IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received February 12, 2019, accepted March 4, 2019, date of publication March 18, 2019, date of current version April 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2905641

A Video Representation Method Based on Multi-View Structure Preserving Embedding for Action Retrieval

KE ZHANG[1], HUI SUN[2], WEILI SHI[1], YUWEN FENG[3], ZHENGANG JIANG[1], AND JIANPING ZHAO[1]

[1]School of Computer Science and Technology, Changchun University of Science and Technology, Changchun 130022, China
[2]College of Humanities and Science, Northeast Normal University, Changchun 130117, China
[3]Jilin University, Changchun 130012, China

Corresponding authors: Zhengang Jiang (jiangzhengang@cust.edu.cn) and Jianping Zhao (1991322678@qq.com)

This work was supported in part by the Fund of Jilin Provincial Science and Technology Department under Grant 20190302112GX, Grant 20190201305JC, Grant 20190201196JC, and Grant 20170204031GX, in part by the Fund of Jilin Provincial Education Department under Grant JJKH20190595KJ, in part by the National Key Research and Development Program of China under Grant 2017YFC0108303, and in part by the National Natural Science Foundation of China under Grant 61602221.

ABSTRACT The content-based video retrieval is a popular topic in computer vision filed, especially, action retrieval. This paper proposes a novel and effective video representation module for content-based action retrieval framework, in which feature learning can be conducted with complementary information and intrinsic structure, where the relationship between appearance feature and geometry can be preserved. Based on multi-view analysis and graph embedding, the target features are generated to minimize the inter-class discrepancy and maximize intra-class discrimination. Applied to the content-based retrieval task, the proposed method can be combined with Euclidean distance for the comparison of low-dimensional features. As demonstrated in the extensive experiments on the benchmark datasets, the performance of the proposed framework is superior to the state-of-the-art methods.

INDEX TERMS Action retrieval, video representation, multi-view analysis, graph embedding.

I. INTRODUCTION

With the explosive growth of video data, the content-based video retrieval (CBVR) has become one of the most active topics in computer vision field due to a wide range of applications [1], [2], such as video surveillance, medical auxiliary therapy and human-computer interaction. According to the content of video, the CBVR domain can be categorized into scene retrieval [3], actor retrieval [4] and action retrieval [5]. Different from the formers, action retrieval can gather all similar actions cross-scene and cross-actor for further action recognition and analysis, arousing an increasing attention from researchers [6].

The content-based action retrieval analyzes the content with the minimum of human participation to gain more accuracy performance than traditional text-based methods when it is lack of label or mistake occurs. However, this task is challenging due to variations of appearance and velocity of actors, size, viewpoint, scene illumination, occlusion, etc. [7].

The goal of content-based video retrieval is to recall database clips which are semantically similar to the query in a ranked order. The video representation aims at minimizing inter-class variance and eliminating the gap between visual features and action understanding. Different from single-image retrieval, both spatial content and temporal content should be exploited for video retrieval. The classical video representation methods are divided into two categories: global representation and local representation. Common global methods encode object silhouettes or motion information in a single feature vector [8]–[10], while local methods characterize the action as a collection of local content after space-time interest point detector and descriptors. 3D Harris corner detector which calculated gradients along x, y and t was the first proposed video detector and extended in [12], [13]. Dollar *et al.* [14] applied Gaussian smoothing

The associate editor coordinating the review of this manuscript and approving it for publication was Mingjun Dai.

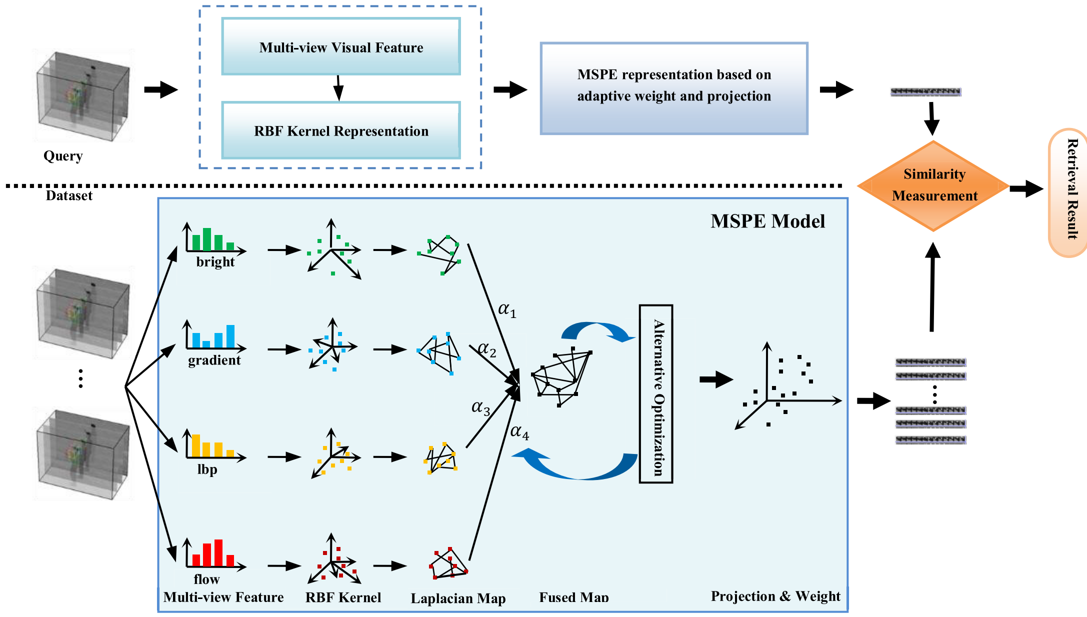

FIGURE 1. The proposed MSPE-based retrieval framework.

kernel and Gabor filtering on both spatial and temporal dimensions individually, defined the location corresponding to the local maxima of the response function as the center of a salient region. Then regions can be described by HOF, HOG3D, 3D SIFT, etc. for action retrieval applications [15], [5].

To avoid the curse of dimensionality issued in high-dimensional data, dimensionality reduction is a powerful tool to improve computational efficiency in machine learning and computer vision. Dimensionality reduction based on manifold learning [16]–[20] aims to explore the intrinsic structure of high-dimensional nonlinear data which is assumed to sample on the low dimensional manifold of the high dimensional space. Generally, these methods based on graph theory are proposed as the instance of spectral embedding methods [21]–[23], for example, Laplacian Eigenmaps [24], Locality Preserving Projection [25] and Neighborhood Preserving Embedding [26]. An affinity matrix with the elements which represent the edge weights of graph is constructed for internal relationship and intrinsic structure representation. The embedding between original feature space and the related low-dimensional space can achieve the low-dimensional representation of high-dimensional data, where various orthogonal perceptual content still belong to the previous directions or clusters.

The real data contains abundant visual contents which provide independently and mutually complementary information from various views. Therefore, exploring internal structure and low-dimensional feature representation of high-dimensional data from multiple viewpoints is benefit for the robustness of feature learning, which effectively avoids noise and occlusion caused by single viewpoint and enhances the retrieval performance. The most state-of-the-art of multi-view methods focus on data clustering [27]–[30]. Some algorithms attempt to learn the manifold structure of high-dimensional data with multi-view analysis, but average contribution of each viewpoint leads to performance degradation due to noise and irrelevant content occurred in some views. Shao *et al.* [31] proposed an adaptive weight adjustment method. However, robust structure constraint is ignored during dimensionality reduction.

