks, an unsupervised approach.

3. Proposed method

In this section, we first introduce multiple Riemannian manifold-valued descriptors capable for describing video data, then introduce single-modality Riemannian subspace clustering methods, and finally introduce the extension of the single-modality method to multimodal, which can be capable of multiple Riemannian manifold-valued descriptors for audio-visual subspace clustering. The detailed algorithmic process is illustrated in Fig. 1. The features from the video frame and optical flow branches are represented as points on the Grassmann manifold, while the audio signal branch features are expressed as points on the SPD manifold. All three manifold-valued descriptors are mapped into a Hilbert space using the kernel method. The fusion of information across the branches is facilitated by applying adaptive weights to each coefficient matrix to construct a shared matrix. This shared matrix is then used in the spectral clustering method to obtain the final clustering results.

3.1. Riemannian manifold-valued descriptors for multiple modalities

Our approach utilizes three distinct branches: (1) Video frame branch, (2) Optical flow branch, and (3) Audio signal branch, each engineered to extract specific features from video data. The Video Frame Branch seizes static visual information, the Optical Flow Branch captures motion patterns, and the Audio Signal Branch extracts auditory information.

3.1.1. Video frame branch

Similar to the approach in Hu and Xu (2022), Piao et al. (2019), Shirazi et al. (2012), Wang et al. (2014, 2018), we perform an SVD decomposition of each video F_i in a video dataset $\mathcal{F} = [F_1, F_2, \dots, F_n]$. The decomposition of $F_i = U\Sigma V^T$ allows us to represent each video F_i as a point on the Grassmann manifold $\mathcal{G}(p, d)$, denoted as $X_i = [u_1, u_2, \dots, u_p]$, by extracting the first $p(p \leq P)$ singular vectors of U. This results in a set of corresponding samples $\mathcal{X} = [X_1, X_2, \dots, X_n]$ on the Grassmann manifold for the video dataset $\mathcal{F}$.

3.1.2. Optical flow branch

Optical flow, which captures motion patterns in video data, is obtained for each video F_i using the LK optical flow method (Baker & Matthews, 2004). Similar to the video frame branch, these features are also represented as points on the Grassmann manifold. Therefore, given an optical flow feature gallery $\mathcal{O} = [O_1, O_2, \dots, O_n]$, we construct the corresponding samples $\mathcal{Y} = [Y_1, Y_2, \dots, Y_n]$ on the Grassmann manifold.

3.1.3. Audio signal branch

Auditory information is typically segmented into specific time windows, with features extracted for each segment independently. Recent research has explored a variety of feature extraction techniques for auditory information. As adopted in some recent works (Albadr et al., 2021, 2022; Yang, Marković, Krenn, et al., 2022), include Mel-frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), Linear Predictive Coding (LPC) (Chauhan, Isshiki, & Li, 2019; Li, Parsan, Wang, et al., 2023; Vafeiadis, Votis, Giakoumis, et al., 2020), MPEG-7 (Muhammad & Alghathbar, 2009), log-Gabor filters (Souli & Lachiri, 2012), and spectrogram. Among these techniques, the spectrogram (Ghandoura, Hjabo, & Al Dakkak, 2021; Yang et al., 2022), which represents each audio signal as a time-frequency image, has garnered particular attention due to its superior experimental performance. Based on this, similar to Chen and Huang (2021), Ghandoura et al. (2021), Yang et al. (2022), we apply the spectrogram function in MATLAB to represent each audio signal as a time-frequency image. The spectrogram's texture information is further processed using the Log-Gabor filter (Field, 1987; Souli & Lachiri, 2012) to extract more discriminative information. Unlike a traditional Gabor filter, the Log-Gabor filter better represents the high-frequency portion and avoids overemphasis of the low-frequency part of the image. For each spectrogram, we compute a corresponding covariance matrix S_i:

$$S_i = \frac{1}{g_i - 1} \sum_{j=1}^{g_i} (G_{ij} - \overline{P}_i)(G_{ij} - \overline{P}_i)^T, \tag{6}$$

where $G_i = [G_{ij}, j = 1, \dots, k_i]$, and k_i is the number of Gabor features, $\overline{P}_i = \frac{1}{g_i} \sum_{j=1}^{g_i} G_{ij}$ is denoted as the mean matrix. The diagonal terms of the covariance matrix reflect the variance of the individual terms of the eigenvector, and the non-diagonal terms of the covariance matrix reflect the correlation between the individual terms of the eigenvector. By capturing the distribution characteristics of the data, the covariance matrix, considered as a point on the SPD manifold, provides a potent representation capability.

3.2. Single-modality Riemannian subspace clustering

In this section, we present the Riemannian subspace clustering for single-modality, hereafter referred to as Single-modality Riemannian Subspace Clustering (SRSC). The features extracted from different branches encapsulate the video from distinct perspectives. Employing three diverse branch strategies, we represent the ith video as the tuple (X_i, Y_i, S_i). It is noteworthy that X_i and Y_i denote points on the Grassmann manifold, whereas S_i, derived from the audio signal branch, resides on the SPD manifold.

Given that Riemannian manifolds lack a vector space structure, the traditional subspace clustering algorithms predicated on Euclidean space are unsuitable for direct application to Riemannian manifolds. Moreover, the subspace clustering algorithms rooted in the self-expressive property implicate numerous linear operations. While considering the fundamental structure of Riemannian manifolds, we propose three solutions to overcome these challenges in subspace clustering on Riemannian manifolds:

(1) Direct construction of a self-expressive-based subspace clustering model within the original Riemannian manifold space. However, the optimization of the solution in the Riemannian manifold space can be non-trivial.

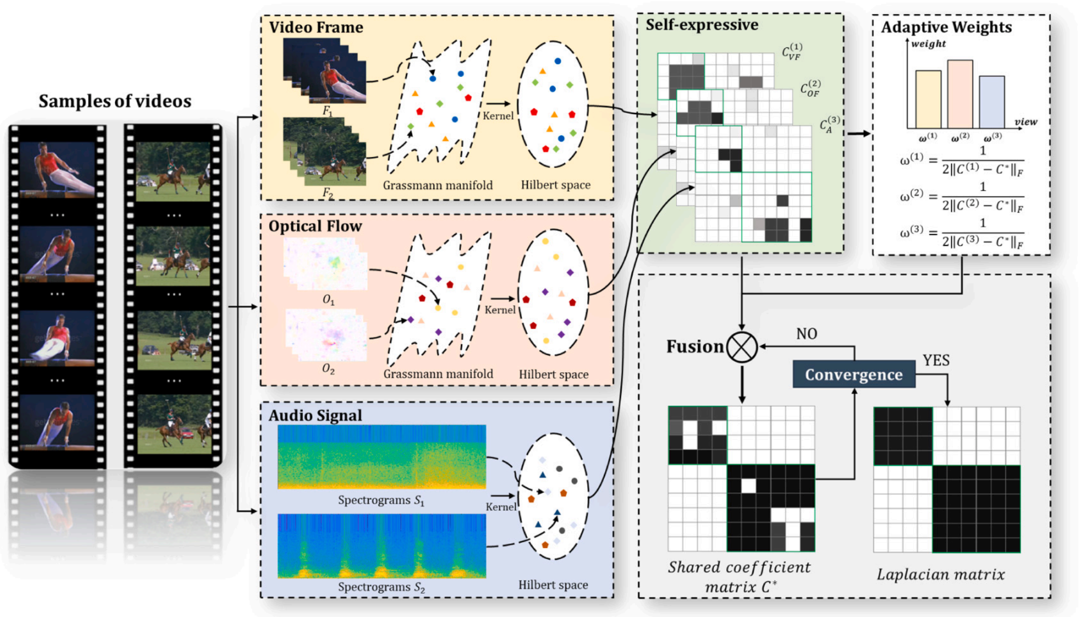


Fig. 1. The workflow of the proposed multi-modality Riemannian subspace clustering method.

(2) Local flattening of Riemannian manifolds via the tangent space, which transmutes the non-Euclidean geometry into a Euclidean structure. However, this approach, only preserving the local structure of Riemannian manifold data via mapping, may neglect the global information and thus can result in sub-optimal clustering performance.

(3) Kernel methods in Euclidean space, which effectively explore data nonlinearity (Ali, Yaseen, Aljanabi, et al., 2023; Wang, Liu, Liu, et al., 2021; Zhang, Kang, Xu, et al., 2022). These methods work on the principle of mapping the data to a high-dimensional feature space, thus providing a more comprehensive representation of data distribution. In addition, kernel functions can be utilized to directly compute the inner product of data in high-dimensional space, bypassing the need to analyze the specific form of the data in such spaces.

As the Reproducing Kernel Hilbert Space (RKHS) is a complete inner product space, the mapping process transforms the nonlinear manifold into a linear space, thus facilitating the resolution of linear operations based on self-expressiveness. Given a data set $\mathcal{R} = [R_1, R_2, \dots, R_n]$ on the Riemannian manifold, where $\mathcal{R}$ can be represented as one of the Riemannian manifold-valued descriptors from the three branches, we define the single-modality Riemannian subspace clustering based on Hilbert kernel space embedding as follows:

$$\min_{C} \lambda \|C\|_1 + \sum_{i=1}^{n} \|\phi(R_i) - \sum_{j=1}^{n} \phi(R_j) c_{ij}\|_F^2, \quad s.t. \quad diag(C) = 0, \tag{7}$$

where $\phi(\cdot)$ denotes a feature mapping function that projects the Riemannian manifold data into RKHS, $C = (c_{ij})_{i,j=1}^{n}$ represents the self-expression coefficient matrix, and λ is a trade-off parameter.

3.2.1. Optimization of SRSC

By applying the kernel trick, Eq. (7) is transformed into the following equivalent problem:

$$\min_{C} \lambda \|C\|_1 + tr(K - 2KC + C^T KC), \quad s.t. \quad diag(C) = 0, \tag{8}$$

where K is the kernel Gram matrix $K = (k_{ij})_{i,j=1}^{n}$. By introducing auxiliary variable A, the above problem can be reformulated as follows:

$$\min_{C,A} \lambda \|C\|_1 + tr(K - 2KA + A^T KA), \quad s.t. \quad A = C - diag(C). \tag{9}$$

It can be solved by alternating direction method of multipliers (ADMM) (Boyd, Parikh, Chu, et al., 2011). The augmented Lagrangian function is given by:

$$\begin{aligned} \mathcal{L} &= (C, A, \Delta) \\ &= \lambda \|C\|_1 + tr(K - 2KA + A^T KA) \\ &+ \frac{\mu}{2} \|A - C + diag(C)\|_F^2 + tr[\Delta^T (A - C + diag(C))], \end{aligned} \tag{10}$$

where Δ is a Lagrange multiplier, and $\mu \geq 0$ is the penalty parameter. We update each of the above variables by minimizing Eq. (10) while keeping the other variables fixed.

Updating A. To update A, we solve the following sub-problem by fixing the other variables.

$$\begin{aligned} &\min_{C,A} - 2tr(KA) + tr(A^T KA) + \frac{\mu}{2} \|A - C + diag(C)\|_F^2 \\ &+ tr[\Delta^T (A - C + diag(C))]. \end{aligned} \tag{11}$$

This is a quadratic optimization problem for A. We set Eq. (11) derivative *w.r.t.* A to zero and get the closed-form solution as:

$$A = (K + \mu I)^{-1} \times (K + \mu C - \Delta), \tag{12}$$

where I is an identity matrix.

Updating C. The update of C can be achieved by solving the following sub-problem (Daubechies, Defrise, & De Mol, 2004; Donoho, 1995):

$$\begin{aligned} &\min_{C} \lambda \|C\|_1 + \frac{\mu}{2} \|A - C + diag(C)\|_F^2 \\ &+ tr[\Delta^T (A - C + diag(C))], \end{aligned} \tag{13}$$

this sub-problem has a closed-form solution given by:

$$\begin{aligned} C &= \Lambda - diag(\Lambda), \\ \Lambda &= T_{\frac{\lambda}{\mu}} (A + \frac{\Delta}{\mu}), \end{aligned} \tag{14}$$

where $T_{\Upsilon}(\cdot)$ is an element-wise soft-thresholding operator that is defined as $T_{\Upsilon}(x) = sign(x) \cdot max(|x| - \Upsilon, 0)$.

Updating Δ. The Lagrangian multiplier is updated according to the following equation:

$$\Delta = \Delta + \mu(A - C + diag(C)). \tag{15}$$

These updating steps are repeated until satisfying the convergence condition or exceed the maximal number of iterations. After obtaining

Algorithm 1 SRSC

```
Input: $\mathcal{R} = [R_1, R_2, ..., R_n]$, $\lambda > 0$, $\mu > 0$, $\mu_{max} = 10^6$, $\rho = 1.5$
Initialize: $A = 0$, $C = 0$, $\Delta = 0$
while not converged do
    Update $A$ according to Eq. (12).
    Update $C$ according to Eq. (14).
    Update $\Delta$ according to Eq. (15).
    Update the penalty variable $\mu := min(\rho\mu, \mu_{max})$;
end while
Apply the spectral clustering algorithm to the affinity matrix $W = (|C| + |C|^T)/2$;
Output: Assignment of the data points to $k$ clusters.
```

the coefficient matrix C, the next step of the algorithm is to find the corresponding clusters. Here, we apply the spectral clustering method to the affinity matrix given by $W = (|C| + |C|^T)/2$ to produce the final clustering results. Algorithm 1 summarizes the steps of SRSC.

