Low-Rank Regularized Multimodal Representation for Micro-Video Event Detection

JING ZHANG, YUTING WU, JINGHUI LIU, PEIGUANG JING, AND YUTING SU
School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
Corresponding author: Peiguang Jing (pgjing@tju.edu.cn)

ABSTRACT Currently, micro-videos are becoming one of the most representative products in the new media age. Although the length of micro-videos is limited to cater to the fast pace of life and are beneficial for rapid distribution, micro-videos are usually recorded in specific scenarios and tend to convey relatively complete events. To more accurately obtain the event types of micro-videos to facilitate potential applications, we propose a low-rank regularized multimodal representation method for micro-video event detection. To solve the less descriptive power of each modality, the latent common representation of micro-videos is obtained by exploiting complementarity among modalities. A considerable gain in accuracy on this basis can be achieved by further considering the low-rank constraint for the lowest-rank intrinsic representation and a flexible label-relaxation strategy for mappings between representations and their correspondences. A newly constructed micro-video dataset is used to verify the advantages of our proposed model. The experimental results demonstrated the superior performance of our proposed method compared with state-of-the-art methods.

INDEX TERMS Micro-video, event detection, multimodal, low-rank representation.

I. INTRODUCTION

Currently, with the rapid development of communication technology and portable devices, micro-videos are becoming a new trend of user-generated content from social media platforms. Compared to traditional forms of text and picture, micro-videos offer a more lively style for recording and exhibiting people’s lives. Micro-videos are first produced with the aim of exploiting the fragmented time of users, and thus, have a very limited time length. For example, TikTok and Instagram are well known micro-video platforms, prescribing a limit of 15 seconds at most for ordinary users.

The study of micro-videos becomes a promising research field attracting increasing attention. Several pioneering efforts have been made to ease the problem of micro-video intelligence analysis. For example, Redi et al. [1] attempted to understand the computational features that are correlated with the creativity of micro-videos. Cao et al. [2] proposed exploiting an attention mechanism and proposed a neural collaborative filtering model to address the problem of micro-video group recommendation. Chen et al. [3] introduced a transductive multimodal learning model to predict micro-video popularity, in which the modality relatedness and insufficiency limitation can be alleviated by finding a common shared subspace. To better estimate the venue category of micro-videos, Nie et al. [4] introduced a deep transfer model to enhance the semantic representations of micro-videos by exploiting external sound knowledge. Liu et al. [5] developed a tree-guided multimodal dictionary learning algorithm to solve the organization problem of online micro-videos. Wei et al. [6] presented a relation-aware driven neural multimodal cooperative learning model to characterize the complex correlations among different modalities. In addition, the analysis of event semantics of micro-videos also plays a significant role in many potential applications, such as automatic annotation, retrieval, and supervision. Nevertheless, very few studies have attempted to solve the event detection problem of micro-videos.

Several limitations could explain the shortage of relevant studies. Firstly, although there are already some relevant work on event recognition in traditional videos and still...
images [7], [8] [9], [10] [11], [12] [13], [14], which reveal high-level semantic cues, such as semantics of object and scene, human poses, and garments, can provide informative priors for event detection. For example, Wang et al. [15] presented an approach for event driven web video summarization by tag localization and key-shot mining. Hong et al. [16] proposed a novel scheme that is able to summarize the content of video search results by mining and threading “key” shots, such that users can get an overview of main content of these videos at a glance. Zhao et al. [17] proposed to recognize video emotions in end-to-end manner based on convolutional neural networks(CNNs). Zhao et al. [18] introduced a novel real-time event detection method by generating an intermediate semantic level from social multimedia data, named microblog clique (MC), which is able to explore the high correlations among different microblogs. As to micro-videos, it not only involves similar integration and interaction of multiple high-level semantics but also exhibits more serious variations across visual appearance and spatial structure, even for the same event category. Specially, a video event usually encompasses the continuous behavior with continuous variations that do not only evolve in temporal domain, but also involve the interactions among multiple objects in spatial domain. Compared with the static concept of objects and scenes, the representation learning of the video events tend to be more complicated. Secondly, due to the short nature of time length and certain practical factors, such as postproduction and editing, irrelevant or noisy information is usually embedded in the source micro-videos, making the valuable cues hidden in the original representations. Furthermore, another significant obstacle to the study of micro-video event detection is the lack of publicly available datasets.