Inspired by the above problems, we present a novel feature learning method which is served as a part of video representation for the content-based retrieval task. The proposed method can preserve the internal structure of high-dimensional data with the inter-class and intra-class constraint. All information is adaptively fused in multi-view feature spaces so that the low-dimensional representation is more discriminative. A series of experiments conducted on the benchmarks demonstrate that the proposed module contributes to action retrieval. This paper is organized as follows. Section 1 introduces the related works. The proposed method is detailed in the next section. In the Section 3, the retrieval experiments and performance evaluation are reported. The conclusions are given in the Section 4.

II. THE PROPOSED METHOD

Different from previous video presentation approaches, Multi-view Structure Preserving Embedding (MSPE) is constructed among various visual spaces to jointly analyze intrinsic structure of high-dimensional data, and learn the discriminative feature representation with complementary information. Therefore, multi-view video content integrating, adaptively fusing and graph embedding learning of high-dimensional data in low-dimensional manifold space for action retrieval is the main contribution of the proposed method. The MSPE-based retrieval framework is illustrated as Figure 1.

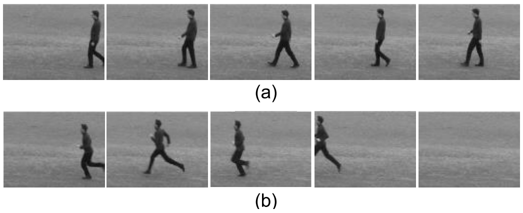

FIGURE 2. The action in 25 sequential frames of: (a) walking and (b) running. (Note: images are extracted with an interval of 5 frames.)

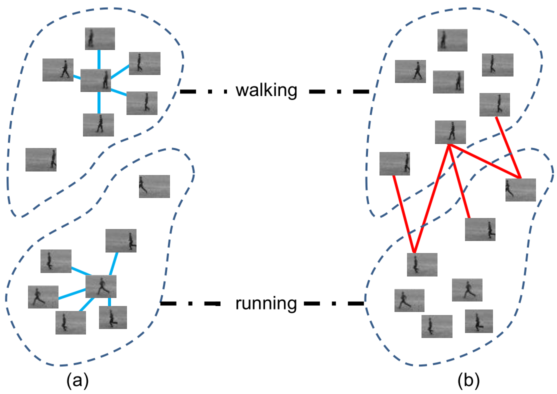

FIGURE 3. An illustration of the single-view (a) intrinsic graph and (b) penalty graph [35].

A. MULTI-VIEW VISUAL FEATURE

Color, shape and texture are the essential visual features [39] which are proven beneficial for content-based image retrieval [40]. However, the above visual feature are not discriminative for some video analysis due to short of motion information. For example, there is no great difference in the appearance between action of 'walking' and 'running', however, variation in velocity forms different motion category. As Figure 2, the action of 'running' achieves a full shot including frame-in and frame-out in the continuous sequence with an interval of five frames. Therefore, the independent views based on color, shape, texture and motion provide complementary information in video description, which is more consistent with human visual machine that multi-view visual features are perceived simultaneously.

For a video $V(x, y, t)$, x and y denote the spatial coordinate component, t is the time coordinate component. To reduce computation complexity, salient regions are detected to represent the key content. In the spatial dimension, the classical methods succeed in detecting interest points with a response function. Extended in spatio-temporal dimension, these methods integrate spatial and temporal response functions to detect video salient regions. According to [14], V can be detected via the response function R, as in (1):

$$R = (V * g * h_{ev})^2 + (V * g * h_{od})^2 \tag{1}$$

where $g(x, y; \sigma_s)$ represents 2D Gaussian smoothing kernel in spatial dimension. h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied along temporal dimension, defined as $h_{ev}(t; \tau_t, \omega_t) = -\cos(2\pi t\omega_t)e^{-t^2/\tau_t^2}$ and $h_{od}(t; \tau_t, \omega_t) = -\sin(2\pi t\omega_t)e^{-t^2/\tau_t^2}$, $\omega_t = 4/\tau_t$, the parameters σ_s and τ_t control the spatial and temporal scale of the detector. According to (1), each interest point roots in the local maxima of the response function, the local regions without distinguishing features cannot recall a response. The salient cuboid concentered on the interest point is extracted with the size of covering most of spatio-temporally contributive microstructure for response function. Therefore, $V(x, y, t)$ can be expressed as a set of local salient cuboids.

The local salient cuboids can preserve spatial structure and temporal consistency. Then multi-view visual features, including brightness, gradient, uniform local binary pattern (ULBP) and optical flow, are used to adequately characterize video content. Brightness [14] is the regularization of original pixel values. Gradient [14] is calculated at each cuboid's location through the channelsx, y, t with the same size as cuboid. ULBP [32] is a simple and efficient method for texture description, which is not only as robust as the classical local binary pattern for illumination, but also contributes to preserve the structure. The 8-neighborhood of each location in the local visual region is carried out the uniform local binary pattern encoding, then the ULBP code for describing texture feature satisfies the range of 0-59. In the methods of instantaneous action description, optical flow occupies the predominant position. Employing the Lucas-Kanade method [33], the variety of corresponding points among adjacent of consistency is calculated in the flow field. Without loss of generality, the gradient and magnitude of velocity are further expressed.

Once multi-view visual features have been extracted, a video-word codebook is built for consistency. In each view m, k-means clustering on dataset feature is performed to form a vocabulary of K different bases. The centroids of clusters are regarded as the symbol of video-word and quantitated the visual features in Euclidean space. Each video sequence can be represented by the co-occurrence histogram of video-words, which models the mutual relation between visual features and the frequency of each bases appeared in the individual video. Then, a nonlinear mapping $\phi(\cdot): X \rightarrow \mathcal{H}$ is performed from the input space X to a high-dimensional feature space $\mathcal{H}$ [34], i.e., $K^m = \phi(X^m) \in \mathbb{R}^{N\times N}$. The multi-view kernel matrix is calculated as $K = \sum_{m=1}^{M} \alpha_m K_m$, s.t. $\sum_{m=1}^{M} \alpha_m = 1$, where α_m denotes the fused weight of the m-th view.

B. DISCRIMINATIVE GRAPH CONSTRUCTURE AND EMBEDDING

The kernel matrix enhances the separability of visual feature, but leading to an exponential growth of feature dimensions. To preserve the structure during the embedding from the high-dimensional feature space, graph theories are adopted to learn the embedding from the multi-view feature with label information. Therefore, a supervised graph is constructed to preserve the structure of high-dimensional data. Following the graph-embedding framework in [35], the single-view intrinsic graph and penalty graph can be defined as (2) and (3), illustrated as Figure 3.