3.3. Multi-modality Riemannian subspace clustering

This section introduces the method for multi-modality Riemannian subspace clustering, henceforth referred to as M-AVSC, which leverages both audio and visual information to perform subspace clustering. For multiple Riemannian manifold-valued descriptors, (X_i, Y_i, Z_i) represent heterogeneous data distributed across distinct manifolds. The kernel method facilitates the embedding of this data into Hilbert space, thereby addressing not only the nonlinearity issue inherent to manifold data but also offering a comprehensive, high-dimensional feature representation.

To provide a clearer representation, let $\mathcal{R}^{(v)} = [R_1^{(v)}, R_2^{(v)}, \dots, R_n^{(v)}]$ symbolize the features of v branches, where $R_i^{(v)}$ denotes the features of the vth branch from the ith video. We express the new feature of the vth branch feature as ϕ_v. Although ϕ_v is typically implicit in kernel methods, for simplicity, we employ it as an explicit feature vector.

The choice of an appropriate kernel function is a pivotal aspect of kernel methods. Particularly for unsupervised cases, the appropriate kernel cannot be selected through a validation set, and we must rely on prior knowledge pertaining to the subspace clustering problem to define K using a simple kernel. For the Riemannian manifold-valued descriptors $\mathcal{R}^{(v)} = [R_1^{(v)}, R_2^{(v)}, \dots, R_n^{(v)}]$, we can acquire the matrix of each coefficient by solving the following equation:

$$\begin{aligned} \min_{\{C^{(v)}\}_{v=1}^{b}} & \sum_{v=1}^{b} (\sum_{i=1}^{n} \| \phi_v(R_i^{(v)}) - \sum_{j=1}^{n} \phi_v(R_j^{(v)}) c_{ij} \|_F^2 + \lambda \| C^{(v)} \|_1), \\ & s.t. \quad diag(C^{(v)}) = 0, v = 1, \dots, b, \end{aligned} \tag{16}$$

where $\phi_v(\cdot)$ signifies the kernel mapping function of each branch feature, b denotes the number of branches, $C^{(v)}$ represents the coefficient matrix for the vth branch feature, $\|C^{(v)}\|_1$ is the sparse constraint term of $C^{(v)}$, and λ is the trade-off parameter. Upon application of Eq. (16), we independently secure the self-expression coefficient matrix for each descriptor. In contrast to the single-modality scenario, the multi-modality setting encompasses multiple coefficient matrices. To discern the diverse contributions made by different Riemannian manifold-valued descriptors, we seek to obtain a shared coefficient matrix C^* predicated on multiple descriptors. The fusion method for the multiple coefficient matrix $C^{(v)}$ is underpinned by two intuitive assumptions:

(1) The $C^{(v)}$ from each branch is viewed as a perturbation of the consistent graph C^*.

(2) To better discern the contributions of different Riemannian manifold-valued descriptors, graphs closer to the consistent graph should be accorded larger weights. To mitigate the impact of low-quality branch features (*e.g.*, video scenes of children playing with piano music added), we endeavor to allocate different weights to different graphs.

Based on these principles, the fusion mechanism can be formulated as follows:

$$\sum_{v=1}^{b} \omega^{(v)} \| C^{(v)} - C^* \|_F^2, v = 1, \dots, b, \tag{17}$$

where the weight $\omega^{(v)}$ denotes the significance of the different Riemannian manifold-valued descriptors. We employ an inverse distance weighting scheme:

$$\omega^{(v)} = \frac{1}{2\| C^{(v)} - C^* \|_F}. \tag{18}$$

As C^* is not known a priori, we can approximate it iteratively. By integrating Eqs. (16) and (17), we can formulate the following objective function:

$$\begin{aligned} \min_{\{C^{(v)}\}_{v=1}^{b}, C^*} & \sum_{v=1}^{b} (\sum_{i=1}^{n} \| \phi_v(R_i^{(v)}) - \sum_{j=1}^{n} \phi_v(R_j^{(v)}) c_{ij} \|_F^2 \\ & + \lambda \| C^{(v)} \|_1 + \omega^{(v)} \| C^{(v)} - C^* \|_F^2), \\ & s.t. \quad diag(C^{(v)}) = 0, v = 1, \dots, b. \end{aligned} \tag{19}$$

Here, λ and $\omega^{(v)}$ are two non-negative parameters used to balance the three terms, and $\omega^{(v)}$ can be updated with iterative adaptive.

3.3.1. Optimization of M-AVSC

To facilitate the expression of the optimization process, Eq. (19) can be transformed into the following form:

$$\begin{aligned} \min_{C^{(v)}, C^*} & \sum_{i=1}^{n} \| \phi_v(R_i^{(v)}) - \sum_{j=1}^{n} \phi_v(R_j^{(v)}) c_{ij} \|_F^2 \\ & + \lambda \| C^{(v)} \|_1 + \omega^{(v)} \| C^{(v)} - C^* \|_F^2, \\ & s.t. \quad diag(C^{(v)}) = 0. \end{aligned} \tag{20}$$

For optimization purposes, the variables are separated. The auxiliary variables $C_1^{(v)}$, $C_2^{(v)}$, and $A^{(v)}$ are introduced, enabling the reformatting of Eq. (20) as:

$$\begin{aligned} \min_{C_1^{(v)}, C_2^{(v)}, A^{(v)}, C^*} & \sum_{i=1}^{n} \| \phi_v(R_i^{(v)}) - \sum_{j=1}^{n} \phi_v(R_j^{(v)}) a_{ij} \|_F^2 \\ & + \lambda \| C_1^{(v)} \|_1 + \omega^{(v)} \| C_2^{(v)} - C^* \|_F^2, \\ s.t. \ A^{(v)} = C_1^{(v)} - diag(C_1^{(v)}), & A^{(v)} = C_2^{(v)}, v = 1, \dots, b, \end{aligned} \tag{21}$$

since $\phi(\cdot)$ is usually implicit, the regularization term in can be extended using the kernel trick method:

$$\begin{aligned} & \sum_{i=1}^{n} \| \phi_v(R_i^{(v)}) - \sum_{j=1}^{n} \phi_v(R_j^{(v)}) a_{ij} \|_F^2 \\ & = tr(K^{(v)} - 2K^{(v)} A^{(v)} + (A^{(v)})^T K^{(v)} A^{(v)}), \end{aligned} \tag{22}$$

where $K^{(v)} = \phi_v(R^{(v)})^T \phi_v(R^{(v)})$ denotes the kernel matrix for the vth branch. We apply the ADMM (Boyd et al., 2011) to solve the optimization problem in Eq. (19) with the augmented Lagrangian function defined as follows:

$$\begin{aligned} \mathcal{L} & = (C_1^{(v)}, C_2^{(v)}, A^{(v)}, C^*, \Delta_1^{(v)}, \Delta_2^{(v)}) \\ & = tr(K^{(v)} - 2K^{(v)} A^{(v)} + (A^{(v)})^T K^{(v)} A^{(v)}) + \lambda \| C_1^{(v)} \|_1 \\ & + \omega^{(v)} \| C_2^{(v)} - C^* \|_F^2 + \frac{\mu_1}{2} \| A^{(v)} - C_1^{(v)} + diag(C_1^{(v)}) \|_F^2 \\ & + \frac{\mu_2}{2} \| A^{(v)} - C_2^{(v)} \|_F^2 + tr[\Delta_1^{(v)^T} (A^{(v)} - C_1^{(v)} + diag(C_1^{(v)}))] \\ & + tr[\Delta_2^{(v)^T} (A^{(v)} - C_2^{(v)})], \end{aligned} \tag{23}$$

where $\{\Delta_i^{(v)}\}_{i=1}^{2} \in \mathbb{R}^{n \times n}$ is the Lagrange multiplier and $\{\mu_i > 0\}_{i=1}^{2}$ denotes the penalty parameter. We alternatively update each of these variables by minimizing Eq. (23) while fixing the other variables.

Algorithm 2 M-AVSC

```
Input: $R = \{R^{(v)}\}_{v=1}^{b}$, $\lambda > 0$, $\mu_1 > 0$, $\mu_2 > 0$, $\mu_{max} = 10^6$, $\rho = 1.5$
Initialize: $A^{(v)} = 0$, $C_1^{(v)} = 0$, $C_2^{(v)} = 0$, $C^* = 0$, $\Delta_1^{(v)} = 0$, $\Delta_2^{(v)} = 0$, $v = 1, ..., b$
while not converged do
    For $v = 1$ to $b$:
        Update $A^{(v)}$ according to Eq. (24).
        Update $C_1^{(v)}$ according to Eq. (26).
        Update $C_2^{(v)}$ according to Eq. (28).
        Update $\Delta_1^{(v)}$ and $\Delta_2^{(v)}$ according to Eq. (29).
    end for
    Update the penalty variable $\mu := min(\rho\mu, \mu_{max})$;
    Update $C^*$ according to Eq. (30).
end while
Output: $C^*$.
```

Updating $A^{(v)}$. By fixing the remaining variables, we can update $A^{(v)}$ by solving the following sub-problem:

$$\begin{aligned} A^{(v)} = &\{K^{(v)} + (\mu_1 + \mu_2)I\}^{-1} \\ &\times \{K^{(v)} + \mu_1 C_1^{(v)} + \mu_2 C_2^{(v)} - \Delta_1^{(v)} - \Delta_2^{(v)}\}. \end{aligned} \tag{24}$$

where I is an identity matrix.

Updating $C_1^{(v)}$. While fixing the remaining variables, the $C_1^{(v)}$ can be updated by solving the following sub-problem:

$$\min_{C_1^{(v)}} \lambda \|C_1^{(v)}\|_1 + \frac{\mu_1}{2}\|A^{(v)} - C_1^{(v)} + diag(C_1^{(v)}) + \frac{\Delta_1^{(v)}}{\mu_1}\|_F^2. \tag{25}$$

This sub-problem has the following closed-form solution:

$$C_1^{(v)} = \Lambda^{(v)} - diag(\Lambda^{(v)}), \tag{26}$$

where $\Lambda^{(v)} = T_{\frac{\lambda}{\mu_1}}(A^{(v)} + \frac{\Delta_1^{(v)}}{\mu_1})$, T is the soft threshold operator that can be defined as $T_Y(x) = sign(x) \cdot max(|x| - Y, 0)$.

Updating $C_2^{(v)}$. The sub-problem to update $C_2^{(v)}$ can be written as:

$$\begin{aligned} \min_{C_2^{(v)}} &\omega^{(v)}\|C_2^{(v)} - C^*\|_F^2 + \frac{\mu_2}{2}\|A^{(v)} - C_2^{(v)}\|_F^2 \\ &+ tr[\Delta_2^{(v)^T}(A^{(v)} - C_2^{(v)})]. \end{aligned} \tag{27}$$

The derivation of the above equation yields the optimal solution of $C_2^{(v)}$:

$$C_2^{(v)} = (2\omega^{(v)} + \mu_2)^{-1}(2\omega^{(v)}C^* + \mu_2 A^{(v)} + \Delta_2^{(v)}). \tag{28}$$

Updating $\Delta^{(v)}$. The Lagrange multiplier is updated according to the following equations:

$$\begin{aligned} \Delta_1^{(v)} &= \Delta_1^{(v)} + \mu_1(A^{(v)} - C_1^{(v)} + diag(C_1^{(v)})), \\ \Delta_2^{(v)} &= \Delta_2^{(v)} + \mu_2(A^{(v)} - C_2^{(v)}). \end{aligned} \tag{29}$$

Updating C^*. The closed-form solution of C^* can be obtained by setting the partial derivative of Eq. (23) with respect to C^* to zero:

$$C^* = \frac{\sum_{v=1}^{m} \omega^{(v)} C^{(v)}}{\sum_{v=1}^{m} \omega^{(v)}}. \tag{30}$$

The iterative update process described above is continued until either convergence is achieved, or the set maximum number of iterations has been exceeded. This update procedure is summarized in Algorithm 2. Convergence is checked at each iteration k by ensuring the following constraints are met: $\{A_k^{(v)} - A_{k-1}^{(v)}\}_\infty \leq \varepsilon$, $\{A^{(v)} - C_1^{(v)} + diag(C_1^{(v)})\}_\infty \leq \varepsilon$, $\{A^{(v)} - C_2^{(v)}\}_\infty \leq \varepsilon$, where $v = 1, \ldots, b$. Upon acquiring the shared coefficient matrix C^*, which is based on multi-branch features as delineated in Algorithm 2, we derive the final clustering results. This is accomplished by employing the spectral clustering method in accordance with the relation $W = (|C^*| + |C^*|^T)/2$.

3.4. Selection of Kernel function

Following the idea of the kernel method in Euclidean space, it is also feasible to embed a manifold in RKHS applicable to linear Building on the concept of the kernel method in Euclidean space, we extend the approach by embedding a manifold in RKHS, thereby rendering it applicable to linear geometry. Given that visual information (*i.e.*, video frame branch and optical flow branch) exists as points on the Grassmann manifold, we adopt the projection kernel in line with Refs. Harandi, Sanderson, Shirazi, et al. (2011), Hu and Xu (2022).