To address these limitations, in this paper, we propose a low-rank regularized multimodal representation learning method for micro-video event detection. We first characterize micro-videos by multiple modalities and eliminate the heterogeneity and redundant information among modalities by maximizing the correlation among different modalities, inspired by some recent low-rank representation [19], [20] [21], [22] [23], [24] and multimodal [25], [26] methods, a considerable gain in detection accuracy then is achieved by further exploring the low-rank representation and the label-relaxed strategies [27]. Specifically, the lowest-rank representation shared by all modalities that captures the intrinsic information is learned by enforcing a low-rank constraint. Meanwhile, the label-relaxed strategy that provides more freedom for the mapping from feature to label spaces is developed to relax the strict binary label assignment into a soft constraint. Moreover, we constructed a newly micro-video event detection dataset to verify the effectiveness of our proposed method.

The rest of this paper is organized as follows. Our proposed algorithm is described in Section 2. The performance evaluation of our proposed algorithm is discussed in Section 3, followed by the conclusion in Section 4.

II. PROPOSED METHOD

Assuming that we are given a set of $N$ micro-videos $\{X_i\}_{i=1}^N$, we have a data matrix $X_k = [x_{k1}^1, x_{k2}^1, \cdots, x_{kN}^1] \in \mathbb{R}^{D_k \times N}$ ($k = 1, 2, \cdots, K$), where $x_{ki}^k \in \mathbb{R}^{D_k}$ is the feature vector of $x_i$ extracted from the $k$-th view, $K$ is the number of views, and $D_k$ is the feature dimension of the $k$-th view.

Micro-video event detection involves a union of multiple types of features. Although multiple representations of micro-videos are complementary and mutually beneficial to each other, the heterogeneity among modalities tends to make them distinct from different subspaces. To eliminate this problem, we first consider mapping the original observations to low-dimensional subspaces through a set of view-specific embedding models $\{W_1, \cdots, W_K\}$. Considering low-rank techniques for denoising [28], we then decompose the mapped observations into the latent low-rank semantic dictionaries $\{U_1, \cdots, U_K\}$ with their corresponding reconstruction errors $\{E_1, \cdots, E_K\}$ to derive a latent common representation $Z$. Accordingly, we formulate the backbone of our proposed method as follows:

$$\min_{Z, W_1, \cdots, W_K} \sum_{i=1}^K \|U_i\|_0 + \gamma_1 \sum_{i=1}^K \|E_i\|_1 + \gamma_2 \sum_{i=1}^K \|E_i\|_{2,1}$$

s.t. $W_i^T X_i = U_i Z + E_i, \quad i = 1, \cdots, K,$ \hspace{1cm} (1)

where $Z$ encodes coefficients of mapped features based on semantic dictionaries; $\|\cdot\|_0$ denotes the nuclear norm to enforce low-rank solutions; $\|\cdot\|_1$ and $\|\cdot\|_{2,1}$ are the $\ell_1$- and $\ell_{2,1}$-norms, which are restricted on the reconstruction errors to simultaneously handle random corruption and maintain the robustness for outliers; $\gamma_1$ and $\gamma_2$ are balanced parameters.

To enhance the representation ability of Eq. (1), we further consider the followings:

1) CORRELATION ANALYSIS

To obtain the common representation of all modalities, we maximize the total correlation across all pairs of modalities while discarding part of the redundant information as follows:

$$\max_{W_1, \cdots, W_K} \sum_{i=1}^K \sum_{j=1}^K \text{tr}(W_i^T X_i X_j^T W_j)$$

s.t. $W_i^T S_{ii} W_i = I, \quad i = 1, \cdots, K,$ \hspace{1cm} (2)

where $I$ is an identity matrix and $S_{ij} = X_i X_j^T$ is defined as the covariance matrix of the $i$-th and the $j$-th modalities.

2) LABEL RELAXATION

We denote the binary label matrix of micro-videos as $Y = [y_1; y_2; \cdots; y_N] \in \{0, 1\}^N \times C$, where $y_i \in \mathbb{R}^C$ is the label vector of the $i$-th micro-video, and $C$ is the number of classes. If the $i$-th micro-video belongs to the $j$-th event type, we denote $y_{ij} = 1$, and zero otherwise. We first present micro-video event detection in the regression form as the backbone for reasons of efficiency, i.e., $\min_A \|Z^T A - Y\|_F^2 + \lambda \|A\|_1$.}

VOLUME 8, 2020

87267
\( \gamma \|A\|_F^2 \), where \( A \) is the coefficient matrix building the connections between latent feature representations and the label matrix, and \( \gamma \) is a balanced parameter. We then consider two improvements: one is to relax the strict binary label assignment into soft label assignments, and the other is that the geometric structure in each modality could be preserved to guide common representation learning.