$$G_{w,ij}^{m} = \begin{cases} \exp\left(-\frac{\left\|x_i^m - x_j^m\right\|^2}{\sigma^2}\right), & \text{if } C(x_i^m) = C(x_j^m) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

$$G_{b,ij}^{m} = \begin{cases} \exp\left(-\frac{\left\|x_i^m - x_j^m\right\|^2}{\sigma^2}\right), & \text{if } C(x_i^m) \neq C(x_j^m) \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where, $C(\cdot)$ is the label information. For the m-th view, G_w^m denotes the intrinsic graph and G_b^m denotes penalty graph. Then the multi-view inter-class and intra-class scatter of the related low-dimensional feature are explored as follows:

$$\begin{aligned} S_w &= \sum_{m=1}^{M} \alpha_m \sum_{i=1}^{N} \sum_{j=1}^{N} \left\| y_i^m - y_j^m \right\|^2 G_{w,ij}^m \\ &= tr\left(W^T \phi(X)(D_w - G_w)\phi(X)^T W\right) \end{aligned} \quad (4)$$

$$\begin{aligned} S_b &= \sum_{m=1}^{M} \alpha_m \sum_{i=1}^{N} \sum_{j=1}^{N} \left\| y_i^m - y_j^m \right\|^2 G_{b,ij}^m \\ &= tr\left(W^T \phi(X)(D_b - G_b)\phi(X)^T W\right) \end{aligned} \quad (5)$$

where, y_i is the low-dimensional representation of kernel feature.$D_w = \sum_{m=1}^{M} \alpha_m D_w^m$ and $D_b = \sum_{m=1}^{M} \alpha_m D_b^m$ are the fused multi-view diagonal matrices with the diagonal elements $D_w^m(i,j) = \sum_j S_w(i,j)$ and $D_b^m(i,j) = \sum_j S_b(i,j)$. Let L_w and L_b be the Laplacian matrices of the above graphs, which $L_w = D_w - G_w$ and $L_b = D_b - G_b$. In this stage, we aim to learn the projection matrix W to enhance the discriminative power of low-dimensional feature y_i while the multi-view weights are optimizing. The object function of Multi-view Structure Preserving Embedding can be defined by maximizing the between-class scatter S_b and minimizing the within-class scatter S_w:

$$\begin{aligned} (W, \alpha) &= \arg\max \frac{tr\left(W^T \phi(X)(D_b - S_b)\phi(X)^T W\right)}{tr\left(W^T \phi(X)(D_w - S_w)\phi(X)^T W\right)} \\ &= \arg\max \frac{tr\left(W^T K L_b K W\right)}{tr\left(W^T K L_w K W\right)} \\ &\text{s.t.} \sum_{m=1}^{M} \alpha_m = 1, \quad \alpha_m \geq 0 \end{aligned} \quad (6)$$

C. OPTIMIZATION

According to the fact that multi-view weight selection and feature learning are mutually correlated and reinforce each other, (6) is optimized in an alternative way [31], [36], [37] which divides the original problem into the following sub-problems.

The first subproblem is to update the projection matrix W. In this stage, it is assumed that multi-view weights α_m have already been optimized, then (6) becomes a generalized eigenvalue problem

$$KL_bKW = \lambda KL_wKW \quad (7)$$

where, $[\lambda_1, \ldots, \lambda_d]$ is the d maximal eigenvalues, which are corresponding to the eigenvector $[w_1, \ldots, w_d]$, then the projection matrix $W \in \mathbb{R}^{N \times d}$ can be represented as follows

$$W = \begin{bmatrix} w_{11} & \cdots & w_{1d} \\ \vdots & \ddots & \vdots \\ w_{N1} & \cdots & w_{Nd} \end{bmatrix} \quad (8)$$

The second subproblem is to update the multi-view weights α_m. By fixing the projection matrix W, we derive a relaxed objective function from (6) to optimize α_m as [36]. In (6), $\frac{tr(W^T KL_b KW)}{tr(W^T KL_w KW)} = \left(\frac{tr(W^T KL_w KW)}{tr(W^T KL_b KW)}\right)^{-1}$ which can be maximum via minimizing the maximum critical value of $\frac{tr(W^T KL_w KW)}{tr(W^T KL_b KW)}$. For convenience, let $Lw_{ijk} = tr\left(W^T K_i L_{w,k} K_j W\right)$ and $Lb_{ijk} = tr\left(W^T K_i L_{b,k} K_j W\right)$. Taking two views as examples, the maximum of $\frac{tr(W^T KL_w KW)}{tr(W^T KL_b KW)}$ can be calculated by the Cauchy-Schwarz inequality (9), as shown at the bottom of the next page, where, $w_{ijk}(\cdot)$ is the function with variables α_1 and α_2. Because of $\alpha_1 + \alpha_2 = 1$, the weights with $\alpha_1^2\alpha_2$ and $\alpha_1\alpha_2^2$ are always smaller than a constant. Then (9) can be converted to solve the coefficient η of $\frac{Lw_{ijk}}{Lb_{ijk}}$, as in (10),

$$\begin{aligned} &\arg\min_{\alpha_1,\alpha_2} \eta_1^r \frac{Lw_{111}}{Lb_{111}} + \eta_2^r \frac{Lw_{222}}{Lw_{111}}, \\ &\text{s.t.} \eta_1 + \eta_2 = 1, \quad \eta_1, \eta_2 \geq 0 \end{aligned} \quad (10)$$

Let $\eta \leftarrow \eta^r$, and $r > 1$, to avoid the situation that only a single-view can be selected. For multi-view, the general form of (10) as follows,

$$\arg\min_{\eta_1,\ldots,\eta_M} \sum_{i=1}^{M} \eta_i^r \frac{Lw_{iii}}{Lb_{iii}}, \text{ s.t.} \sum_{i=1}^{M} \eta_i = 1, \eta_i \geq 0 \quad (11)$$

With the Lagrangian multiplier η,

$$J(\beta, \zeta) = \sum_{i=1}^{M} \eta_i^r \frac{Lw_{iii}}{Lb_{iii}} - \zeta \left(\sum_{i=1}^{M} \eta_i - 1\right) \quad (12)$$

Let the derivatives of $J(\eta, \zeta)$ with respect to η_i and ζ, i.e.

$$\begin{cases} \frac{\partial J(\eta,\zeta)}{\partial \eta_i} = r\eta_i^{r-1} \frac{Lw_{iii}}{Lb_{iii}} - \zeta = 0 \\ \frac{\partial J(\eta,\zeta)}{\partial \zeta} = \sum_{i=1}^{M} \eta_i - 1 = 0 \end{cases} \quad (13)$$

According to (13),

$$\eta_i = \frac{(Lb_{iii}/Lw_{iii})^{\frac{1}{r-1}}}{\sum_{j=1}^{M} (Lb_{iii}/Lw_{iii})^{\frac{1}{r-1}}}, \quad i = 1, \ldots, M \quad (14)$$

Since $\frac{\eta_i^r}{\eta_j^r} = \frac{\alpha_i^3 Lw_{iii}}{\alpha_j^3 Lw_{jjj}}$, (15) can update α,

$$\alpha_i = \frac{(\eta_i^r / Lw_{iii})^{\frac{1}{3}}}{\sum_{j=1}^{M} (\eta_i^r / Lw_{iii})^{\frac{1}{3}}}, \quad i = 1, \ldots, M \quad (15)$$

Algorithm 1 MSPE-Based Video Representation

Input: video dataset $V = \{V_1, \ldots, V_i, \ldots V_j, \ldots, V_N\}$ and parameters K,σ, r, d

Output: the projection matrix W and multi-view weight vector α

1: Extract multiple features X, and perform a nonlinear mapping to $X \rightarrow \mathcal{H}$

2: Compute the Laplacian matrices L_w^m and L_b^m via inter-class and intra-class scatter for each view

3: Initialize $\alpha_m = 1/M$, $t = 0$

4: **Repeat**

5: Compute the multi-view kernel matrix $K = \sum_{m=1}^{M} \alpha_m K_m$ and the multi-view Laplacian matrix $L_w = \sum_{m=1}^{M} \alpha_m L_w^m$: $L_b = \sum_{m=1}^{M} \alpha_m L_b^m$

6: Update projection matrix $(W)_t$ by Eq. (7)

$$KL_bK(W)_t = (\lambda)_t KL_wK(W)_t$$

7: Update weight vector $(\alpha)_t$ by (14) and (15)

$$(\alpha_i)_t = \frac{((\eta_i^r)_t / Lw_{iii})^{\frac{1}{3}}}{\sum_{j=1}^{M} ((\eta_i^r)_t / Lw_{iii})^{\frac{1}{3}}}$$

8: $t = t + 1$

9:**Until** Eq. (6) converges or $t = t_{max}$

The third subproblem is about initialization and post-processing. In the first round of iteration, the multi-view weight α can be assumed that the efforts of each view information are equal and initialized as $\alpha_m = 1/M, \forall m = 1, \ldots, M$. Fast convergence and accurate performance of the algorithm is benefit of the initialized parameter selection with some prior knowledge. After several round of iterations referred to the above alternative optimization, the iteration stops when the number of iterations reaches the maximum t_{max} or the iteration converges. Then the final multi-view discriminative feature can be represented in the low-dimensional space with the fused weights.