Projection Kernel: This kernel is defined in Harandi, Salzmann, Jayasumana, et al. (2014), Harandi et al. (2011). Given two samples X_i and X_j located on Grassmann manifolds, the projection kernel can be expressed as follows:

$$K_{Proj}(X_i, X_j) = \|X_i^T X_j\|_F^2 = tr[(X_i X_i^T)(X_j X_j^T)]. \tag{31}$$

The ensuing equation validates its non-negativity. For all $[X_1, \ldots, X_n] \in \mathcal{G}(d,p)$ and $[b_1, \ldots, b_n] \in \mathbb{R}$, we can express:

$$\begin{aligned} &\sum_{i,j=1}^{n} b_i b_j \|X_i^T X_j\|_F^2 = \sum_{i,j=1}^{n} b_i b_j tr(X_i X_i^T X_i X_j^T) \\ &= tr(\sum_{i=1}^{n} b_i X_i X_i^T)^2 = \|\sum_{i=1}^{n} b_i X_i X_i^T\|_F^2 \geq 0. \end{aligned} \tag{32}$$

For the audio signal branch, which can be conceptualized as a point on the SPD manifold, we employ the inner product kernel, in keeping with Refs. Hu and Wu (2020), Jayasumana, Hartley, Salzmann, et al. (2013).

Inner Product Kernel: Leveraging the Frobenius norm and the polarization formula, the inner product of two n-dimensional SPD manifold data sets S_i and S_j in the tangent space can be defined thus:

$$< log(S_i), log(S_j) >= tr[log(S_i) \cdot log(S_j)], \tag{33}$$

where $tr[\cdot]$ denotes the trace of the matrix. In line with Ref. Jayasumana et al. (2013), the kernel function on Sym_d^+ can be derived:

$$K_{Inner}(S_i, S_j) = tr(log(S_i)log(S_j)). \tag{34}$$

However, according to Mercer's theorem, the kernel function must be positive definite to ensure a valid RKHS. Hence, for all $[S_1, \ldots, S_n] \in Sym_d^+$ and $[b_1, \ldots, b_n] \in \mathbb{R}$:

$$\sum_{i,j=1}^{n} b_i b_j tr[log(S_i) \cdot log(S_j)] = \|\sum_{i}^{n} b_i log(S_i)\|_F^2 \geq 0. \tag{35}$$

In conclusion, the kernel function is positive definite and thus satisfies the stipulations of Mercer's theorem.

4. Experimental result and analysis

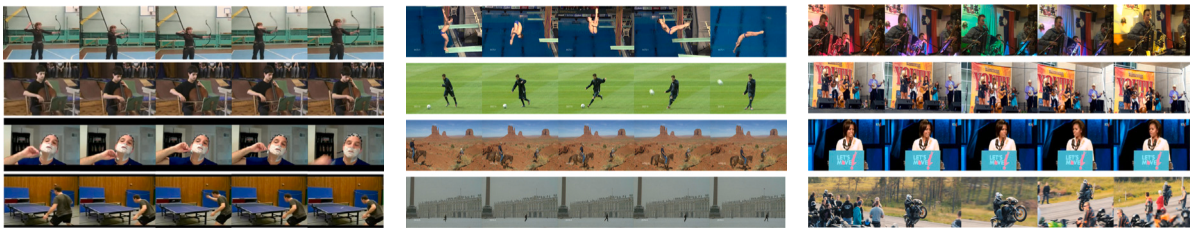
4.1. Evaluation metrics

Six widely used clustering metrics: accuracy, Normalized Mutual Information (NMI), precision, recall, F-score, and Adjusted Rand Index (ARI), are applied. Notably, higher values indicate better clustering performance (Manning, Raghavan, & Schütze, 2008; Xie, Tao, Zhang, et al., 2018).

Accuracy provides an intuitive understanding of the overall correctness of the clustering. Denote r_i as the clustering result and l_i as the ground truth, the accuracy is estimated by:

$$Accuracy = \frac{\sum_{i=1}^{N} \delta(r_i, map(l_i))}{N}, \tag{36}$$

where $\delta(x, y)$ is the delta function that equals 1 if $x = y$ and equals zero otherwise, and $map(l_i)$ is the mapping function that perutes clustering results l_i to the equivalent labels from the dataset. The best mapping can be found by employing the Kuhn–Munkres algorithm.



(a) UCF-101. (b) UCF-sport. (c) AVE.

Fig. 2. Sample frames of different datasets. Each row represents a sequence of frames within a video.

NMI quantifies the mutual information shared between the clustering results and the true labels, we can calculate it following:

$$NMI(U,V) = 2 \times \frac{I(U,V)}{H(U)+H(V)}, \tag{37}$$

where $I(U,V)$ is the mutual information between U and V. For clustering, U and V are clustering results and true labels, respectively. $H(U)$ and $H(V)$ are the entropies of U and V, respectively.

Precision measures the homogeneity within a cluster, gauging the extent to which data points, predominantly from a single true class, are clustered together. Recall, on the other hand, quantifies how comprehensively the data points from a particular true class are grouped together across clusters. The formulas for precision and recall are as follows:

$$Precision = \frac{TP}{TP+FP}, \tag{38}$$

$$Recall = \frac{TP}{TP+FN}, \tag{39}$$

where TP and TN denote the number of true positives and true negatives, FP and FN denote the number of false positives and false negatives. After obtaining precision and recall, we can calculate F-score following:

$$F\text{-}score = 2 \times \frac{Recall \times Precision}{Recall + Precision}, \tag{40}$$

ARI evaluates the consistency of all possible combinations of data point pairs between the clustering results and the true labels. Then the ARI is defined by:

$$ARI = \frac{RI - E(RI)}{max(RI) - E(RI)}, \tag{41}$$

where RI $= \frac{TP+TN}{TP+TN+FP+FN}$ denotes Rand Index. By examining the clustering results through multiple lenses, we aim to capture a more nuanced understanding of the algorithm's efficacy.

All experiments are implemented in MATLAB R2018a software on the computer with an Intel(R) Core (TM) i7 CPU (2.8 GHz) and 8.0 GB RAM. To avoid the effect of randomness caused by the k-means algorithm, each experiment is repeated 20 times, and the average clustering performance with standard deviation is reported. The implementation codes for our approach are available at https://github.com/infinite-hwb/mavsc.

4.2. Description of datasets

In this research, three significant datasets have been utilized: UCF-101 (Soomro, Zamir, & Shah, 2012)[1], UCF-Sport (Rodriguez, Ahmed, & Shah, 2008)[2], and the Audio-Visual Event (AVE) dataset (Tian, Shi, Li, et al., 2018)[3]. For the UCF-101 dataset, not all videos contain audio information. To facilitate a more comprehensive evaluation, we select ten categories and choose 20 individual videos from each category.

Table 1
Summary of the datasets used in this work.

Dataset	Video num.	Category num.	Contains audio
UCF-101	200	10	Yes
UCF-Sport	150	10	No
AVE	4,143	28	Yes

These datasets are rich sources of video sequences derived from various real-world activities and events. While UCF-101 and AVE datasets are collated from YouTube, the UCF-Sport dataset is collated from television channels. To maintain uniformity across experiments, the video clips from all datasets are converted to grayscale images, with varying dimensions based on computational constraints. Table 1 summarizes the key details of these datasets, and few examples from these three dataset are presented in Fig. 2.

4.3. Description of comparative methods

We evaluate the clustering performance of our proposed method in comparison with several other representative methods, namely G-KM (Shirazi et al., 2012), G-LRR (Wang et al., 2014), G-DNLR (Piao et al., 2019), and G-OKSC (Hu & Xu, 2022). We reproduce the code for G-KM and G-LRR based on the description provided in Shirazi et al. (2012), Wang et al. (2014), respectively. The original programs for G-DNLR and G-OKSC are obtained directly from the authors of Hu and Xu (2022), Piao et al. (2019).

In this context, SRSC stands as a single-modality Riemannian subspace clustering method, which is tested using information from three distinct aspects: **v**ideo frames (SRSC(V)), **o**ptical flow (SRSC(O)), and **a**udio signal data (SRSC(A)). Concurrently, M-AVSC is introduced as a multi-modal Riemannian subspace representation method with the ability to exploit audio-visual data for video clustering. The efficacy of SRSC, M-AVSC, and all other comparison methods is tested across three widely accepted public datasets: UCF-101 (Soomro et al., 2012), UCF-Sport (Rodriguez et al., 2008), and AVE (Tian et al., 2018).

4.4. Performance on UCF-101 dataset

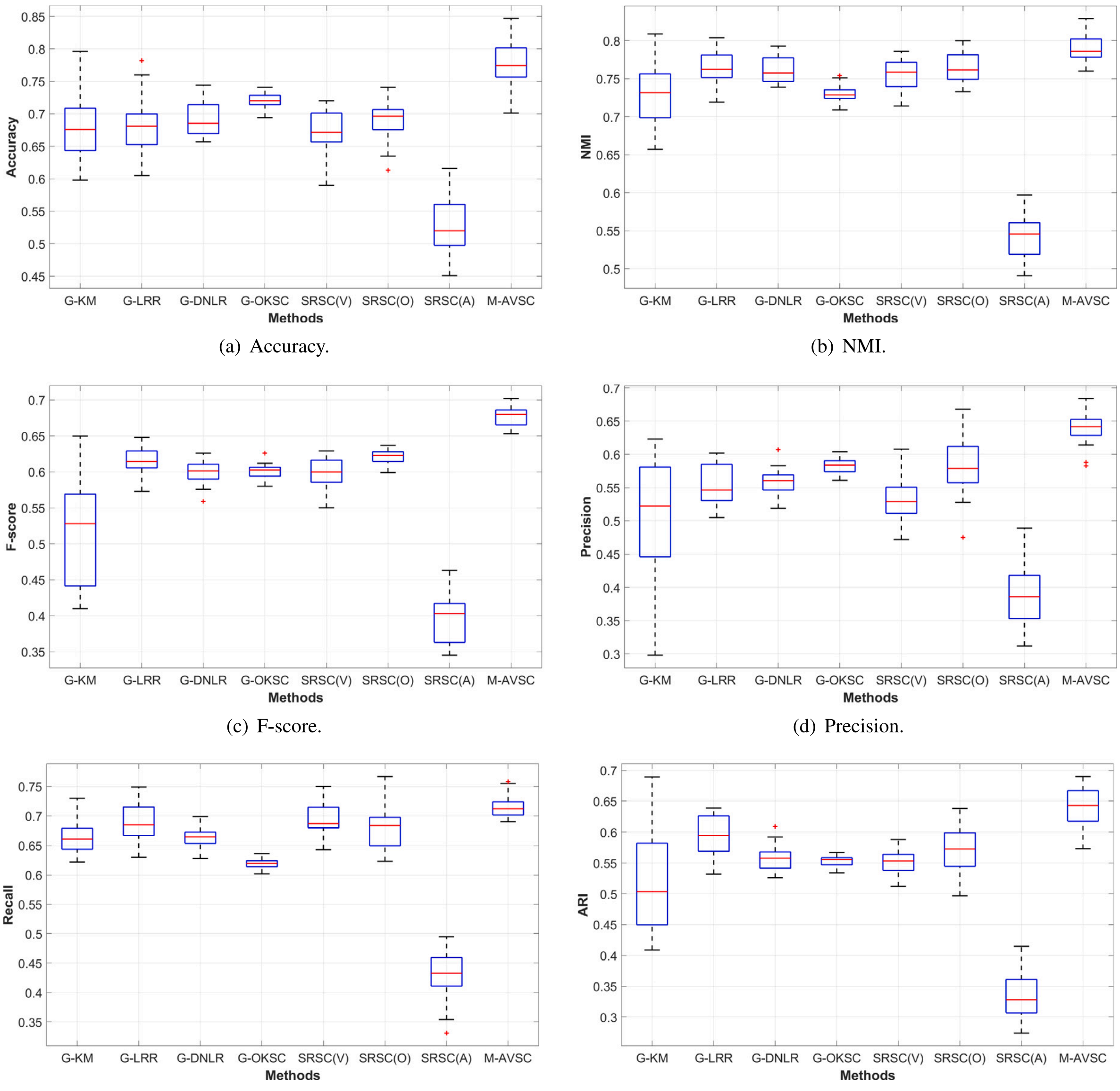
The comparison algorithms, namely G-KM, G-LRR, G-DNLR, and G-OKSC, employed raw frame features of the video for their experimental results. Table 2 presents the clustering performance of all the algorithms on the UCF-101 dataset across 20 independent experiments. The proposed M-AVSC method demonstrated superior performance. More specifically, the average accuracy, NMI, F-score, precision, recall, and ARI of M-AVSC show improvements over the second-best method by 4%, 2%, 6%, 6%, 3%, 6%, respectively.

Fig. 3 provides a more granular view of the performance distributions of all algorithms across the evaluation metrics. Notably, the median performance of M-AVSC consistently emerged at the top echelons, signifying its dominant efficacy. Moreover, the relatively narrower interquartile range of M-AVSC across these metrics underscores its stable performance in the experiments. To shed light on the specific range

[1] https://www.crcv.ucf.edu/data/UCF101.php

[2] https://www.crcv.ucf.edu/data/UCF_Sports_Action.php

[3] https://sites.google.com/view/audiovisualresearch



(a) Accuracy.

(b) NMI.

(c) F-score.

(d) Precision.

(e) Recall.

(f) ARI.