For the first improvement, the widely used binary label assignment is too rigid to provide more freedom in the mapping from features to label spaces. To address this problem, we introduce a nonnegative label-relaxation matrix \( M \) that combines with the indicator matrix \( B \) to form a slack label matrix \( Y + B \odot M \), where \( \odot \) is the Hadamard product operator. \( B \) is set to \( B_{ij} = 1 \) when \( Y_{ij} = 1 \), and \( B_{ij} = -1 \) otherwise.

Therefore, the label-relaxed regression model can be reformulated as follows:

\[
\min_{A,M} \|Z^T A - (Y + B \odot M)\|_F^2 + \delta \|A\|_F^2 \quad (3)
\]

For the second improvement, a straightforward method is to construct a common shared graph Laplacian to preserve the geometric structure consistency among modalities, i.e., \( \sum_{k=1}^K \text{tr}(A^T Z L^k Z^T A) = \text{tr}(A^T Z L Z^T A) \), where \( L = \sum_{k=1}^K L^k = \sum_{k=1}^K D^k - S^k \), \( L^k \) is the graph Laplacian matrix for the \( i \)-th view, \( S^k \) is the weight matrix computed by the Gaussian similarity function, and \( D^k \) is the diagonal degree matrix with \( D^k_{ii} = \sum_j S^k_{ij} \). The main idea of the graph regularizer is that if the two samples are similar in original feature space, the two samples should also be close neighbors in the latent low-rank space, and vice versa. Different from the general methods of calculating the Laplacian matrix in feature space, in our proposed algorithm, a unified graph structure is calculated build connections between feature and labels in order to ensure the reliability of the prediction results and the robustness of our proposed algorithm.

By combining the low-rank multimodal embedding and the label-relaxed discriminant learning terms, we formulate the compact optimization problem as follows:

\[
\begin{align*}
\min_{W,U,Z,E,M} &\|U\|_F + \gamma_1 \|E\|_{2,1} + \gamma_2 \|E\|_1 + \delta \|A\|_F^2 \\
&\quad + \gamma \text{tr}(A^T Z L Z^T A) - \beta \text{tr}(W^T S W) \\
&\quad + \varphi \|Z^T A - (Y + B \odot M)\|_F^2 \\
\text{s.t.} &\quad W^T X = U Z + E, \quad W^T S W = I \quad (4)
\end{align*}
\]

The meaning of each symbol in the objective function is summarized in TABLE 1.

3) OPTIMIZATION
Due to the nonconvex property of Eq. (4), it is difficult to achieve the global optimal solution. We develop an effective
optimization algorithm by iteratively updating each variable with fixing the other variables. We first introduce two reflected variables $E_1$ and $E_2$ to obtain the augmented Lagrangian function [29] as follows:

$$\|U\|_* + \gamma_1 \|E_{1,2}\|_1 + \gamma_2 \|E_2\|_1 + \delta \|A\|_F^2 - \beta Tr(W^T SW) + \varphi \|Z^T A - (Y + B \otimes M)\|_F^2 + \gamma Tr(A^T ZLZ^T A) + \frac{\mu}{2} \left( \left\| W^T X - UZ - E + \frac{Y_1}{\mu} \right\|_F^2 + \|E - E_1 + \frac{Y_2}{\mu}\|_F^2 \right) + \frac{\mu}{2} \left( \|E - E_2 + \frac{Y_3}{\mu}\|_F^2 + \left\| W^T \hat{S}W - I + \frac{Y_4}{\mu}\right\|_F^2 \right)$$

where $Y_1, Y_2, Y_3$ and $Y_4$ are Lagrange multipliers and $\mu > 0$ is a penalty parameter.

*Update $W$: We can update $W$ as the following scheme:

$$W = \arg \min_W \frac{\mu}{2} \left( \|W^T X - UZ - E + Y_1\|_F^2 + \beta Tr(W^T SW) \right)$$

By setting the derivative of Eq. (5) regarding $W$ to zero, we obtain

$$(\mu - 2\beta)\hat{S}^{-1}SW + 2Wy = \mu\hat{S}^{-1}XG_1$$

where $G_1 = (UZ + E - Y_1/\mu)^T$.