The proposed video representation based on Multi-view Structure Preserving Embedding is in Algorithm 1.

III. EXPERIMENTS

In this section, a series of experiments based on the proposed method for content-based video retrieval are carried out. To provide a fair experimental environment, the benchmark datasets and evaluation indicators are introduced in the front. Meanwhile, we provide the related parameter analysis and performance comparison to analyze the superiority of our method. All experiments are in Matlab and executed on a computer with Intel Corei7-2600CPU @ 3.40 GHz and 64 GB physical memory.

Algorithm 2 MSPE-based Retrieval

Input: the query video V_q, the projection matrix W and multi-view weight vector α

Output: retrieval result

1: Extract multiple features X_q, and perform the nonlinear mapping

2: With α, compute the fused kernel matrix K_q

3: With W, generate the low-dimensional feature representation $Y'_q = K_q W$

4: Measure the similarity and rank

5: Feedback videos from database

A. DATASETS AND EVALUATION INDICATORS

For the query video, the retrieval process aims to find out its similar targets belonging to the same category. A representative and discriminative feature representation is conducive to enhance the retrieval performance. For the proposed MSPE-based retrieval framework, the first task is to learn the optimal projection matrix and multi-view weights. Then, each video clip, including the query and the clips in the dataset, can be represented in the low-dimensional space. The similarity between low-dimensional features can be computed with Euclidean distance and ranked in descending order. The clips in the dataset with high similarity score fall into the same category of the query in theory and are feedback to the user. The above retrieval process can be illustrated as Algorithm 2.

The performance of the proposed method is evaluated on several public benchmarks: FACE dataset [14], KTH dataset [13], UCF YouTube dataset [38] and MOUSE dataset [14], which provide abundant video content including facial emotion, human action and animal behavior. The FACE dataset is consisted of 6 expressions, i.e. anger, disgust, fear, joy, sadness and surprise, which are collected from 2 individuals under 2 lighting setups. There are 598 clips and 1168 clips in KTH dataset and UCF YouTube dataset respectively.

$$\begin{aligned}\frac{tr(W^T KL_wKW)}{tr(W^T KL_bKW)} &= \frac{tr(W^T(\alpha_1K_1+\alpha_2K_2)(\alpha_1L_{w,1}+\alpha_2L_{w,2})(\alpha_1K_1+\alpha_2K_2)W)}{tr(W^T(\alpha_1K_1+\alpha_2K_2)(\alpha_1L_{b,1}+\alpha_2L_{b,2})(\alpha_1K_1+\alpha_2K_2)W)}\\ &= \frac{\alpha_1^3Lw_{111}+2\alpha_1^2\alpha_2Lw_{121}+\alpha_1\alpha_2^2Lw_{221}+\alpha_1^2\alpha_2Lw_{112}+2\alpha_1\alpha_2^2Lw_{122}+\alpha_2^3Lw_{222}}{\alpha_1^3Lb_{111}+2\alpha_1^2\alpha_2Lb_{121}+\alpha_1\alpha_2^2Lb_{221}+\alpha_1^2\alpha_2Lb_{112}+2\alpha_1\alpha_2^2Lb_{122}+\alpha_2^3Lb_{222}}\\ &\le \sum_{i,j,k\in\{1,2\}} w_{i,j,k}(\alpha_1,\alpha_2)\frac{Lw_{ijk}}{Lb_{ijk}}\end{aligned} \tag{9}$$

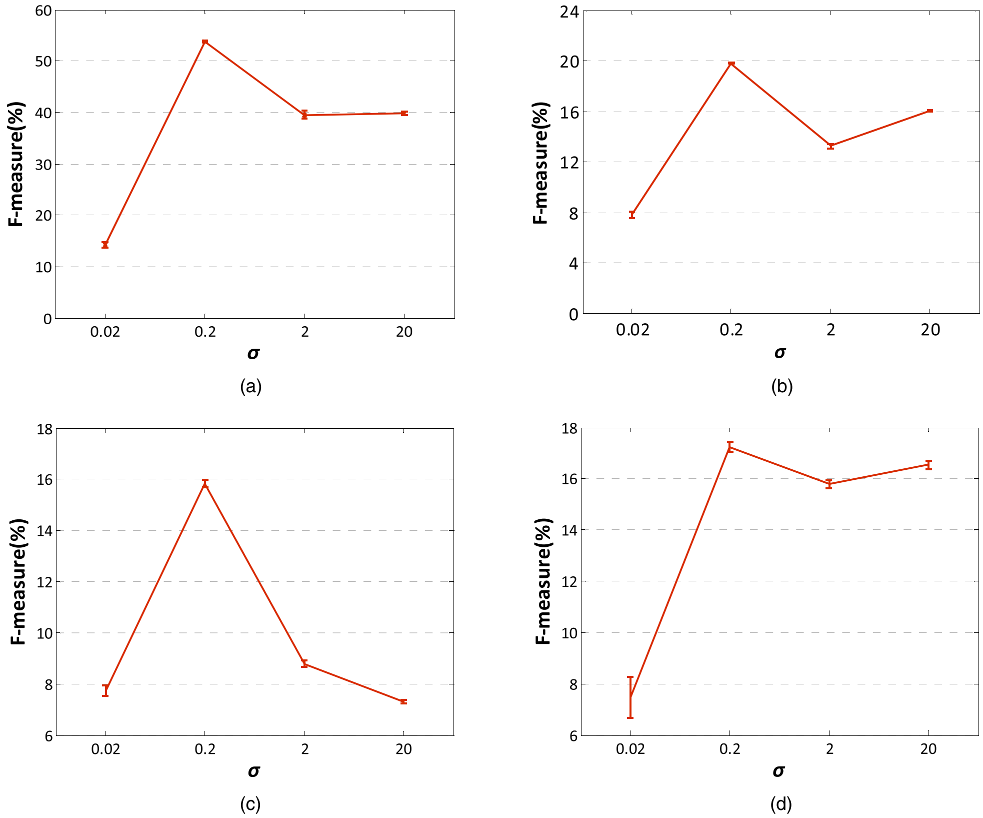

FIGURE 4. F-measure under various hot kernel parameter σ on dataset: (a) FACE, (b) KTH, (c) UCF YouTube, and (d) MOUSE.

KTH dataset records 25 individuals engaged in the activities: walking, jogging, clapping, waving, boxing and running. UCF YouTube dataset groups into 11 action categories: basketball shooting, biking, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking and walking with a dog. The MOUSE dataset involves 406 clips filmed a same mouse at different points a day with various behaviors like drinking, eating, exploring, grooming and sleeping. The above datasets are challenging due to large variations in camera motion, object appearance, action type, sample size, etc.