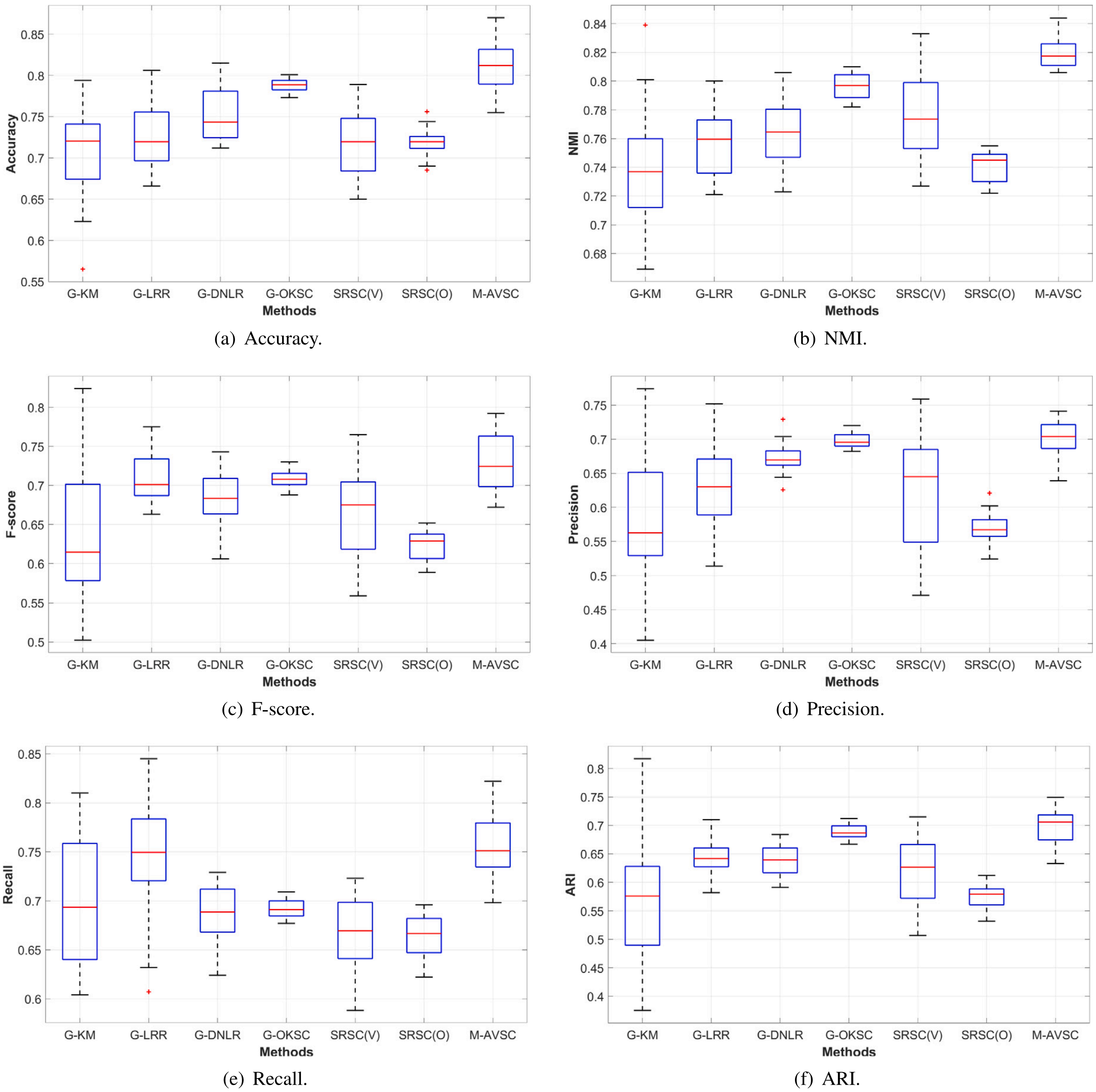
Fig. 3. Box plot representation of clustering metrics for evaluated methods on the UCF-101.

Table 2
The average performance and standard deviation (values in parentheses) of 20 test run on the UCF-100 dataset (10 categories). The best results are highlighted in bold.

Metrics	G-KM	G-LRR	G-DNLR	G-OKSC	SRSC(V)	SRSC(O)	SRSC(A)	M-AVSC
Accuracy	$0.670_{(0.07)}$	$0.680_{(0.04)}$	$0.698_{(0.03)}$	$0.723_{(0.01)}$	$0.677_{(0.03)}$	$0.705_{(0.03)}$	$0.519_{(0.04)}$	$\mathbf{0.761}_{(0.04)}$
NMI	$0.741_{(0.04)}$	$0.769_{(0.02)}$	$0.760_{(0.02)}$	$0.732_{(0.01)}$	$0.762_{(0.02)}$	$0.756_{(0.02)}$	$0.542_{(0.03)}$	$\mathbf{0.794}_{(0.02)}$
F-score	$0.574_{(0.07)}$	$0.621_{(0.02)}$	$0.602_{(0.02)}$	$0.599_{(0.01)}$	$0.602_{(0.02)}$	$0.622_{(0.01)}$	$0.399_{(0.04)}$	$\mathbf{0.679}_{(0.02)}$
Precision	$0.507_{(0.09)}$	$0.563_{(0.03)}$	$0.551_{(0.02)}$	$0.577_{(0.01)}$	$0.533_{(0.04)}$	$0.579_{(0.04)}$	$0.378_{(0.04)}$	$\mathbf{0.641}_{(0.04)}$
Recall	$0.669_{(0.03)}$	$0.694_{(0.03)}$	$0.664_{(0.02)}$	$0.622_{(0.01)}$	$0.694_{(0.04)}$	$0.672_{(0.03)}$	$0.426_{(0.04)}$	$\mathbf{0.723}_{(0.02)}$
ARI	$0.521_{(0.08)}$	$0.576_{(0.03)}$	$0.556_{(0.02)}$	$0.555_{(0.01)}$	$0.554_{(0.02)}$	$0.578_{(0.04)}$	$0.332_{(0.04)}$	$\mathbf{0.642}_{(0.03)}$

of performance achieved by M-AVSC on individual metrics: its accuracy spanned between 0.701 to 0.847, NMI ranged from 0.753 to 0.829, F-score varied between 0.653 to 0.702, precision is between 0.583 to 0.702, recall fluctuated from 0.634 to 0.758, and ARI oscillated from 0.573 to 0.691. These metrics further emphasize the method's consistency and superior clustering performance across the board.

M-AVSC leverages multiple Riemannian manifold-valued descriptors, exploiting audio-visual information and yielding better performance. G-LRR extends low-rank representation onto Grassmann manifolds, employing subspace clustering techniques for video clustering tasks. As a crucial baseline, G-LRR's clustering performance significantly outperformed G-KM, highlighting the advantage of self-expression-based learning methods. It is important to note that some videos might contain additional background music or narration. Interestingly, the clustering results of G-DNLR and G-OKSC exceeded G-LRR, attributable to the introduction of structure constraint on the coefficient matrix. More complex loss functions result from more structural constraints, requiring additional parameters to balance individual loss



(a) Accuracy.

(b) NMI.

(c) F-score.

(d) Precision.

(e) Recall.

(f) ARI.

Fig. 4. Box plot representation of clustering metrics for evaluated methods on the UCF-sport.

Table 3
The average performance and standard deviation (values in parentheses) of 20 test run on the UCF-Sport dataset. The best results are highlighted in bold.

Metrics	G-KM	G-LRR	G-DNLR	G-OKSC	SRSC(V)	SRSC(O)	M-AVSC
Accuracy	$0.703_{(0.08)}$	$0.731_{(0.04)}$	$0.744_{(0.03)}$	$0.791_{(0.01)}$	$0.744_{(0.06)}$	$0.714_{(0.03)}$	$\mathbf{0.798}_{(0.04)}$
NMI	$0.742_{(0.05)}$	$0.764_{(0.02)}$	$0.765_{(0.02)}$	$0.802_{(0.01)}$	$0.778_{(0.03)}$	$0.744_{(0.01)}$	$\mathbf{0.821}_{(0.01)}$
F-score	$0.622_{(0.09)}$	$0.685_{(0.04)}$	$0.684_{(0.03)}$	$0.713_{(0.01)}$	$0.657_{(0.06)}$	$0.615_{(0.02)}$	$\mathbf{0.732}_{(0.03)}$
Precision	$0.577_{(0.11)}$	$0.633_{(0.05)}$	$0.679_{(0.03)}$	$0.698_{(0.01)}$	$0.641_{(0.07)}$	$0.572_{(0.02)}$	$\mathbf{0.708}_{(0.03)}$
Recall	$0.682_{(0.06)}$	$0.750_{(0.05)}$	$0.689_{(0.03)}$	$0.691_{(0.01)}$	$0.676_{(0.04)}$	$0.665_{(0.02)}$	$\mathbf{0.761}_{(0.03)}$
ARI	$0.573_{(0.10)}$	$0.645_{(0.04)}$	$0.647_{(0.03)}$	$0.687_{(0.01)}$	$0.616_{(0.06)}$	$0.571_{(0.02)}$	$\mathbf{0.702}_{(0.03)}$

terms. M-AVSC, with fewer parameters, achieved the best clustering performance.

Compared to SRSC(A), SRSC(V) delivered better clustering performance, underscoring the importance of visual information for clustering performance. SRSC(O) show a slight performance improvement over SRSC(V), indicating that optical flow features assist in capturing discriminative representations in the action video dataset. The best performance is achieved by M-AVSC, suggesting that the fusion of multiple descriptors can enhance performance.

4.5. Performance on UCF-sport dataset

Table 3 showcases the comparison results of different methods on the UCF-Sport dataset. The Grassmann manifold is a superior Riemannian manifold-valued descriptor for videos. The UCF-Sport videos, shot from different angles and spanning a wide range of scenes, present a challenge for any classic clustering method. Interestingly, M-AVSC outperforms other comparison methods, indicating that multiple-view descriptors positively influence clustering performance. In this instance,

Table 4
The average performance and standard deviation (values in parentheses) of 20 test run on the AVE dataset. The best results are highlighted in bold.

Metrics	G-KM	G-LRR	G-DNLR	G-OKSC	SRSC(V)	SRSC(O)	SRSC(A)	M-AVSC
Accuracy	$0.276_{(0.01)}$	$0.260_{(0.01)}$	$0.279_{(0.01)}$	$0.280_{(0.01)}$	$0.280_{(0.01)}$	$0.281_{(0.01)}$	$0.122_{(0.01)}$	$\mathbf{0.302}_{(0.01)}$
NMI	$0.441_{(0.01)}$	$0.380_{(0.01)}$	$0.450_{(0.00)}$	$0.442_{(0.01)}$	$0.460_{(0.01)}$	$0.417_{(0.01)}$	$0.418_{(0.01)}$	$\mathbf{0.461}_{(0.01)}$
F-score	$0.211_{(0.01)}$	$0.192_{(0.01)}$	$0.208_{(0.01)}$	$0.200_{(0.01)}$	$0.209_{(0.01)}$	$0.199_{(0.01)}$	$0.199_{(0.01)}$	$\mathbf{0.231}_{(0.01)}$
Precision	$0.189_{(0.01)}$	$0.169_{(0.00)}$	$0.207_{(0.01)}$	$0.190_{(0.01)}$	$0.198_{(0.01)}$	$0.200_{(0.01)}$	$0.200_{(0.01)}$	$\mathbf{0.219}_{(0.01)}$
Recall	$0.219_{(0.01)}$	$0.200_{(0.01)}$	$0.210_{(0.01)}$	$0.220_{(0.01)}$	$0.220_{(0.01)}$	$0.196_{(0.01)}$	$0.201_{(0.01)}$	$\mathbf{0.240}_{(0.01)}$
ARI	$0.168_{(0.01)}$	$0.148_{(0.01)}$	$0.183_{(0.01)}$	$0.173_{(0.01)}$	$0.182_{(0.01)}$	$0.171_{(0.01)}$	$0.161_{(0.01)}$	$\mathbf{0.198}_{(0.01)}$

M-AVSC does not employ audio information (since UCF-Sport does not provide audio tracks) and uses only video frame and optical flow information. If only visual information is used, SRSC(V) is somewhat analogous to G-LRR. However, compared to G-LRR, SRSC(V) shows a marginal performance improvement, which could be attributed to the utilization of the kernel method. Mapping points on the Grassmann manifold to Hilbert space enables the capture of a richer feature representation. By amalgamating information from two perspectives, M-AVSC surpasses G-LRR by 7%, 1%, 3%, 5%, 1%, and 3% in accuracy, NMI, F-score, precision, recall, and ARI, respectively, evidencing that optical flow is beneficial for action video clustering.

Fig. 4 provides an in-depth view of the performance distributions of all methodologies across the evaluation performance. When examining the box plots of M-AVSC in comparison to other methods, it is evident that M-AVSC consistently displays a higher median, reflecting its performance. Additionally, a more constrained interquartile range for M-AVSC's box plot indicates the method's stability, with a significant portion of the data lying within a narrow range. The median performance of M-AVSC consistently ranks in the upper tiers across these metrics. Furthermore, the relatively tighter interquartile range of M-AVSC emphasizes its consistent performance throughout the experiments. Breaking down M-AVSC's performance: its accuracy spans from 0.755 to 0.870, NMI ranges from 0.806 to 0.844, F-score varies from 0.672 to 0.792, precision is between 0.639 and 0.741, recall lies from 0.698 to 0.822, and ARI ranges between 0.633 and 0.749. Overall, M-AVSC displays both consistent and stable results across the board (see Fig. 5).