*Update $U$: We can update $U$ as follows:

$$U = \arg \min_U \|U\|_* + \frac{\mu}{2} \left( \|W^T X - UZ - E + Y_1\|_F^2 \right)$$

$$= \arg \min_U \|U\|_* + \frac{\mu}{2} \left( \|W - \hat{U}\|_F^2 + \langle \nabla_U, U - \hat{U} \rangle \right)$$

where $\hat{U}$ is the optimal solution in the previous iteration.

*Update $Z$: We can update the terms independent of $Z$ as follows:

$$Z = \arg \min_Z \varphi \left( \left\| Z^T A - (Y + B \otimes M)\right\|_F^2 + \gamma Tr(A^T ZLZ^T A) \right) + \frac{\mu}{2} \left( \|W^T X - UZ - E + Y_1\|_F^2 \right)$$

By setting the derivative of Eq. (8) regarding $W$ to zero, we obtain the closed form solution of Eq. (8)

$$\mu(AA^T)^{-1}U^T UZ + 2Z(\varphi I + \gamma L) = (AA^T)^{-1}(2\varphi A(Y + B \otimes M)^T + \mu U^T \hat{G})$$

where $\hat{G} = W^T X - E + \frac{Y_1}{\mu}$. The above equation can be obtained by solving the Lyapunov equation [31].

*Update $E$: We can update $E$ as the following scheme:

$$E = \arg \min_E \frac{\mu}{2} \left( \|W^T X - UZ - E + Y_1\|_F^2 + \|E - E_1 + \frac{Y_2}{\mu}\|_F^2 + \|E - E_2 + \frac{Y_3}{\mu}\|_F^2 \right)$$

The closed form solution of Eq. (10) is obtained as

$$E = \frac{1}{\lambda} (UZ - W^T X + E_1 + E_2 - \frac{Y_1}{\mu} - \frac{Y_2}{\mu} - \frac{Y_3}{\mu})$$

*Update $E_1$: We can update $E_1$ as the following scheme:

$$E_1 = \arg \min_{E_1} \frac{\mu}{2} \left( \|E_{1,2}\|_1 + \|E - E_1 + \frac{Y_2}{\mu}\|_F^2 \right)$$

where $D_{\gamma/\mu}[-]$ is the self-defined operator. Given an arbitrary matrix $Q = [q_1, q_2, \ldots, q_i, \ldots]$ and the parameter $\theta$, the operator $D_\theta[Q]$ is defined as follows:

$$D_\theta[Q] (:, i) = \begin{cases} \frac{\|q_i - \theta\| - \|q_i\|}{\|q_i\|} q_i & \text{if } \theta < \|q_i\| \\ 0 & \text{otherwise} \end{cases}$$

*Update $E_2$: We can update $E_2$ as follows:

$$E_2 = \arg \min_{E_2} \frac{\mu}{2} \left( \|E_{2,3}\|_1 + \|E - E_2 + \frac{Y_3}{\mu}\|_F^2 \right)$$

The closed form solution of Eq. (11) is obtained as

$$E_2 = S_{\gamma/\mu}[E + Y_3/\mu]$$

*Update $A$: We can update $A$ as the following scheme:

$$A = \arg \min_A \varphi \left( \left\| Z^T A - (Y + B \otimes M)\right\|_F^2 + \delta \|A\|_F^2 + \gamma Tr(A^T ZLZ^T A) \right)$$

The optimal solution of Eq. (16) can be obtained as follows:

$$M = \max(Q \odot B, 0)$$

where $Q = Z^T A - Y$.