During the retrieval process, each video of the database is regarded as the query according to the leave-one-out rule and recalls 12 similar clips. The test is repeated 10 times. The final result is the mean value of 10 times. Precision-recall curve (PR curve) is adopted to analyze the user satisfaction to the retrieval results. The effectiveness is evaluated in terms of Precision, Recall, and F-measure, computed as follows:

$$Precision = \frac{1}{N_r}\sum_{n=1}^{N_r}\psi\left(C\left(V_q\right), C\left(V_n\right)\right) \tag{16}$$

$$Recall = \frac{1}{N_b}\sum_{n=1}^{N_b}\psi\left(C\left(V_q\right), C\left(V_n\right)\right) \tag{17}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{18}$$

where, $\psi\left(C\left(V_q\right), C\left(V_n\right)\right) = \begin{cases} 1, & C\left(V_q\right) = C\left(V_n\right) \\ 0, & else \end{cases}$,

$C(\cdot)$ represents the label information of the query V_q and feedback V_n, N_r is the total feedback number to user, N_b is the total number of similar actions in the dataset.

B. PARAMETER ANALYSIS

Reasonable parameters selection can promote the discriminability of feature representation. There are three parameters playing a great role in the MSPE-based retrieval method, i.e. hot kernel parameter, view parameter and dimensionality scale. In this subsection, we discuss and analyze these parameters by the means of grid search and cross validation. Then, the optimum parameter is selected according to the mean and variance of F-measure in retrieval results of 10 times.

The hot kernel parameter σ occurs in the graph construction and determines the attenuation rate of similarity function. As in (2) and in (3), the relationship of any pair of samples weakens with increasing σ. When $\sigma \to \infty$, the weight of each view is almost the same, so that the intrinsic graph and penalty graph easily blend together. Therefore, a relative small σ value represents the weight difference which can preserve the inter-class compactness and the intra-class separability. However, $\sigma \to 0$ is forbidden due to the constraint of discriminative structure. In the proposed framework, $\sigma = 0.2$ is selected according to the outstanding F-measure illustrated in Figure 4.

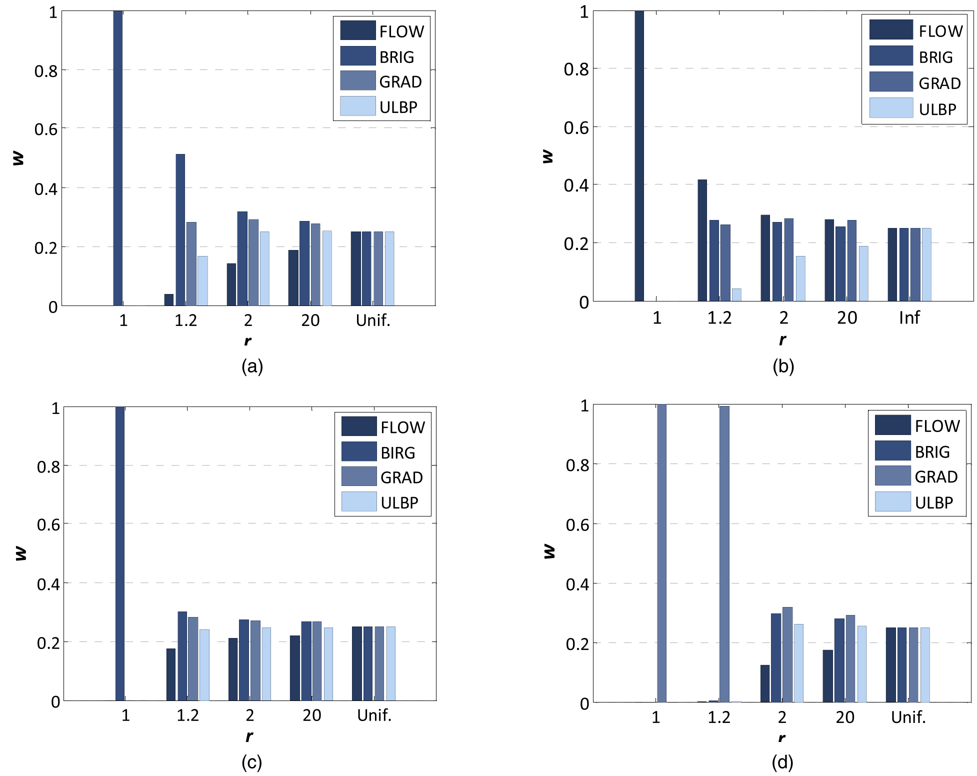

FIGURE 5. Weight distribution under various view parameter *r* on dataset: (a) FACE, (b) KTH, (c) UCF YouTube, and (d) MOUSE.

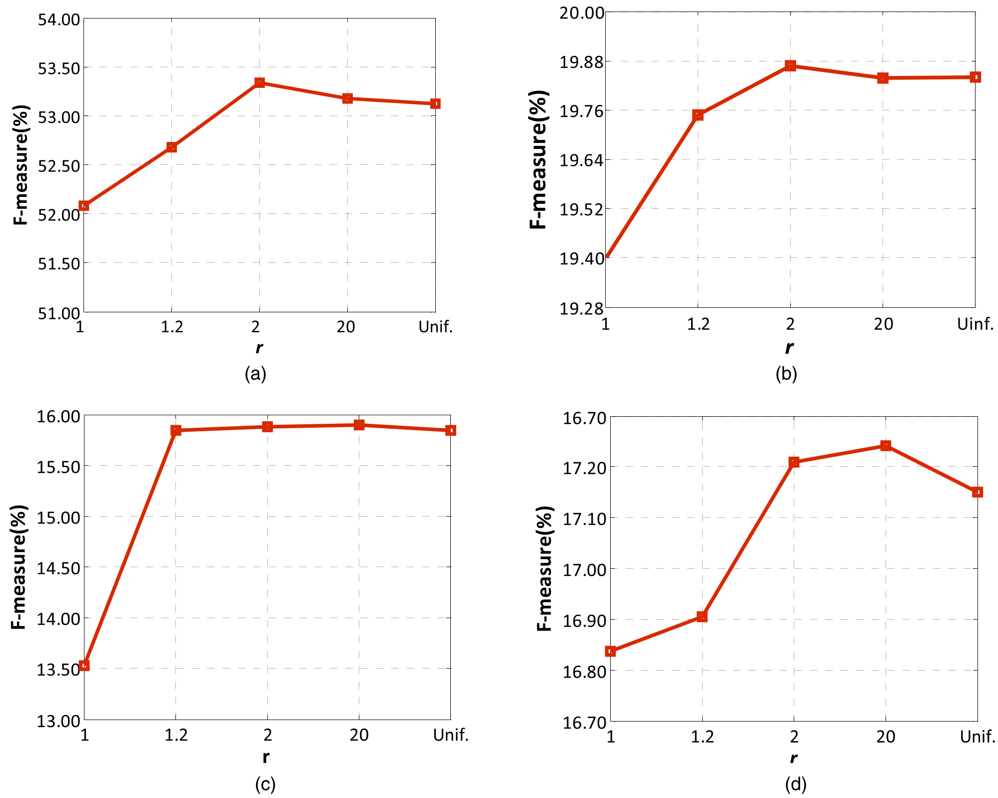

FIGURE 6. F-measure under various view parameter *r* on dataset: (a) FACE, (b) KTH, (c) UCF YouTube, and (d) MOUSE.