4.6. Performance on AVE dataset

As the number of samples increases, the clustering performance of many exceptional clustering methods tends to decrease significantly. The AVE dataset is 27.6 times larger than the previous UCF-Sports dataset. Beyond its larger scale, the AVE dataset also presents a challenge due to the presence of more complex internal scene differences. That is, intra-class distances of videos within the same category are notably varied.

Table 4 summarizes the best clustering performance for each comparison method in the corresponding experiment. It is evident that compared to G-KM, G-LRR, G-DNLR, G-OKSC, and SRSC, which apply single-modality Riemannian manifold-valued descriptors, M-AVSC delivers superior clustering performance. This outcome underscores that multiple Riemannian manifold-valued descriptors can extract more discriminative information, thereby enhancing clustering performance. The clustering performances by G-KM, G-LRR, G-DNLR, and G-OKSC are very similar, illustrating the challenge presented by the AVE dataset. Despite the larger data scale, larger number of classes, and more complex interclass relationships, G-DNLR and G-OKSC do not yield better results, even with structural constraints imposed on the coefficient matrix. M-AVSC, however, outperforms the second-best method by 2%, 1%, 2%, 1%, 2%, and 2% in accuracy, NMI, F-score, precision, recall, and ARI, respectively, thanks to the exploitation of multiple Riemannian manifold-valued descriptors.

Fig. 5 provides a more granular view of the performance distributions of all algorithms across the evaluation metrics. To shed light on the specific range of performance achieved by M-AVSC on individual metrics: its accuracy spanned between 0.293 to 0.312, NMI ranged from 0.441 to 0.472, F-score varied between 0.220 to 0.243, precision is between 0.200 to 0.233, recall fluctuated from 0.228 to 0.253, and ARI oscillated from 0.184 to 0.207.

4.7. Statistical significance analysis

To further substantiate the effectiveness of our method, we conduct a statistical significance test. The results, as illustrated in Fig. 6, showcase the p-values representing the clustering accuracy differences between M-AVSC and the other compared algorithms across the three datasets. Drawing from conventions in Fu, Yang, Chen, and Zhang (2022), Zhong and Pun (2020), a significance level of 0.05 is employed. A p-value lower than this threshold suggests that the performance difference between the two compared algorithms is statistically significant.

From the visualizations in Fig. 6, it is evident that all p-values, irrespective of the dataset, fall below the 0.05 threshold. This consistency not only indicates the statistical significance of the performance disparities between our method and the other algorithms but also reinforces the efficacy of our approach across diverse datasets.

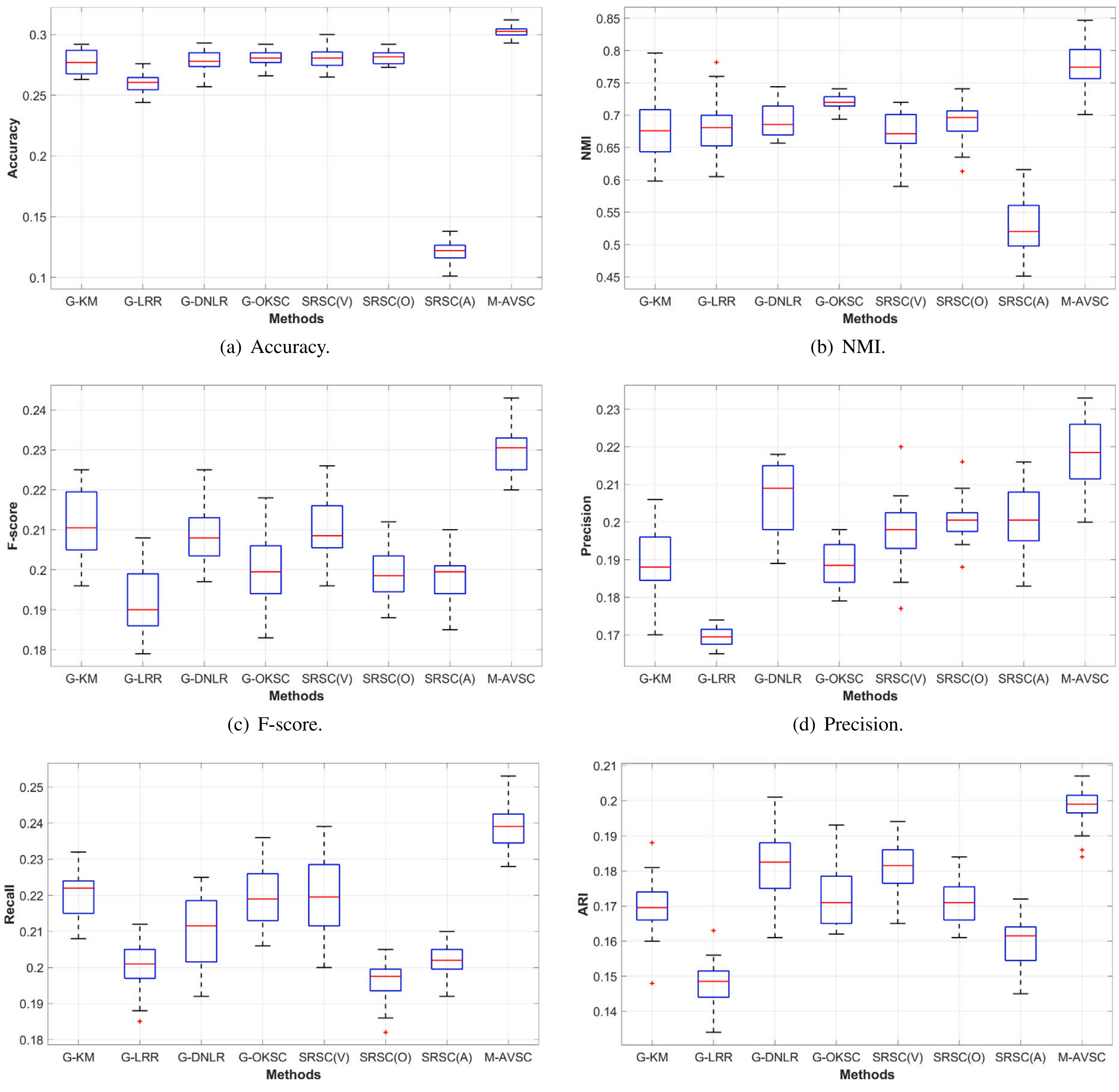
4.8. Ablation study

We conduct three ablation studies to assess the components of our approach. First, we compare diverse video feature representations with our Riemannian manifold-valued descriptors to confirm their effectiveness. Given the nascent state of audio subspace clustering research, we test SRSC's capability with audio signals. Finally, we evaluate combinations of our descriptors for distinct branching features, highlighting the strength of our M-AVSC.

4.8.1. Effectiveness of the visual clustering strategy with Riemannian manifold-valued descriptors

Feature extraction and representation are paramount in video content analysis. To comprehensively capture the video's intricate details, various advanced feature extraction strategies are used such as Lukas-Kanade (Baker & Matthews, 2004), Motion Boundary Histogram (MBH) (Dalal, Triggs, & Schmid, 2006), Histogram of Oriented Gradients (HOG) (Dalal & Triggs, 2005; Lowe, 2004), Histogram of Optical Flow (HOF) (Perš, Sulić, Kristan, et al., 2010) and autoencoders (Ji, Zhang, Li, et al., 2017). The distinct advantage of Riemannian Manifold-valued Descriptors is their ability to reduce each video to a fixed-dimension subspace. This unique property allows for the seamless integration of Riemannian Manifold-valued Descriptors with clustering techniques, something that other feature extraction methods cannot achieve.

To compare the performance of our method with other representations, we duplicate the last frame of videos with fewer frames repeatedly, ensuring that all videos maintain consistent dimensions. Such videos are marked with an asterisk (*) in the table. To be consistent with the SRSC method and to ensure a fair comparison, we apply the commonly used Gaussian kernel function after extracting all the features and ensured that the kernel matrix had a value between 0–1. Table 5 shows the clustering performance obtained on the UCF-Sport dataset using different feature extraction methods.



(a) Accuracy.

(b) NMI.

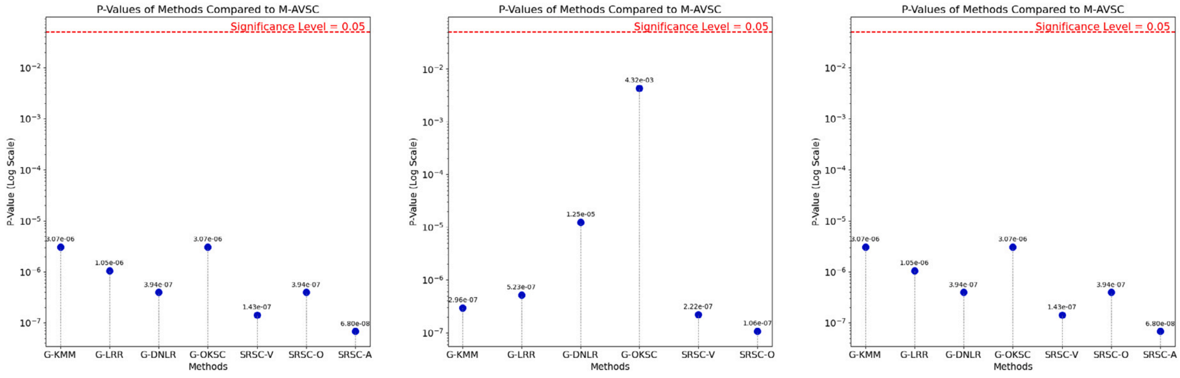
(c) F-score.

(d) Precision.

(e) Recall.

(f) ARI.

Fig. 5. Box plot representation of clustering metrics for evaluated methods on the AVE.



(a) UCF-101.

(b) UCF-sport.

(c) AVE.

Fig. 6. *p*-values of clustering Accuracy between M-AVSC and other methods on the three datasets.

Table 5
Comparative performance analysis of visual feature extraction methods for clustering on the UCF-Sport dataset.

ID	Feature	Accuracy	NMI	F-score	Precision	Recall	ARI
1	RAW[a]	0.477$_{(0.03)}$	0.550$_{(0.03)}$	0.361$_{(0.03)}$	0.283$_{(0.04)}$	0.508$_{(0.06)}$	0.263$_{(0.04)}$
2	LK[a]	0.542$_{(0.04)}$	0.562$_{(0.03)}$	0.412$_{(0.03)}$	0.375$_{(0.04)}$	0.462$_{(0.04)}$	0.336$_{(0.04)}$
3	MBH[a]	0.676$_{(0.05)}$	0.753$_{(0.03)}$	0.594$_{(0.05)}$	0.514$_{(0.06)}$	**0.709**$_{(0.05)}$	0.538$_{(0.06)}$
4	HOG[a]	0.577$_{(0.03)}$	0.604$_{(0.02)}$	0.486$_{(0.03)}$	0.483$_{(0.03)}$	0.490$_{(0.03)}$	0.426$_{(0.03)}$
5	HOF[a]	0.637$_{(0.05)}$	0.705$_{(0.03)}$	0.574$_{(0.05)}$	0.527$_{(0.07)}$	0.633$_{(0.04)}$	0.519$_{(0.06)}$
6	Auto-encoder[a]	0.348$_{(0.03)}$	0.379$_{(0.02)}$	0.221$_{(0.01)}$	0.164$_{(0.01)}$	0.340$_{(0.02)}$	0.095$_{(0.01)}$
7	SRSC(V)	**0.744**$_{(0.06)}$	**0.778**$_{(0.03)}$	**0.657**$_{(0.06)}$	**0.641**$_{(0.07)}$	*0.676*$_{(0.04)}$	**0.616**$_{(0.06)}$
8	SRSC(O)	*0.714*$_{(0.03)}$	*0.744*$_{(0.01)}$	*0.615*$_{(0.02)}$	*0.572*$_{(0.02)}$	0.665$_{(0.02)}$	*0.571*$_{(0.02)}$

[a] Denotes frame padding is employed.

Initially, we evaluate the raw visual features alongside LK optical flow features. Serving as our baseline, their performance, although insightful, remained suboptimal. To further improve the representation of video content, we have explored a range of other feature extraction methods. Among them, MBH describes the boundary motion of objects and performs well in capturing dynamic changes in videos. Meanwhile, HOG and HOF are used for object detection and describing motion patterns in video content, respectively. These methods perform well on certain evaluation criteria, but there is still a gap in their performance compared to MBH. Notably, we also experiment with an auto-encoder that utilizes deep learning techniques. In this model, a fully-connected layer is inserted between the encoding and decoding layers, and the features from this layer are directly analyzed as video features. Although auto-encoders are widely acknowledged as robust feature extraction tools in image and video analysis, their performance in this specific context proves underwhelming. This may be due to the choice of model parameters. From Table 5, we observe that while MBH achieves the best recall score, SRSC(V) demonstrates improvements over MBH by 6.8%, 2.5%, 6.3%, 12.7%, and 7.8% in accuracy, NMI, F-score, precision, and ARI, respectively. Moreover, SRSC(O) also secures second-best performance across most of the metrics. Based on these findings, we conclude that the visual clustering strategy employing Riemannian manifold-valued descriptors is indeed robust.