## III. EXPERIMENTAL RESULTS AND ANALYSIS

### A. DATASET AND EXPERIMENT SETTING

We constructed a micro-video event detection dataset from Flickr through its public API.\(^3\) We collected 20 event types with reference to TRECVID MED 2013 [33] and MED 2014, including “Biking”, “Camping”, “Cooking”, “Concert”, “Dance”, “Diving”, “Drawing”, “Driving”, “Football”,

3https://www.flickr.com/services/api/.
Algorithm 1 The Optimization of Objective Function

Input: Feature matrix X, Label matrix Y
Initialization: $E_0 = E_t_0 = E_{2,0} = Y_{1,0} = Y_{2,0} = Y_{3,0} = Y_{4,0} = 0$, $Z_0 = [I, 0]$, $\mu = 10^{-2}$, $\mu_{max} = 10^8$, $t = 0$, $t_{max} = 50$, $\rho = 1.01$, $\varphi = 0.7$, $\psi = 0.01$, $\gamma_1 = \gamma_2 = \gamma = \phi = 10^{-4}$, $D_0 = 512$, Reduce dimension of $X$ to get $Q$, $W_0 = (X/Q)^T$, $U_0 = Q/Z_0$, $A_0 = Y/Z_0$.

1) Set the convergence condition of the algorithm:
$$V(t) = \|Z_t - Z_{t-1}\|_F^2 < 10^{-5};$$

2) While not converged do

3) Fix the others and update $W_{t+1}$:
$$Y_t = \min_{W_t} \frac{1}{2} \|U - V_t + \mu_{t+1} \mu\|_F^2 + \frac{1}{t_{t+1}} \|U\|_F^2;$$

4) Fix the others and update $U_{t+1}$:
$$U_{t+1} = \min_{U_t} \frac{1}{2} \|V_{t+1} - U_t - \mu_{t+1} \mu\|_F^2 + \frac{1}{t_{t+1}} \|U_t\|_F^2;$$

5) Fix the others and update $Z_{t+1}$:
$$Z_{t+1} = \min_{Z_t} \frac{1}{2} \|V_{t+1} - Z_t - \mu_{t+1} \mu\|_F^2 + \frac{1}{t_{t+1}} \|Z_t\|_F^2;$$

6) Fix the others and update $E_{t+1}$:
$$E_{t+1} = \min_{E_t} \frac{1}{2} \|Z_t - W_t - \mu_{t+1} \mu\|_F^2 + \frac{1}{t_{t+1}} \|E_t\|_F^2;$$

7) Fix the others and update $E_{2,t+1}$:
$$E_{2,t+1} = \min_{E_2} \frac{1}{2} \|Z_t - W_t - \mu_{t+1} \mu\|_F^2 + \frac{1}{t_{t+1}} \|E_2 - E_t\|_F^2 + \frac{1}{t_{t+1}} \|Y_{t+1} - \mu_{t+1} \mu\|_F^2;$$

8) Fix the others and update $A_{t+1}$:
$$A_{t+1} = (\mu Z_t + \mu_{t+1} \mu I + \mu \phi Z_t + \mu \lambda Z_t) -1 \mu \phi Z_t + \mu \lambda Z_t;$$

9) Fix the others and update $M_{t+1}$:
$$M_{t+1} = \max (R_t \odot B_t, 0);$$

10) Fix the others and update $Y_{1,t+1}$, $Y_{2,t+1}$, $Y_{3,t+1}$, $Y_{4,t+1}$, and $\mu_{t+1}$:
$$Y_{1,t+1} = \mu (W_{t+1}^T X - U_{t+1} Z_{t+1} - E_{t+1}),$$
$$Y_{2,t+1} = \mu (E_{t+1} - E_{t+1}),$$
$$Y_{3,t+1} = \mu (E_{t+1} - E_{t+1}),$$
$$Y_{4,t+1} = \mu (W_{t+1}^T S W_{t+1} - 1),$$
$$\mu_{t+1} = \min (\rho \mu_{t+1}, \mu_{max}).$$

11) Check the convergence conditions.

Output: $U$, $W$, $A$

All experiments were repeated three times for random training and test partitions and were reported as the average performance. Moreover, the trade-off parameters were all tuned by a grid-search strategy. We adopted the mean average precision (mAP), recall, and precision as metrics to measure the accuracy of the results.

B. FEATURE EXTRACTION

Because of the strong performance of the deep convolutional neural network (CNN) [34]–[36] [37], [38]in various visual recognition tasks, we extracted deep object and scene representations to characterize micro-videos. In more detail, we employed the two well-trained VGNNets [39] as our feature extractors, which were trained with 1.2 million images of 1,000 object classes from the ImageNet dataset [39] and 1.8 million images of 365 scene categories from the Places365-standard dataset [40]. For each feature extractor, the outputs of the fully connected layer fc6 were treated as high-level semantic representations [41], which resulted in 2048D feature vectors. Moreover, to obtain more robust representations, we adopted a keyframe-based strategy [42], i.e., the final features were obtained by conducting average pooling to fuse features extracted from keyframes.