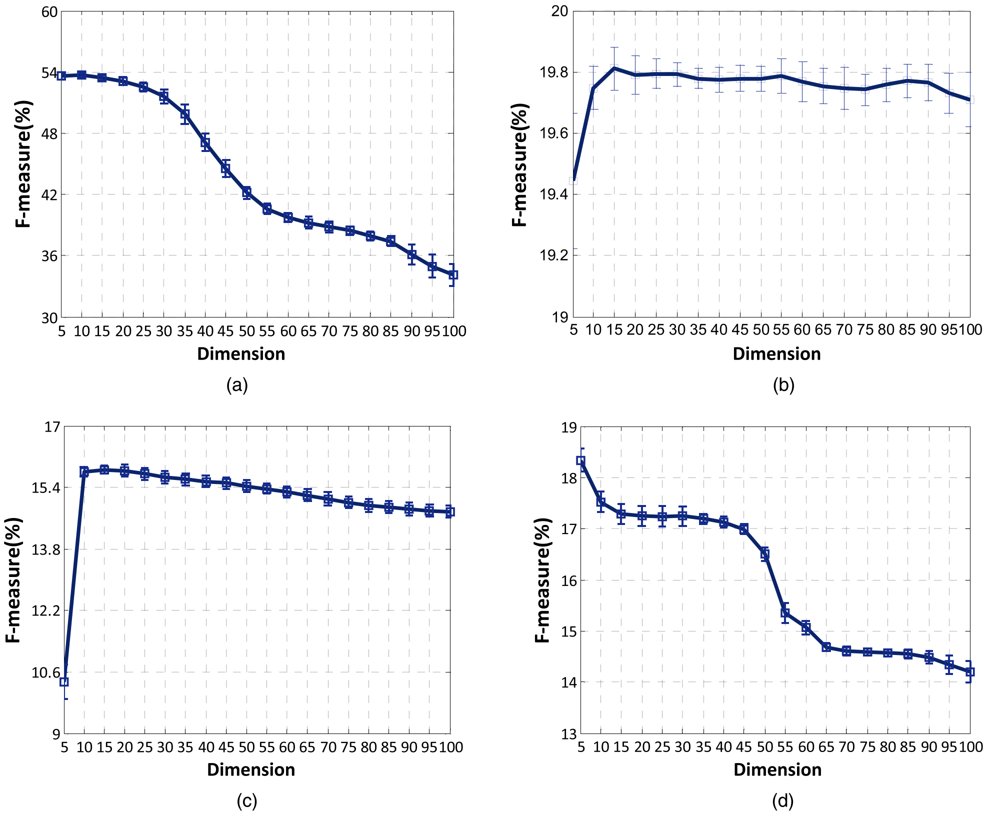

FIGURE 7. F-measure under various dimensionality scale d on dataset: (a) FACE, (b) KTH, (c) UCF YouTube, and (d) MOUSE.

TABLE 1. Performance comparison on face dataset.

Performance	BRIG	GRAD	ULBP	FLOW	MFVS	KMP	MSPE
Percision(%)	75.46	75.07	72.42	62.44	75.79	93.83	97.33
Recall(%)	28.30	28.15	27.16	23.41	28.42	35.19	36.50

To guarantee the low-dimensional feature representation with multi-view information, we introduce the view parameter r to control the weight coefficient of fused view. Combined (14) and (15), the relation of the view and r can be expressed as in (19). If $r = 1$, the completely sparse solution will reduce the complementary information from multi-view, even leading to single-view choice, illustrated as in (20). With the increasing r, the i-th view will get a smaller inter- class scatter and a larger intra-class scatter than the j-th, that means the i-th view is more important for content description. When $r \to \infty$, $\alpha_i/\alpha_j \to 1$, i.e. the view difference is suppressed. Therefore, the view parameter r is determined by the rule of both independence and complementarity. Some priori knowledge for r choice guides the weight optimization towards rapid convergence. Figure 5 and Figure 6 illustrate the F-measure of single-view, multi-view and average view, indicating that $r = 2$ satisfies the proposed method.

$$\frac{\alpha_i}{\alpha_j} = \left(\frac{\eta_i^r}{\eta_i^r} \cdot \frac{Lw_{jjj}}{Lw_{iii}}\right)^{\frac{1}{3}}$$

$$= \left(\left(\frac{Lb_{iii}}{Lb_{jjj}}\right)^{\frac{r}{r-1}} \cdot \left(\frac{Lw_{jjj}}{Lw_{iii}}\right)^{\frac{r+1}{r-1}}\right)^{\frac{1}{3}}, \quad r > 1 \tag{19}$$

$$\alpha_i = \begin{cases} 1, & i = \arg\max_{i'} \left(\frac{Lw_{iii'}}{Lb_{iii'}}\right), \\ 0, & \text{otherwise} \end{cases} \quad r = 1 \tag{20}$$

During the generalized eigenvalue decomposition in (7), the dimensionality scale d is immediately in charge of the number of eigenvalues and the size of projection matrix. Low-dimensional and high-discriminative feature representation is the key for efficiency retrieval. As the Figure7, the dimensionality scale d is varying in the range of 5 to 100. The feature redundancy will weaken the discrimination.

C. EXPERIMENTAL RESULTS

In order to test the performance of the proposed method applied to content-based action retrieval task, the classified retrieval is discussed on the benchmarks and illustrated in Figure 8. Each PR curve indicates the retrieval results belonging to the same category of the query. Theoretically,

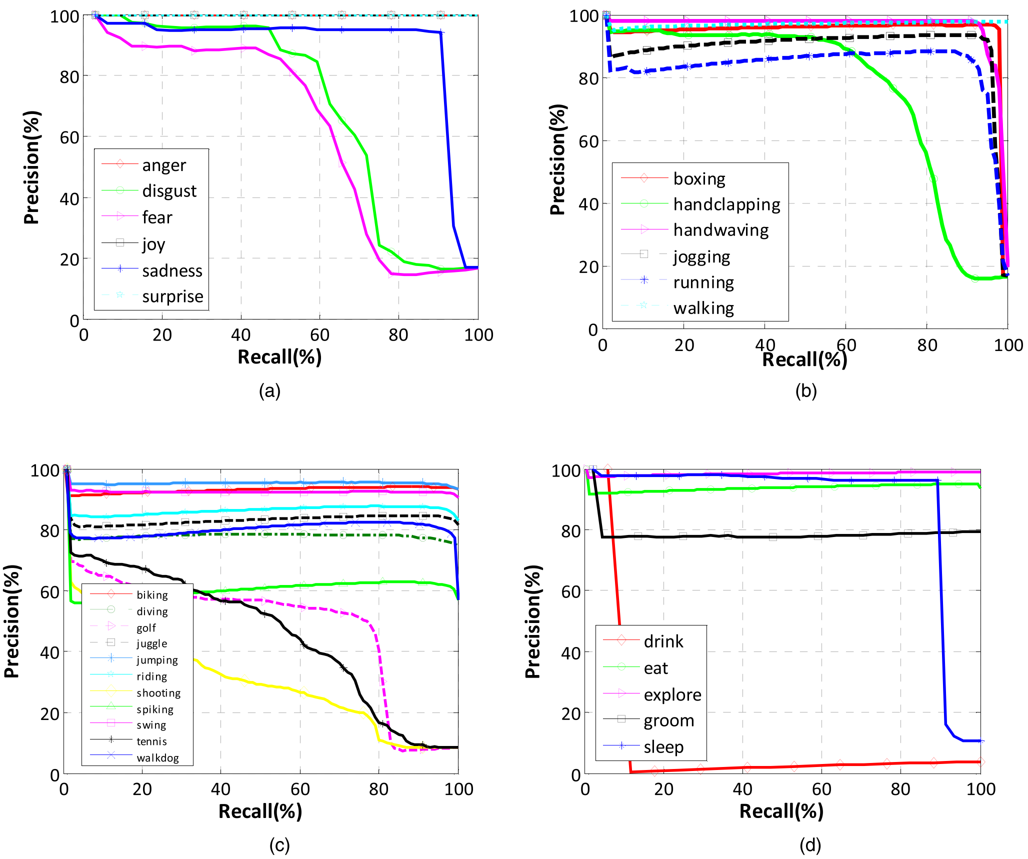

FIGURE 8. PR curve of classified retrieval on dataset: (a) FACE, (b) KTH, (c) UCF YouTube, and (d) MOUSE.