4.8.2. *Effectiveness of the audio signal clustering strategy with Riemannian manifold-valued descriptors*

When a visual event correlates with a prominent audio signature, this audio can be interpreted as a label depicting the visual content. Several classic subspace representation methods (Elhamifar & Vidal, 2013; Patel & Vidal, 2014) have been dedicated to image clustering tasks, but few have ventured to apply these techniques to audio signal clustering. To underscore the proficiency of SRSC in processing audio signals, we have replicated these classical methods for comparison against SRSC. Two different datasets are used for this comparison to establish a more robust validation. The UrbanSound8K dataset (Salamon, Jacoby, & Bello, 2014), widely used for environmental sound classification, encompasses data from 10 different categories. Additionally, we apply the UCF-101 dataset (consisting of 10 categories), extracting solely the audio signals from the videos as inputs for the different models.

For the spectrogram $T = [t_1, t_2, \ldots, t_n]$, where t_i represents the spectrogram for the ith audio sequence, the clustering results of the original audio signal can be obtained through sparse subspace representation, dubbed as Sparse Subspace Clustering for Audio Signal (SSC(A)). The corresponding loss function can be expressed as follows:

$$\min_{C} \lambda\|C\|_1 + \|T - TC\|_F^2, \quad s.t. \quad diag(C) = 0, \tag{42}$$

where C denotes the coefficient matrix.

Furthermore, as the kernel approach aids in unveiling the nonlinear structure inherent to high-dimensional data, the spectrograms are projected into the RKHS for a richer representation. This leads to the definition of Kernelized Sparse Subspace Clustering for Audio Signal (KSSC(A)). Specifically, the sparse subspace representation based on Hilbert space embedding for the spectrograms is as follows:

$$\min_{C} \lambda\|C\|_1 + \|\phi(T) - \phi(T)C\|_F^2, \quad s.t. \quad diag(C) = 0. \tag{43}$$

Finally, we employ SRSC to address the audio signal clustering problem. SSC(A) and KSSC(A) can be regarded as ablation studies for SRSC(A). For a fair comparison, all contrasted methods use the same data.

Table 6
Comparative audio clustering efficacy on UrbanSound8K and UCF-101 datasets.

(a) UrbanSound8K

Metrics	SSC(A)	KSSC(A)	SRSC(A)
Accuracy	0.297$_{(0.02)}$	0.286$_{(0.02)}$	**0.427**$_{(0.03)}$
NMI	0.242$_{(0.02)}$	0.247$_{(0.02)}$	**0.416**$_{(0.02)}$
F-score	0.138$_{(0.01)}$	0.150$_{(0.01)}$	**0.268**$_{(0.03)}$
Precision	0.125$_{(0.01)}$	0.116$_{(0.01)}$	**0.252**$_{(0.02)}$
Recall	0.154$_{(0.01)}$	0.215$_{(0.03)}$	**0.286**$_{(0.02)}$
ARI	0.030$_{(0.01)}$	0.022$_{(0.01)}$	**0.180**$_{(0.03)}$

(b) UCF-101

Metrics	SSC(A)	KSSC(A)	SRSC(A)
Accuracy	0.310$_{(0.02)}$	0.330$_{(0.02)}$	**0.519**$_{(0.04)}$
NMI	0.282$_{(0.02)}$	0.302$_{(0.02)}$	**0.542**$_{(0.03)}$
F-score	0.192$_{(0.01)}$	0.210$_{(0.01)}$	**0.399**$_{(0.04)}$
Precision	0.180$_{(0.01)}$	0.174$_{(0.01)}$	**0.378**$_{(0.04)}$
Recall	0.206$_{(0.02)}$	0.267$_{(0.03)}$	**0.426**$_{(0.04)}$
ARI	0.100$_{(0.02)}$	0.107$_{(0.01)}$	**0.332**$_{(0.04)}$

As shown in Table 6, the best clustering performance is achieved by SRSC(A), indicating that the strategy with Riemannian manifold-valued descriptors is more adept at extracting discriminative information from the audio signal. Additionally, KSSC(A) yields better clustering performance than SSC(A), possibly due to the nonlinear features present in the spectrogram. Given that SRSC(A) yields the best clustering performance, we infer that the strategy adopted in the audio branch of our multimodal Riemannian subspace clustering design is indeed reliable.

4.8.3. *Different combinations of multiple Riemannian manifold-valued descriptors*

In an effort to further validate our method's efficacy and analyze the fusion effects of each branching feature, we carry out ablation experiments. Table 7 presents the results of the ablation experiments conducted on the UCF-101 dataset. Using information from multiple modalities yields superior clustering performance than methods utilizing a single viewpoint (for instance, solely the video frame). When employing data from a single view, there is no need to utilize a fusion strategy to obtain a shared C^*.

According to Table 7, SRSC with a single viewpoint feature occasionally performs better when exploiting optical flow features as opposed to using only the original video frame features. This is because optical flow features can more effectively capture the relationship between the sequential frames of the video. Interestingly, the best

Table 7
Ablation study on the UCF-101 dataset. The upper part of the table presents the results of 20 independent runs of SRSC using single viewpoint features, while the lower part of the table showcases the results of 20 independent runs of M-AVSC using multiple viewpoint features.

ID	Feature	Accuracy	NMI	F-score	Precision	Recall	ARI
1	Video Frame	$0.677_{(0.03)}$	$0.762_{(0.02)}$	$0.602_{(0.02)}$	$0.533_{(0.04)}$	$0.694_{(0.04)}$	$0.554_{(0.02)}$
2	Optical Flow	$0.705_{(0.03)}$	$0.756_{(0.02)}$	$0.622_{(0.01)}$	$0.579_{(0.04)}$	$0.672_{(0.03)}$	$0.578_{(0.04)}$
3	Audio Signal	$0.519_{(0.04)}$	$0.542_{(0.03)}$	$0.399_{(0.04)}$	$0.378_{(0.04)}$	$0.426_{(0.04)}$	$0.332_{(0.04)}$
4	Video Frame+Optical Flow	$0.737_{(0.03)}$	$0.759_{(0.02)}$	$0.632_{(0.03)}$	$0.591_{(0.04)}$	$0.680_{(0.03)}$	$0.590_{(0.04)}$
5	Video Frame+Audio Signal	$0.735_{(0.05)}$	$0.785_{(0.02)}$	$0.668_{(0.03)}$	$0.636_{(0.04)}$	$0.704_{(0.02)}$	$0.631_{(0.04)}$
6	Optical Flow+Audio Signal	$0.704_{(0.06)}$	$0.760_{(0.03)}$	$0.628_{(0.04)}$	$0.592_{(0.05)}$	$0.668_{(0.04)}$	$0.585_{(0.05)}$
7	Video Frame+Optical Flow+Audio Signal	$\mathbf{0.761}_{(0.04)}$	$\mathbf{0.794}_{(0.02)}$	$\mathbf{0.679}_{(0.02)}$	$\mathbf{0.641}_{(0.04)}$	$\mathbf{0.723}_{(0.02)}$	$\mathbf{0.642}_{(0.03)}$

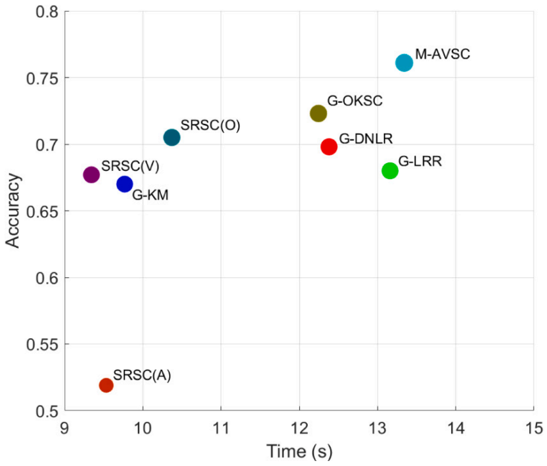


Fig. 7. Performance trade-off between runtime and accuracy for various clustering methods on the UCF-101 dataset.

performance is not achieved by clustering the video using only its audio signal.

For the clustering performance that incorporates information from multiple viewpoints, there is a trend that increasing the number of viewpoints can improve the clustering performance. Compared to employing only the video frame, using both video frame and audio information improves accuracy, NMI, F-score, precision, recall, and ARI by 6%, 3%, 7%, 11%, 1%, and 8%, respectively. This supports the cognitive behavior study asserting that the application of both audio and visual information can lead to more accurate inferences.

The optimal clustering performance is achieved by employing the raw frame + optical flow + audio signal. This suggests that using these features can complement each other effectively, thereby enhancing the clustering performance.

4.9. Timing experiments

As illustrated in Fig. 7, although M-AVSC has a slightly longer execution time, it achieves the highest accuracy among all the compared methods, with a score of 0.761. This emphasizes its ability to ensure high performance while maintaining relative efficiency. The marked improvement of M-AVSC over the unimodal SRSC accentuates the pivotal role of audio-visual data integration in video clustering. As for the trade-off between fast execution and accuracy, SRSC(V) is the fastest but its accuracy is relatively low. Overall, M-AVSC demonstrates significant advantages in the realm of video clustering.

4.10. Parameter selection

This section discusses the parameter selection for M-AVSC, which primarily includes λ and $\omega^{(v)}$. $\omega^{(v)}$ is adaptively adjusted by the inverse of the distance. Initially, we assign the same value μ to all constraints (μ_1 and μ_2). A grid search technique is employed to fine-tune parameters μ and λ. We designate μ as one of the values in the set [10,30,50,70,90,150,200,500] and λ as one of the values in the set [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9]. We subsequently arrange and combine the values of parameters μ and λ to generate a "grid". Fig. 8 illustrates the impact of adjusting both μ and λ on the algorithm's clustering accuracy, showing that M-AVSC is not highly sensitive within the corresponding value range.

The loss function Eq. (19) includes only one equilibrium parameter, λ. Here, we fix μ at 50 for all datasets and then modify the size of λ. Fig. 9 depicts the accuracy scores achieved by adjusting the parameter λ on the three datasets. We establish λ as 0.35, 0.6, and 0.65 for the UCF-101, UCF-sport, and AVE datasets, respectively. The variation of the adaptive parameters $\omega^{(v)}$ per iteration is represented in Fig. 10. The red line shows the weight of each iteration from C^1 through the raw frame, the blue line signifies the weight of each iteration from C^2 through the optical flow, and the black line demonstrates the weight of each iteration from C^3 through the auditory information. As Fig. 10 exhibits, similar to the ablation study, the weight value of C^3 does not attain higher weights, which are generally assigned to C^1 and C^2. As an example, the performance of SRSC(O) in Table 2 is higher than SRSC(V), and the weights $\omega^{(v)}$ assigned to C^2 in Fig. 10(a) is higher than C^1, which indicates that adaptive weights can work well. Most notably, the adaptive parameter $\omega^{(v)}$ gradually converges to a specific interval as the number of iterations increases.

4.11. Convergence analysis

Despite Eq. (23) not being a jointly convex problem across all variables, and thus a global optimal solution still standing as an unresolved issue, we tackle Eq. (23) by employing ADMM (Algorithm 2). Given that each sub-problem is convex, we obtain optimal solutions for each, thereby achieving notable convergence with Algorithm 2. The convergence of each sub-problem is expounded upon as follows.

To optimize variables separately, we introduce auxiliary variables $C_1^{(v)}$, $C_2^{(v)}$, and $A^{(v)}$ while updating $C_1^{(v)}$, $C_2^{(v)}$, and $A^{(v)}$, which simplifies the acquisition of their closed-form solutions.

Similarly, when updating $\omega^{(v)}$, we can deduce the closed-form solution of $\omega(v)$ by resolving Eq. (18), which poses as a linear function of $\omega^{(v)}$.

Consequently, we need only validate that Algorithm 2 converges for C^*. An update to C^* delivers a closed-form solution, and the convergence can also be established based on the lemma below.