C. EXPERIMENT RESULTS

1) EVALUATION ON CONVERGENCE

We investigated the convergence of our proposed alternating algorithm. In addition to the objective function values, we also calculated the convergence of $Z$ by measuring the variance between two consecutive measurements since it is directly used to perform event detection. FIGURE 4 shows the convergence curves based on our proposed optimization algorithm. From the figure, we can see that both curves decreased rapidly as the number of iterations increased. In particular, the solutions stabilized the convergence curve after approximately 40 iterations, which guarantees the stability of our proposed optimization algorithm.

2) EVALUATION ON PARAMETERS

Among all parameters, the parameters $\alpha$, $\beta$, and $\varphi$ played significant roles in our proposed method since they constrained the impact of geometrical structure consistency, correlation maximization, and label relaxation, respectively. FIGURE 5(a), FIGURE 5(b), and FIGURE 5(c) show the detection performance with different values of parameters on our proposed method. In our experiments, the parameters $\alpha$, $\beta$, and $\varphi$ were selected using a grid-search strategy in a heuristic manner, ranging from 0.00001 to 1, 0.01 to 100, and 0.0001 to 10, respectively. From the figures, we can observe that the best performance was obtained when $\alpha = 0.001$, $\beta = 0.7$, and $\varphi = 0.01$. If these three parameters were too large, they led to a noticeable performance drop. We also changed the percentage of labeled samples from 30% to 100%. As depicted in FIGURE 5(d), we can see that the performance improved significantly with sufficient labeled information.
3) EVALUATION ON TRAINING SAMPLES

This section mainly discusses the influence of changing the size of training set on the micro-video event detection results based on low-rank multimodal representation learning. Specifically, a new training set is constructed by randomly extracting samples from the original training set according to different proportions, and the size of the original test set is unchanged. The experiments are carried out without kernel function and with kernel function (linear kernel) respectively, and the changes of detection results are shown in FIGURE 6(a) and FIGURE 6(b).

From FIGURE 6(a), we can see the detection results without kernel function algorithm. When the training set increases to 70%, the mAP and Precision curve begin to flatten. On the one hand, it shows that increasing the size of training set can effectively improve the accuracy of detection results when the size of training set is small. But on the other hand, it also shows that there is a limit to the improvement of the detection results. When the size of training set reaches a certain amount, the accuracy of the detection results almost do not improve. This shows that there is a threshold for the accuracy of the detection results with the changes of the size of training set. From FIGURE 6(b), we can see the detection results with kernel function algorithm. When the size of training set increases gradually, the mAP and Precision curves keep
4) EVALUATION ON ABLATION ANALYSIS

In order to analyze the necessity of each component in our objective function, we calculated the changes averaged experiment results by removing different modules. Specifically, we conducted experiments as follows: Specifically, the objective function is processed as follows:

- noM: we removed the label-relaxation matrix $M$ to test the performance of by setting $M = 0$.
- noGraph: we remove the introduced graph regularization operator, namely the Laplacian matrix by setting $\gamma = 0$.
- noMCCA: we eliminated the influence of correlation analysis by setting $\beta = 0$. The final recognition results of each evaluation standard after removing different ingredients are shown in TABLE 2.

The following conclusions can be drawn from TABLE 2: 1) the most influential component is noMCCA, which encodes complementarity and correlation among modalities in our proposed algorithm; 2) the prediction results of both noM and noGraph are worse than the complete algorithm, which shows the necessity of introducing graph regularizer and label-relaxation matrix; 3) the result of noM is worse than that of noGraph, illustrating that the graph regularization term produce significant influence than the label relaxation term.

5) COMPARISON WITH STATE-OF-THE-ART METHODS

We compare our proposed method with several state-of-the-art methods, including support vector machine (SVM) [43], supervised regularization-based robust subspace (SRRS) [44], low-rank common subspace (LRCS) [45], discriminative transfer subspace learning (DTSL) [46], robust multi-view subspace learning (RMSSL) [47], collective low-rank subspace (CLRS) [48], Nearly-isotonic support vector machine (NI-SVM) [49], bilevel semantic representation (B-LSA) [8], Deep pARallel Sequence with sparse constraint (EASTERN) [50], Deep tRansfer model (DARE) [4], 3D CNN (C3D) [51], and inflated 3D ConvNet (I3D) [52].