TABLE 2. Performance comparison on kth dataset.

Performance	BRIG	GRAD	ULBP	FLOW	MFVS	KMP	MSPE
Percision(%)	74.31	72.59	50.09	72.85	73.60	90.82	92.17
Recall(%)	8.95	8.74	6.03	8.77	8.86	10.94	11.10

TABLE 3. Performance comparison on ucf youtube dataset.

Performance	BRIG	GRAD	ULBP	FLOW	MFVS	KMP	MSPE
Percision(%)	36.52	37.89	32.17	35.31	38.63	73.82	78.49
Recall(%)	3.98	4.16	3.51	3.88	4.24	8.27	8.82

TABLE 4. Performance comparison on mouse dataset.

Performance	BRIG	GRAD	ULBP	FLOW	MFVS	KMP	MSPE
Percision(%)	77.72	78.97	78.88	78.90	82.68	87.04	89.13
Recall(%)	9.07	9.38	9.32	9.53	10.17	9.75	10.04

precision is proportional to recall like the category 'anger', 'joy' and 'surprise' in Figure 8(a). However, the feedback recalls some negative samples leading to a lower precision, such as the curve of 'boxing' in Figure 8(b) and 'tennis' in Figure 8(c). Meanwhile, the imbalance category number aggravates a worse performance when the feedback number is fixed, for example, 'drink' in Figure 8(d) with a total of 17 samples. In general, the MSPE-based retrieval method has the advantage of feedback with the same category in earlier.

To further demonstrate the performance, we compare MSPE with the state-of-the-art. In Table 1 to Table 4, the

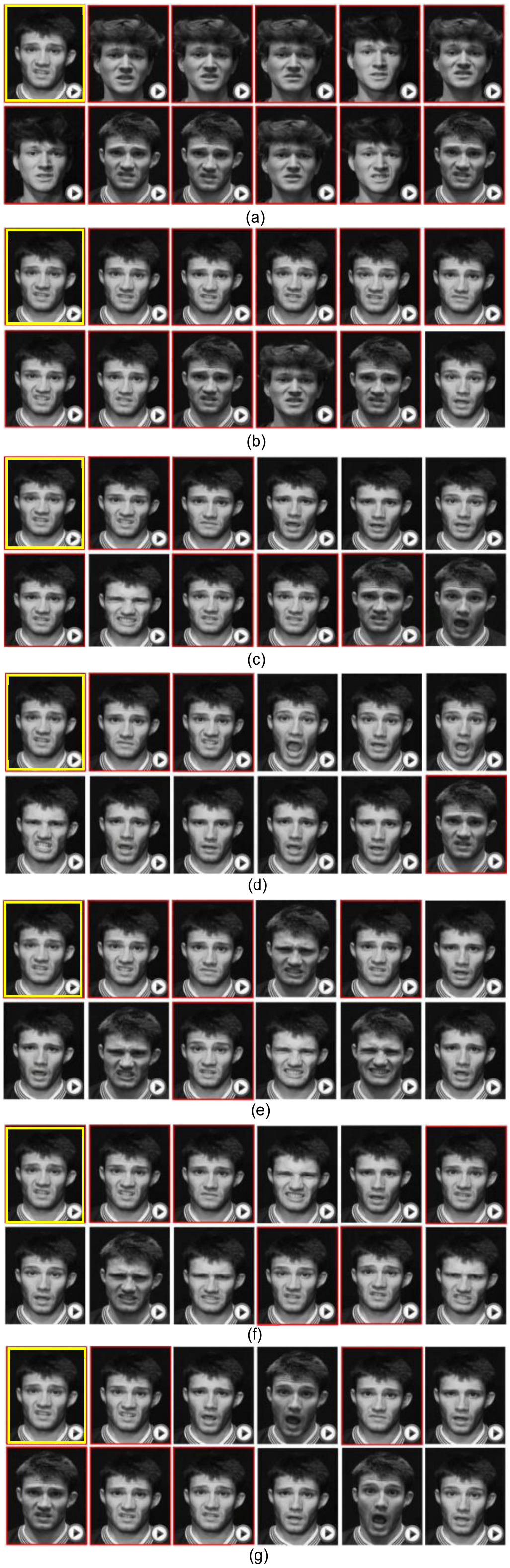

FIGURE 9. Feedback of the query ‘disgust’ with various method: (a) MSPE, (b) KMP, (c) MFVS, (d) ULBP, (e) GRID, (f) BRIG, and (g) FLOW. (Note: the query with yellow box and the exact retrieval results with red box.)

retrieval based on BRIG, GRAD, ULBP, FLOW, MFVS, KMP and MSPE are illustrated. The single-view method BRIG, GRAD, ULBP and FLOW accounts for brightness, appearance, texture and motion information [14]. In fact, different content in the real data arouses different visual attention. For example, the movements of facial expressions are less drastic than body action so that FLOW holds low efficacy on FACE and inversely on MTH. As mentioned in Figure 5, the above priori knowledge is conducive to the initialization of view parameter to reach fast convergence of the iteration. MFVS stands for multi-view feature vector splicing which averages view contribution. Although MFVS combines multi-view information, the redundancy feature is emerged. However, the multi-view method is still superior to the single-view, especially for UCF YouTube dataset with the real scene in Table 3. During the graph embedding, KMP only preserves the inter-class structure. Therefore, MSPE on basis of complementary view, adaptive view weight scheme and the favorable structure preserving embedding has a more discriminative low-dimensional feature than KMP with the same parameter setting.

Figure 9 provides the visual comparison of MSPE and the others with the query ‘disgust’. We receive 12 clips and evaluate the retrieval result according to the GroundTruth. In Figure 9, the feedback of MSPE-based retrieval method are all belonging to the same category of the query, but the other method only recall partial positive samples and multi-view methods have advantage over the single-view. In particular, MSPE-based method utilizes complementary vision and structure information to realize cross-individual expression retrieval.

IV. CONCLUSION

In this paper, the Multi-view Structure Preserving Embedding module (MSPE) is proposed for video representation in content-based action retrieval task. MSPE is a unified framework in which both multi-view analysis and structure preserving embedding are explored. Brightness, appearance, texture and motion information are jointly described but interdependently represented in the defined adaptive way. Meanwhile, mining the multi-view inter-class and intra-class relationship preserves the internal structure of high-dimensional data during graph embedding. Then, a low-dimensional and high-discriminative feature can be carried out in the retrieval task. Although the experimental datasets are too abundant to challenging, experiments results indicated that MSPE-based method can retrieve precisely and comprehensively and have a high performance as being satisfactory. Since most online customer videos hold seldom useful labels, a semi-supervised/unsupervised MSPE method will be extended in the future works.

REFERENCES

[1] M. Ramezani and F. Yaghmaee, “A review on human action analysis in videos for retrieval applications,” *Artif. Intell. Rev.*, vol. 46, no. 4, pp. 485–514, Dec. 2016.

[2] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache, “A smart atlas for endomicroscopy using automated video retrieval,” *Med. Image Anal.*, vol. 15, no. 4, pp. 460–476, 2011.

[3] X. Sun *et al.*, “Place retrieval with graph-based place-view model,” in *Proc. ACM Conf. Multimedia Inf. Retr.*, 2008, pp. 268–275.