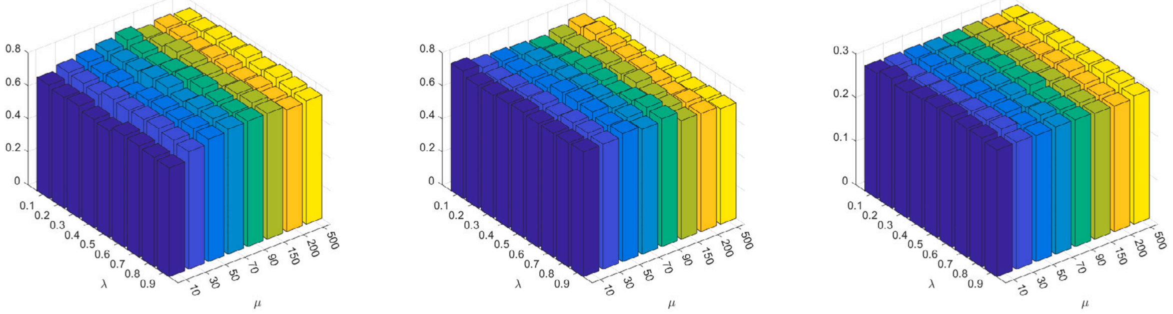
Lemma 1. *For any non-zero matrix $P \in R^{n\times n}$ and $Q \in R^{n\times n}$, the subsequent inequality is upheld:*

$$\|P\|_F - \frac{\|P\|_F^2}{2\|Q\|_F} \le \|Q\|_F - \frac{\|Q\|_F^2}{2\|Q\|_F}. \tag{44}$$

Theorem 1. *In every iteration of Algorithm 2, the update C^* will diminish the value of the objective function until it converges.*

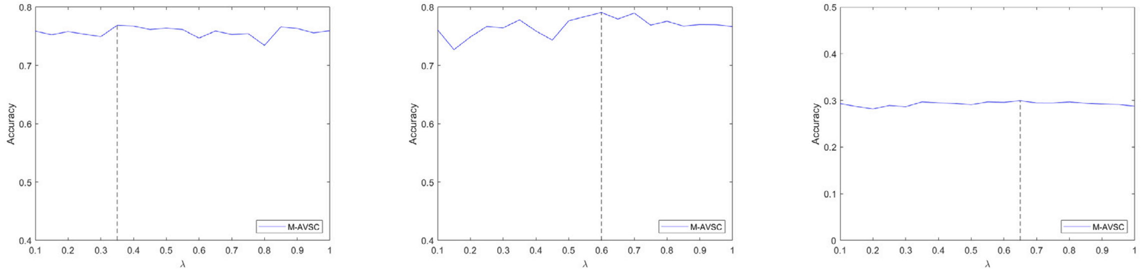
Proof. Let the tth and $(t+1)$-th iteration results of the shared coefficient matrix be denoted by C and $\hat{C}$, respectively. We can infer:

$$\sum_{v=1}^{b} \frac{\|C^{(v)} - \hat{C}^*\|_F^2}{2\|C^{(v)} - C^*\|_F^2} \le \sum_{v=1}^{b} \frac{\|C^{(v)} - C^*\|_F^2}{2\|C^{(v)} - C^*\|_F^2}. \tag{45}$$



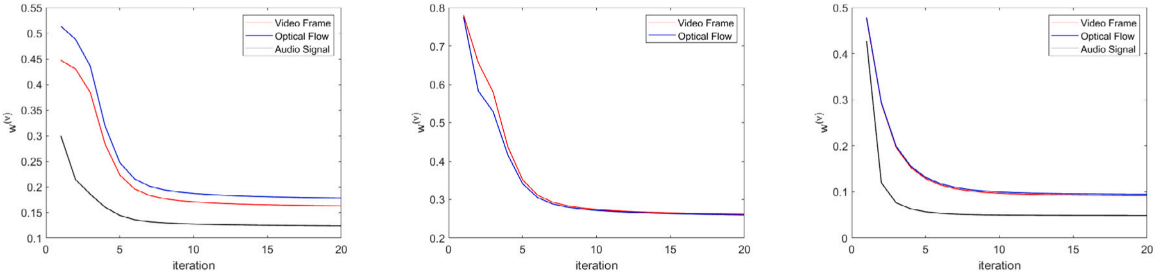
(a) UCF-101. (b) UCF-sport. (c) AVE.

Fig. 8. The effect of changing the parameters μ and λ on the accuracy of clustering. (a) Clustering accuracy with the parameter μ and λ varying on the UCF-101; (b) Clustering accuracy with the parameter μ and λ varying on the UCF-sport; (c) Clustering accuracy with the parameter μ and λ varying on the AVE.



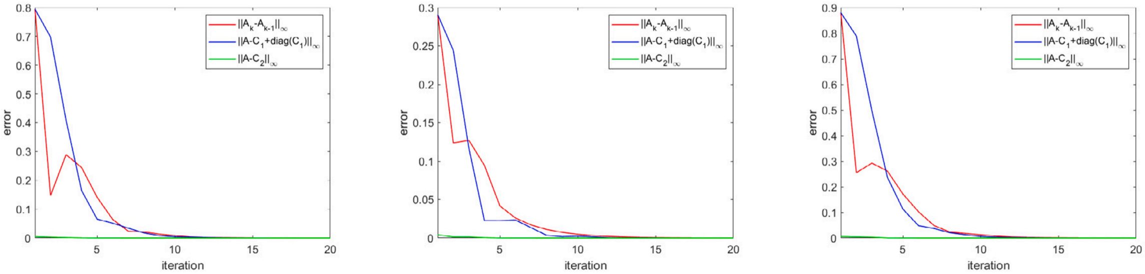
(a) UCF-101. (b) UCF-sport. (c) AVE.

Fig. 9. Parameter selections in three dataset experiment. (a) Clustering accuracy with the parameter λ varying on the UCF-101; (b) Clustering accuracy with the parameter λ varying on the UCF-sport; (c) Clustering accuracy with the parameter λ varying on the AVE.



(a) UCF-101. (b) UCF-sport. (c) AVE.

Fig. 10. The change of the parameter $\omega^{(v)}$ at each iteration on three different datasets.



(a) UCF-101. (b) UCF-sport. (c) AVE.

Fig. 11. The convergence curve of M-AVSC on three datasets. Each subfigure has the x-axis representing the number of iterations and the y-axis symbolizing the sum of normalized errors across three views for M-AVSC.

As per Lemma 1, we deduce:

$$\begin{aligned}&\sum_{v=1}^{b}\|C^{(v)}-\hat{C}^{*}\|-\sum_{v=1}^{b}\frac{\|C^{(v)}-\hat{C}^{*}\|_F^2}{2\|C^{(v)}-C^{*}\|_F^2}\\ &\leq\sum_{v=1}^{b}\|C^{(v)}-C^{*}\|-\sum_{v=1}^{b}\frac{\|C^{(v)}-C^{*}\|_F^2}{2\|C^{(v)}-C^{*}\|_F^2}.\end{aligned}\tag{46}$$

By summing both Eqs. (45) and (46) over both sides, we establish:

$$\sum_{v=1}^{b}\|C^{(v)}-\hat{C}^{*}\|\leq\sum_{v=1}^{b}\|C^{(v)}-C^{*}\|,\tag{47}$$

thereby completing the proof.

Convergence of the algorithm is additionally confirmed through experimentation. We examine convergence by verifying the following constraints in each iteration k: $\{A_k^{(v)}-A_{k-1}^{(v)}\}_\infty \leq \epsilon$, $\{A^{(v)}-C_1^{(v)}+diag(C_1^{(v)})\}_\infty \leq \epsilon$, $\{A^{(v)}-C_2^{(v)}\}_\infty \leq \epsilon$, where $v=1,\ldots,b$. In the conducted experiments, we set the maximum value$T = 50$ and the convergence error tolerance ϵ=0.001. Fig. 11 illustrates the errors of the three views $\{A_k^{(v)}-A_{k-1}^{(v)}\}_\infty$, $\{A^{(v)}-C_1^{(v)}+diag(C_1^{(v)})\}_\infty$, $\{A^{(v)}-C_2^{(v)}\}_\infty$. The errors are normalized and summed across three views. The results reflect that the curve undergoes a sharp decline after 20 iterations before stabilizing, thereby showcasing the impressive convergence of the algorithm.

5. Conclusion

In the realm of real-world scenarios, audio and visual modalities inherently interplay to provide a more comprehensive and enriched representation of the environment. This study delves into the amalgamation of these modalities, aiming to harness the joint power of audio-visual data for enhanced video clustering. In this paper, we propose a single-modality Riemannian subspace clustering technique and later expand it to a multi-modality approach, uniquely emphasizing the fusion of audio-visual data. This method is a marked departure from its predecessors as it accentuates the significance of audio-visual data fusion in video clustering, underpinned by our novel loss function tailored for shared coefficient matrices.

In evaluations conducted on three distinct video datasets, our proposed M-AVSC method demonstrated superior performance in video clustering compared to existing methods. This enhanced performance is attributed to the effective application of multiple Riemannian manifold-valued descriptors, which successfully encapsulate multimodal video information within a unified structure. Additionally, extensive parameter testing confirmed the stability of our method across a wide range of parameter settings. While our methods effectively utilize audio-visual synergy, they rely on predefined manifold-based representations, which may not fully capture the varying complexities of diverse real-world data. Considering this, we would like a fusion of Grassmann manifolds with deep learning architectures in the future. Such an integration, especially with structures like convolutional autoencoders, promises a more adaptable and scalable method for data representation and processing.

CRediT authorship contribution statement

Wenbo Hu: Conceptualization, Software, Writing – original draft, Writing – review & editing. **Hongjian Zhan:** Supervision, Validation, Writing – review & editing. **Yinghong Tian:** Supervision, Validation, Writing – review & editing. **Yujie Xiong:** Supervision, Validation, Writing – review & editing. **Yue Lu:** Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my data/code in the paper.

Annex A. Definition of Grassmann manifolds.

Generally, a manifold is considered a topological space whose local neighborhood is approximated as a Euclidean space (Absil, Mahony, & Sepulchre, 2009). Two typical manifolds are regarded as quotients of the special orthogonal group $\mathcal{SO}_p$ are the compact Stiefel manifold and the Grassmann manifold (Absil, Mahony, & Sepulchre, 2004; Edelman, Arias, & Smith, 1998), where $\mathcal{SO}_p$ is the smooth differential manifold (Lie group structure) of all $p \times p$ orthogonal matrices with determinant +1 (subspace of the orthogonal group $(\mathcal{O}_d)$).

Definition 1. The Stiefel manifold $\mathcal{S}(d,p), d \geq p$, is a Riemannian manifold composed of all $d \times p$ orthonormal matrices $\{Y \in \mathbb{R}^{d\times p} : Y^T Y = I_p\}$, where I_p denotes the $p \times p$ identity matrix.

In contrast to the Stiefel manifold, the basis selection for the subspace of the Grassmann manifold is non-unique. A point on the Grassmann manifold, $\mathcal{G}_{p,d}$, is a p-dimensional linear subspace of the d-dimensional Euclidean space, identified by an orthogonal basis. All orthonormal matrices that span the same subspace are considered equivalent, which allows interpretation of each point on the Grassmann manifold as an equivalent point on the Stiefel manifold.

Definition 2. A point on the Grassmann manifold can be represented by an orthonormal matrix $Y \in \mathbb{R}^{d\times p}$, where the columns span the corresponding subspace.

Consequently, we can deduce the Stiefel representation of the Grassmann manifold:

$$\mathcal{G}_{p,d}=\{span(Y) : Y\in\mathbb{R}^{d\times p} : Y^T Y = I_p\}.\tag{48}$$

The Grassmann manifold possesses a Riemannian structure that allows for the performance of calculus operations. Given that the Grassmann manifold is smooth and curved, applying the Euclidean metric directly would be inappropriate. The commonly used distance measure on the Grassmann manifold is the embedding distance (Harandi, Sanderson, Shen, et al., 2013).

Annex B. Definition of SPD manifolds.

Symmetric positive definite matrices are well-known for their powerful representation capabilities. These can be obtained by constructing a covariance matrix of image features. It is important to note that the symmetric positive definite matrix contains a Riemannian manifold structure, making the SPD matrix space non-linear.

Definition 3. For a space of $d \times d$ SPD matrices, denote as Sym_d^+, the mathematical expression of the SPD manifold is:

$$Sym_d^+=\{S\in\mathbb{R}^{d\times d} : x^T S x>0, \exists x\in\mathbb{R}^d-\{0_d\}\}.\tag{49}$$

Given that Sym_d^+ forms a convex cone within the d^2 dimensional Euclidean space, the Riemannian metric serves as a more accurate distance measure on Sym_d^+. Generally, there are two popular Riemannian metrics proposed on Sym_d^+, which include the affine-invariant Riemannian metric (Pennec, Fillard, & Ayache, 2006) and the log-Euclidean Riemannian metric (Arsigny, Fillard, Pennec, et al., 2005, 2006).

References

Absil, P. A., Mahony, R., & Sepulchre, R. (2004). Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematica*, *80*(2), 199–220. http://dx.doi.org/10.1023/b:acap.0000013855.14971.91.

Absil, P., Mahony, R., & Sepulchre, R. (2009). *Optimization algorithms on matrix manifolds*. Princeton University Press, https://dl.acm.org/doi/abs/10.5555/1557548.

Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., et al. (2021). Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In *Advances in neural information processing systems*: *vol. 34*, (pp. 24206–24221). https://api.semanticscholar.org/CorpusID:233346984.

Albadr, M. A. A., et al. (2021). Extreme learning machine for automatic language identification utilizing emotion speech data. In *International conference on electrical, communication, and computer engineering*. http://dx.doi.org/10.1109/icecce52056.2021.9514107.

Albadr, M. A. A., et al. (2022). Particle swarm optimization-based extreme learning machine for covid-19 detection. *Cognitive Computation*, 1–16. http://dx.doi.org/10.1007/s12559-022-10063-x.

Ali, A. H., Yaseen, M. G., Aljanabi, M., et al. (2023). Transfer learning: A new promising techniques. *Mesopotamian Journal of Big Data*, 31–32. http://dx.doi.org/10.58496/mjbd/2023/004.

Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., & Tran, D. (2020). Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 9758–9770, https://api.semanticscholar.org/CorpusID:208513596.

Arandjelovic, R., & Zisserman, A. (2017). Look, listen and learn. In *International Conference on Computer Vision* (pp. 609–617). http://dx.doi.org/10.1111/j.1464-410x.2011.10134.x.

Arsigny, V., Fillard, P., Pennec, X., et al. (2005). *Fast and simple computations on tensors with log-Euclidean metrics*. INRIA, https://inria.hal.science/inria-00070423/document.