TABLE 3 shows the recognition performance compared with several state-of-the-art algorithms. Obviously, our method outperformed all the other algorithms in terms of mAP, precision, and recall. Benefiting from the multiview property, RMSL, CLRS and our proposed algorithm performed much better than SRRS, LRCS, and DTSL. Our algorithm outperformed RMSSL and CLRS. One possible reason is that the geometric structure in each modality is preserved in our model to guide better representation learning. In the deep learning method, we choose two convolutional neural networks, C3D and 13D for event detection and action recognition. The results are very common. At the same time, we choose four methods NI-SVM, B-LSA, EASTERN
and DARE, which use deep concepts to detect micro-video events. The results are better than those of C3D and I3D, which shows that the potential semantic features in these videos can represent complex events better.

IV. CONCLUSION

In this paper, we proposed a label-relaxed multimodal low-rank representation algorithm to address micro-video event detection. To learn more intrinsic representations of micro-videos, the complementarity among modalities was explored to address the less descriptive power of each modality. The low-rank and label-relaxed constraints were further exploited to ensure more intrinsic representations. Experiment results conducted on our constructed dataset verified its superior performance. Considering the complex pattern among different modalities, especially in scenarios with modality missing. In the future, we will develop a novel and flexible multimodal representation framework to deal with the problem of missing modalities.

REFERENCES

[1] M. Redi, N. O'Hare, R. Schifanella, M. Trevisiol, and A. Jaimis, “6 seconds of sound and vision: Creativity in micro-videos,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 4722–4729.

[2] D. Cao, X. He, L. Miao, Y. An, C. Yang, and R. Hong, “Attentive group recommendation,” in Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Jun. 2018, pp. 645–654.

[3] J. Chen, X. Song, L. Nie, X. Wang, H. Zhang, and T.-S. Chua, “Micro tells macro: Predicting the popularity of micro-videos via a transductive model,” in Proc. ACM Multimedia Conf., 2016, pp. 998–907.

[4] L. Nie, X. Wang, J. Zhang, X. He, H. Zhang, R. Hong, and Q. Tian, “Enhancing micro-video understanding by harnessing external sounds,” in Proc. ACM Multimedia Conf., 2017, pp. 1192–1200.

[5] M. Liu, L. Nie, X. Wang, Q. Tian, and B. Chen, “Online data organizer: Micro-video categorization by structure-guided multimodal dictionary learning,” IEEE Trans. Image Process., vol. 28, no. 3, pp. 1235–1247, Mar. 2019.

[6] Y. Wei, X. Wang, W. Guan, L. Nie, Z. Lin, and B. Chen, “Neural multimodal cooperative learning toward micro-video understanding,” IEEE Trans. Image Process., vol. 29, pp. 1–14, 2020.

[7] X. Liu, M. Li, C. Tang, J. Xia, J. Xiong, L. Liu, M. Kloft, and E. Zhu, “Efficient and effective regularized incomplete multi-view clustering,” IEEE Trans. Pattern Anal. Mach. Intell., early access, Feb. 18, 2020. doi: 10.1109/TPAMI.2020.2974828.

[8] X. Chang, Z. Ma, Y. Yang, Z. Zeng, and A. G. Hauptmann, “Bi-level semantic representation analysis for multimedia event detection,” IEEE Trans. Cybern., vol. 47, no. 5, pp. 1180–1436, Jan. 2019.

[9] Z. Ma, X. Chang, Z. Xu, N. Sebe, and A. G. Hauptmann, “Joint attributes and event analysis for multimedia event detection,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 7, pp. 2921–2930, Oct. 2017.

[10] L. Wang, Z. Wang, Y. Qiao, and L. Van Gool, “Transferring deep object and scene representations for event recognition in still images,” Int. J. Comput. Vis., vol. 126, nos. 2–4, pp. 390–409, Apr. 2018.

[11] W. Xie, H. Yao, X. Sun, T. Han, S. Zhao, and T.-S. Chua, “Discovering latent discriminative patterns for multi-mode event representation,” IEEE Trans. Multimedia, vol. 21, no. 6, pp. 1425–1436, Jun. 2019.

[12] S. Zhao, H. Yao, Y. Gao, G. Ding, and T.-S. Chua, “Predicting personalized image emotion perceptions in social networks,” IEEE Trans. Affect. Comput., vol. 9, no. 4, pp. 526–540, Oct. 2018.