[4] Y. Xu, B. Ma, R. Huang, and L. Lin, "Person search in a scene by jointly modeling people commonness and person uniqueness," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 937–940.

[5] J. Qin, L. Liu, M. Yu, Y. Wang, and L. Shao, "Fast action retrieval from videos via feature disaggregation," *Comput. Vis. Image Understand.*, vol. 156, pp. 104–116, Mar. 2017.

[6] S. Jones and L. Shao, "Content-based retrieval of human actions from realistic video databases," *Inf. Sci.*, vol. 236, no. 1, pp. 56–65, Jul. 2013.

[7] R. Ji, H. Yao, and X. Sun, "Actor-independent action search using spatiotemporal vocabulary with appearance hashing," *Pattern Recognit.*, vol. 44, no. 3, pp. 624–638, 2011.

[8] L. Shao, S. Jones, and X. Li, "Efficient search and localization of human actions in video databases," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 3, pp. 504–512, Mar. 2014.

[9] S. Jones and L. Shao, "A multigraph representation for improved unsupervised/semi-supervised learning of human actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 820–826.

[10] A. Ciptadi, M. S. Goodwin, and J. M. Rehg, "Movement pattern histogram for action recognition and retrieval," in *Proc. IEEE Conf. Eur. Conf. Comput. Vis. (ECCV)*, vol. 8690, 2014, pp. 695–710.

[11] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. IEEE Conf. Comput. Vis. ICCV)*, Oct. 2003, vol. 64, nos. 2–3, pp. 432–439.

[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[13] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit.*, Aug. 2004, vol. 3, no. 17, pp. 32–36.

[14] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. IEEE Int. Workshop Vis. Surveill. Perform. Eval. Tracking Surveill.*, Oct. 2005, pp. 65–72.

[15] F. Paez, J. A. Vanegas, and F. A. Gonzalez, "An evaluation of NMF algorithm on human action video retrieval," in *Proc. Symp. Signals, Images Artif. Vis. (STSIVA)*, 2013, pp. 1–4.

[16] Y. Jia, M. Salzmann, and T. Darrell, "Factorized latent spaces with structured sparsity," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2010, pp. 982–990.

[17] F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[18] M. Yang, L. Zhang, X. Feng, and D. Zhang, "Fisher discrimination dictionary learning for sparse representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nov. 2011, pp. 543–550.

[19] X. C. Lian, Z. Li, B. L. Lu, and L. Zhang, "Max-margin dictionary learning for multiclass image categorization," in *Proc. IEEE Conf. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2010, pp. 157–170.

[20] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," *J. Mach. Learn. Res.*, vol. 12, pp. 3371–3412, Jan. 2011.

[21] Y. Yi, J. Wang, W. Zhou, C. Zheng, J. Kong, and S. Qiao, "Non-negative matrix factorization with locality constrained adaptive graph," *IEEE Trans. Circuits Syst. Video Technol.*, to be published. doi: 10.1109/TCSVT.2019.2892971.

[22] Y. Yi, S. Qiao, W. Zhou, C. Zheng, Q. Liu, and J. Wang, "Adaptive multiple graph regularized semi-supervised extreme learning machine," *Soft Comput.*, vol. 22, no. 11, pp. 3545–3562, Jun. 2018.

[23] H. Zhang, N. M. Nasrabadi, Y. Zhang, and T. S. Huang, "Joint dynamic sparse representation for multi-view face recognition," *Pattern Recognit.*, vol. 45, no. 4, pp. 1290–1298, 2012.

[24] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14, 2002, pp. 585–591.

[25] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, vol. 16, no. 1, pp. 186–197.

[26] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2. Oct. 2005, pp. 1208–1213.

[27] Q. Yin, S. Wu, R. He, and L. Wang, "Multi-view clustering via pairwise sparse subspace representation," *Neurocomputing*, vol. 156, pp. 12–21, May 2015.

[28] Y.-M. Xu, C.-D. Wang, and J.-H. Lai, "Weighted multi-view clustering with feature selection," *Pattern Recognit.*, vol. 53, pp. 25–35, May 2016.

[29] X. Cai, F. Nie, and H. Huang, "Multi-view k-means clustering on big data," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 2598–2604.

[30] G. Tzortzis and A. Likas, "Kernel-based weighted multi-view clustering," in *Proc. IEEE 12th Int. Conf. Data Mining (ICDM)*, Dec. 2012, vol. 5, no. 1, pp. 675–684.

[31] L. Shao, L. Liu, and M. Yu, "Kernelized multiview projection for robust action recognition," *Int. J. Comput. Vis.*, vol. 118, no. 2, pp. 115–129, 2016.

[32] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[33] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Int. Joint Conf. Artif. Intell.*, 1981, vol. 73, no. 3, pp. 674–679.

[34] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering," *Pattern Recognit.*, vol. 41, no. 1, pp. 176–190, 2008.

[35] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[36] J. C. Bezdek and R. J. Hathaway, "Some notes on alternating optimization," in *Advances in Soft Computing* (Lecture Notes in Computer Science), vol. 2275, no. 4. London, U.K.: Springer-Verlag, 2002, pp. 288–300.

[37] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and Gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1715, Oct. 2007.

[38] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, vol. 38, no. 10, pp. 1996–2003.

[39] M. Zhang, Y. Pang, Y. Wu, Y. Du, H. Sun, and K. Zhang, "Saliency detection via local structure propagation," *J. Vis. Commun. Image Represent.*, vol. 52, pp. 131–142, Apr. 2018.

[40] K. Zhang, F. Zhang, J. Lu, Y. Lu, J. Kong, and M. Zhang, "Local structure co-occurrence pattern for image retrieval," *J. Electron. Imag.*, vol. 25, no. 2, 2016, Art. no. 023030.

KE ZHANG was born in Jilin, China, in 1988. She received the B.S. degree in computer science and information engineering from Harbin Normal University, China, in 2011, and the Ph.D. degree from the School of Computer Science and Information Technology, Northeast Normal University, in 2010. She is currently a Lecturer with the School of Computer Science and Technology, Changchun University of Science and Technology. Her research interests include feature learning, saliency detection, image retrieval, and action recognition.

HUI SUN received the B.E. degree from the Computer School, Jilin University, China, in 2005, and the M.S. degree from the College of Computer Science and Information Technology, Northeast Normal University, in 2010. Her research interests include feature learning and motion analysis.

WEILI SHI was born in Jiangsu, China, in 1981. He received the M.S. and Ph.D. degrees from the Changchun University of Science and Technology, China, in 2012 and 2017, respectively. He is an Associate Professor and a Researcher with the Medical Image Computing Laboratory, Jilin. His research interests include human behavior analysis, computer-aided diagnosis, and medical image retrieval.

YUWEN FENG was born in Jilin, China. She currently studies and researches with Jilin University, China. Her research interests include intelligent computation, information processing, feature extraction, and behavior analysis.

ZHENGANG JIANG received the B.S. and M.S. degrees from the Changchun University of Science and Technology, China, in 1997 and 2000, respectively, and the Ph.D. degree from Nagoya University, Japan. He is currently a Professor and a Doctoral Tutor with the School of Computer Science and Technology, Changchun University of Science and Technology. His research interests include artificial intelligence, machine learning, and computer-aided diagnosis.

JIANPING ZHAO received the degree from the Changchun University of Science and Technology, China, where he is currently a Professor and a Doctoral Tutor with the School of Computer Science and Technology. His research interests include machine learning, video understanding, handwriting identification, medical image processing, and data mining.