Arsigny, V., Fillard, P., Pennec, X., et al. (2006). Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *56*(2), 411–421. http://dx.doi.org/10.1002/mrm.20965.

Baker, S., & Matthews, I. (2004). Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, *56*(3), 221–255. http://dx.doi.org/10.1023/b:visi.0000011205.11775.fd.

Boyd, S., Parikh, N., Chu, E., et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Machine Learning*, *3*(1), 1–122. http://dx.doi.org/10.1561/9781601984616.

Chauhan, N., Isshiki, T., & Li, D. (2019). Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database. In *International conference on computer and communication systems* (pp. 130–133). http://dx.doi.org/10.1109/ccoms.2019.8821751.

Chen, Q., & Huang, G. (2021). A novel dual attention-based BLSTM with hybrid features in speech emotion recognition. *Engineering Applications of Artificial Intelligence*, *102*, Article 104277. http://dx.doi.org/10.1016/j.engappai.2021.104277.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Conference on computer vision and pattern recognition* (pp. 886–893). http://dx.doi.org/10.1109/cvpr.2005.177.

Dalal, N., Triggs, B., & Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *European conference on computer vision* (pp. 428–441). http://dx.doi.org/10.1007/11744047_33.

Daubechies, I., Defrise, M., & De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, *57*(11), 1413–1457. http://dx.doi.org/10.1002/cpa.20042.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, *41*(3), 613–627. http://dx.doi.org/10.1109/18.382009.

Edelman, A., Arias, T. A., & Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, *20*(2), 303–353. http://dx.doi.org/10.1137/s0895479895290954.

Elhamifar, E., & Vidal, R. (2013). Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(11), 2765–2781. http://dx.doi.org/10.1109/tpami.2013.57.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, *4*(12), 2379–2394. http://dx.doi.org/10.1364/josaa.4.002379.

Fu, L., Yang, J., Chen, C., & Zhang, C. (2022). Low-rank tensor approximation with local structure for multi-view intrinsic subspace clustering. *Information Sciences*, *606*, 877–891.

Gao, R., Oh, T.-H., Grauman, K., & Torresani, L. (2020). Listen to look: Action recognition by previewing audio. In *International conference on computer vision and pattern recognition* (pp. 10457–10467). http://dx.doi.org/10.1109/cvpr42600.2020.01047.

Ghandoura, A., Hjabo, F., & Al Dakkak, O. (2021). Building and benchmarking an arabic speech commands dataset for small-footprint keyword spotting. *Engineering Applications of Artificial Intelligence*, *102*, Article 104267. http://dx.doi.org/10.1016/j.engappai.2021.104267.

Harandi, M. T., Salzmann, M., Jayasumana, S., et al. (2014). Expanding the family of Grassmannian kernels: An embedding perspective. In *European conference on computer vision* (pp. 408–423). http://dx.doi.org/10.1007/978-3-319-10584-0_27.

Harandi, M. T., Sanderson, C., Shen, C., et al. (2013). Dictionary learning and sparse coding on Grassmann manifolds: An extrinsic solution. In *IEEE international conference on computer vision* (pp. 3120–3127). http://dx.doi.org/10.1109/iccv.2013.387.

Harandi, M. T., Sanderson, C., Shirazi, S., et al. (2011). Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching. In *International conference on computer vision and pattern recognition* (pp. 2705–2712). http://dx.doi.org/10.1109/cvpr.2011.5995564.

Hu, D., Li, X., Mou, L., Jin, P., Chen, D., Jing, L., et al. (2020). Cross-task transfer for geotagged audiovisual aerial scene recognition. In *European conference on computer vision* (pp. 68–84). http://dx.doi.org/10.1007/978-3-030-58586-0_5.

Hu, W., & Wu, X. (2020). Multi-geometric sparse subspace clustering. *Neural Processing Letters*, *52*(1), 849–867. http://dx.doi.org/10.1007/s11063-020-10274-z.

Hu, W., & Xu, T. (2022). One-step kernelized sparse clustering on Grassmann manifolds. *Multimedia Tools and Applications*, *81*, 31017–31038. http://dx.doi.org/10.1007/s11042-022-12495-x.

Jayasumana, S., Hartley, R., Salzmann, M., et al. (2013). Kernel methods on the Riemannian manifold of symmetric positive definite matrices. In *International conference on computer vision and pattern recognition* (pp. 73–80). http://dx.doi.org/10.1109/cvpr.2013.17.

Ji, P., Zhang, T., Li, H., et al. (2017). Deep subspace clustering networks. *Advances in Neural Information Processing Systems*, https://api.semanticscholar.org/CorpusID:3470712.

Kazakos, E., Nagrani, A., Zisserman, A., & Damen, D. (2019). Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *International conference on computer vision and pattern recognition* (pp. 5492–5501). http://dx.doi.org/10.1109/iccv.2019.00559.

Kudithipudi, D., Aguilar-Simon, M., Babb, J., et al. (2022). Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, *4*(3), 196–210. http://dx.doi.org/10.1038/nmi.2022.15.

Li, Y., Parsan, A., Wang, B., et al. (2023). A multi-tasking model of speaker-keyword classification for keeping human in the loop of drone-assisted inspection. *Engineering Applications of Artificial Intelligence*, *117*, Article 105597. http://dx.doi.org/10.1016/j.engappai.2022.105597.

Liu, Z., Hu, D., Wang, Z., et al. (2023). LatLRR for subspace clustering via reweighted Frobenius norm minimization. *Expert Systems with Applications*, *224*, Article 119977. http://dx.doi.org/10.1016/j.eswa.2023.119977.

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., & Ma, Y. (2012). Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 171–184. http://dx.doi.org/10.1201/b20190-13.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*, 91–110. http://dx.doi.org/10.1023/b:visi.0000029664.99615.94.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Syngress Publishing, http://informationretrieval.org.

Morgado, P., Li, Y., & Nvasconcelos, N. (2020). Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 4733–4744, https://api.semanticscholar.org/CorpusID:226237128.

Morgado, P., Vasconcelos, N., & Misra, I. (2021). Audio-visual instance discrimination with cross-modal agreement. In *International conference on computer vision and pattern recognition*. http://dx.doi.org/10.1109/cvpr46437.2021.01229.

Muhammad, G., & Alghathbar, K. (2009). Environment recognition from audio using MPEG-7 features. In *International conference on embedded and multimedia computing* (pp. 1–6). http://dx.doi.org/10.1109/em-com.2009.5402978.

Owens, A., & Efros, A. A. (2018). Audio-visual scene analysis with self-supervised multisensory features. In *European conference on computer vision* (pp. 631–648). http://dx.doi.org/10.1007/978-3-030-01231-1_39.

Patel, V. M., & Vidal, R. (2014). Kernel sparse subspace clustering. In *IEEE international conference on image processing* (pp. 2849–2853). http://dx.doi.org/10.1109/icip.2014.7025576.

Pennec, X., Fillard, P., & Ayache, N. (2006). A Riemannian framework for tensor computing. *International Journal of Computer Vision*, *66*(1), 41–66. http://dx.doi.org/10.1007/s11263-005-3222-z.

Perš, J., Sulić, V., Kristan, M., et al. (2010). Histograms of optical flow for efficient representation of body motion. *Pattern Recognition Letters*, *31*(11), 1369–1376, https://api.semanticscholar.org/CorpusID:15952865.

Pham, L., Schindler, A., Schutz, M., Lampert, J., Schlarb, S., & King, R. (2022). Deep learning frameworks applied for audio-visual scene classification. *Data Science–Analytics and Applications*, 39–44. http://dx.doi.org/10.31219/osf.io/6hxrq.

Piao, X., Hu, Y., Gao, J., et al. (2019). Double nuclear norm based low rank representation on Grassmann manifolds for clustering. In *Conference on computer vision and pattern recognition* (pp. 12075–12084). http://dx.doi.org/10.1109/cvpr.2019.01235.

Rodriguez, M. D., Ahmed, J., & Shah, M. (2008). Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *International conference on computer vision and pattern recognition* (pp. 1–8). http://dx.doi.org/10.1109/cvpr.2008.4587727.

Salamon, J., Jacoby, C., & Bello, J. P. (2014). A dataset and taxonomy for urban sound research. In *ACM International conference on multimedia* (pp. 1041–1044). http://dx.doi.org/10.1145/2647868.2655045.

Scheliga, S., Kellermann, T., Lampert, A., et al. (2023). Neural correlates of multisensory integration in the human brain: An ALE meta-analysis. *Reviews in the Neurosciences*, *34*(2), 223–245. http://dx.doi.org/10.1234/rns.2023.45.

Senocak, A., Kim, J., Oh, T.-H., et al. (2023). Event-specific audio-visual fusion layers: A simple and new perspective on video understanding. In *Winter conference on applications of computer vision* (pp. 2237–2247). http://dx.doi.org/10.1109/wacv56688.2023.00227.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905. http://dx.doi.org/10.1049/cp.2012.0440.

Shirazi, S., Harandi, M. T., Sanderson, C., et al. (2012). Clustering on Grassmann manifolds via kernel embedding with application to action analysis. In *IEEE international conference on image processing* (pp. 781–784). http://dx.doi.org/10.1109/icip.2012.6466976.

Song, Q., Sun, B., & Li, S. (2022). Multimodal sparse transformer network for audio-visual speech recognition. *IEEE Transactions on Neural Networks and Learning Systems*, http://dx.doi.org/10.1109/tnnls.2022.3163771.

Song, K., Yao, X., Nie, F., et al. (2021). Weighted bilateral K-means algorithm for fast co-clustering and fast spectral clustering. *Pattern Recognition*, *109*, Article 107560. http://dx.doi.org/10.1016/j.patcog.2020.107560.

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402. https://arxiv.org/pdf/1212.0402.pdf.

Souli, S., & Lachiri, Z. (2012). Environmental sound classification using log-gabor filter. In *International conference on signal processing*: *vol. 1*, (pp. 144–147). http://dx.doi.org/10.1109/icosp.2012.6491621.

Tagliasacchi, M., Gfeller, B., de Chaumont Quitry, F., & Roblek, D. (2020). Pre-training audio representations with self-supervision. *Signal Processing Letters*, *27*, 600–604. http://dx.doi.org/10.1109/lsp.2020.2985586.

Tan, H., Zhou, Y., Tao, Q., et al. (2021). Bioinspired multisensory neural network with crossmodal integration and recognition. *Nature Communications*, *12*(1), 1120. http://dx.doi.org/10.1038/s41467-021-21404-z.

Tian, Y., Shi, J., Li, B., et al. (2018). Audio-visual event localization in unconstrained videos. In *European conference on computer vision* (pp. 247–263). http://dx.doi.org/10.1007/978-3-030-01216-8_16.

Vafeiadis, A., Votis, K., Giakoumis, D., et al. (2020). Audio content analysis for unobtrusive event detection in smart homes. *Engineering Applications of Artificial Intelligence*, *89*, Article 103226. http://dx.doi.org/10.1016/j.engappai.2019.08.020.

Wang, B., Hu, Y., Gao, J., Sun, Y., & Yin, B. (2014). Low rank representation on Grassmann manifolds. In *Asian conference on computer vision* (pp. 81–96). http://dx.doi.org/10.1007/978-3-319-16865-4_6.

Wang, B., Hu, Y., Gao, J., et al. (2018). Cascaded low rank and sparse representation on Grassmann manifolds. In *International joint conference on artificial intelligence* (pp. 2755–2761). http://dx.doi.org/10.24963/ijcai.2018/382.

Wang, S., Liu, X., Liu, L., et al. (2021). Late fusion multiple kernel clustering with proxy graph refinement. *IEEE Transactions on Neural Networks and Learning Systems*, http://dx.doi.org/10.1109/tnnls.2021.3117403.

Wang, S., Mesaros, A., Heittola, T., & Virtanen, T. (2021). A curated dataset of urban scenes for audio-visual scene analysis. In *IEEE international conference on acoustics, speech and signal processing* (pp. 626–630). http://dx.doi.org/10.1109/icassp39728.2021.9415085.

Wang, J., Wu, B., Ren, Z., et al. (2023). Multi-scale deep multi-view subspace clustering with self-weighting fusion and structure preserving. *Expert Systems with Applications*, *213*, Article 119031. http://dx.doi.org/10.1016/j.eswa.2022.119031.

Xie, Y., Tao, D., Zhang, W., et al. (2018). On unifying multi-view self-representations for clustering by tensor multi-rank minimization. *International Journal of Computer Vision*, *126*(11), 1157–1179. http://dx.doi.org/10.1007/s11263-018-1086-2.

Yang, K., Marković, D., Krenn, S., et al. (2022). Audio-visual speech codecs: Rethinking audio-visual speech enhancement by re-synthesis. In *International conference on computer vision and pattern recognition* (pp. 8227–8237). http://dx.doi.org/10.1109/cvpr52688.2022.00805.

Zhang, Q., Kang, Z., Xu, Z., et al. (2022). Spaks: Self-paced multiple kernel subspace clustering with feature smoothing regularization. *Knowledge-Based Systems*, *253*, Article 109500. http://dx.doi.org/10.1016/j.knosys.2022.109500.

Zhong, G., & Pun, C.-M. (2020). Subspace clustering by simultaneously feature selection and similarity learning. *Knowledge-Based Systems*, *193*, Article 105512, https://api.semanticscholar.org/CorpusID:213962199.