[13] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, “Continuous probability distribution prediction of image emotions via multitask shared sparse regression,” IEEE Trans. Multimedia, vol. 19, no. 3, pp. 632–645, Mar. 2017.

[14] S. Zhao, G. Ding, Y. Gao, and J. Han, “Approximating discrete probability distribution of image emotions by multi-modal features fusion,” Transfer, vol. 1000, no. 1, pp. 4669–4675, 2017.
[39] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[40] R. Raturi, “Adapting deep features for scene recognition utilizing places database,” in Proc. 2nd Int. Conf. Inventive Comput. Commun. Technol. (ICICCT), Apr. 2018, pp. 487–495.

[41] S. Wang, T. Liu, and L. Tan, “Automatically learning semantic features for defect prediction,” in Proc. 38th Int. Conf. Softw. Eng. (ICSE), 2016, pp. 297–308.

[42] C. Dang and H. Radha, “RPCA-KFE: Key frame extraction for video using robust principal component analysis,” IEEE Trans. Image Process., vol. 24, no. 11, pp. 3742–3753, Nov. 2015.

[43] J. A. Suykens and J. Vandewalle, “Least squares support vector machine classifiers,” Neural Process. Lett., vol. 9, no. 3, pp. 293–300, 1999.

[44] S. Li and Y. Fu, “Learning robust and discriminative subspace with low-rank constraints,” IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 11, pp. 2160–2173, Nov. 2016.

[45] Z. Ding and Y. Fu, “Low-rank common subspace for multi-view learning,” in Proc. IEEE Int. Conf. Data Mining, Dec. 2014, pp. 110–119.

[46] Y. Xu, X. Fang, J. Wu, X. Li, and D. Zhang, “Discriminative transfer subspace learning via low-rank and sparse representation,” IEEE Trans. Image Process., vol. 25, no. 2, pp. 850–863, Feb. 2016.

[47] Z. Ding and Y. Fu, “Robust multi-view subspace learning through dual low-rank decompositions,” in Proc. AAAI Conf. Artif. Intell., 2016, pp. 1181–1187.

[48] Z. Ding and Y. Fu, “Robust multiview data analysis through collective low-rank subspace,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 5, pp. 1986–1997, May 2018.

[49] X. Chang, Y.-L. Yu, Y. Yang, and E. P. Xing, “Semantic pooling for complex event analysis in untrimmed videos,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 8, pp. 1617–1632, Aug. 2017.

[50] M. Liu, L. Nie, M. Wang, and B. Chen, “Towards micro-video understanding by joint sequential-sparse modeling,” in Proc. ACM Multimedia Conf., 2017, pp. 970–978.

[51] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4489–4497.

[52] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6299–6308.

JING ZHANG received the B.S. and M.S. degrees from Sichuan University, Chengdu, China, in 1994 and 1997, respectively, and the Ph.D. degree from Tianjin University, Tianjin, China, in 2002. She is currently an Associate Professor with the School of Electrical and Information Engineering, Tianjin University. Her research interests include computer vision, multimedia content analysis, and digital image / video processing.

YUTING WU received the B.S. degree in computer science from the School of Electrical and Information Engineering, Tianjin University, Tianjin, China, in 2018. He is currently pursuing the M.S. degree with Tianjin University. His research interests include computer vision, micro-video event detection, tensor decomposition, and deep learning. In 2018, he jointly did thesis research in Columbia University in the city of New York.

JINGHUI LIU received the M.S. degree in computer science from the School of Electrical and Information Engineering, Tianjin University, Tianjin, China, in 2020. His research interests include micro-video event detection and deep learning.

PEIGUANG JING received the M.S. and Ph.D. degrees from Tianjin University, Tianjin, China, in 2013 and 2018, respectively. From 2014 to 2015, he was a Visiting Student with the National University of Singapore. He is currently an Assistant Professor with the School of Electrical and Information Engineering, Tianjin University. His current research interests include multimedia computing, signal processing, and machine learning.

YUTING SU received the B.S., M.S., and Ph.D. degrees from Tianjin University, Tianjin, China, in 1995, 1998, and 2001, respectively. He is currently a Professor with the School of Electrical and Information Engineering, Tianjin University. His research interests include computer vision, multimedia content analysis, information security, and tensor decomposition.

* * *