IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 26, 2020, accepted December 2, 2020, date of publication December 9, 2020,
date of current version December 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3043412

Behavior Recognition Based on Category Subspace in Crowded Videos

CHUNHUA DENG[1,2], XIAOGE KANG[1,2], ZIQI ZHU[1,2], (Member, IEEE), AND SHIQIAN WU[2], (Senior Member, IEEE)

[1]School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China
[2]Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430065, China

Corresponding author: Ziqi Zhu (zhuzq@wust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61602350, in part by the Provincial Natural Science Fund of Hubei under Grant 2018 CFB195, in part by the Provincial Education Office Science and Technology Research Project Youth Talent Project of Hubei under Grant Q20181104, in part by the Defense Pre-Research Fund of Wuhan University of Science and Technology under Grant GF201814, and in part by the Key Laboratory of Image Information Processing and Intelligent Control of the Ministry of Education under Grant IPIC2018-04.

ABSTRACT Crowd behavior refers to a collective behavior composed of two or more individuals who influence, interact, and depend on each other for a specific goal. Compared with an ordinary crowd behavior, the probability of a dangerous crowd behavior is much smaller. Video-based crowd behavior recognition can be categorized as one multi-label classification task, which is characterized by complex scenes and imbalanced samples. Aimed at tackling problems of imbalanced samples and multi-label task, a classification method of associative subspace is proposed. For a single category (called main category) with fewer samples, this paper generates a special subspace wherein it is relatively easy to distinguish these samples by association with other categories. A classifier that can weaken the main category and strengthen relationship between the main category and other categories is designed in the subspace. Therefore, the main category can contribute to reducing dependence on the number of samples with the above-mentioned classifier in the corresponding subspace. In order to make full use of the relevant information concerning categories, multi-label information is further injected into spatio-temporal features of video action representation. Experiments on a challenging WWW dataset show that both the proposed subspace method and multi-label information fusion mechanism are efficient.

INDEX TERMS Multi-label, subspace, imbalanced samples, associative subspace.

I. INTRODUCTION

Crowded videos of the same category may contain different scenes, different numbers of people, different fields of vision, making the crowded video classification task very challenging [1]–[4]. In the real world, dangerous crowd behaviors are unlikely to occur, which makes it difficult to collect sufficient video data. However, dangerous crowd behaviors (stampede, riots, etc.) can cause huge loss of property and lives easily. Even more, existing crowded video data sets are very poorly balanced [1] including numerous samples of general crowded videos but very few dangerous ones. Therefore, it is very important to investigate the classification of imbalanced crowded videos.

Crowded videos often contain multiple events and behaviors, thus making the issue of video-based crowd behavior recognition a multi-label classification task. Due to the emergence of the large crowd behavior video dataset (WWW), a series of multi-label crowd behavior recognition algorithms have been proposed [1]–[3]. Among the algorithms above, Shao *et al.* [1] developed a multi-task deep model for joint learning by combining appearance and motion features for a better crowd understanding. Besides, based on category dependencies, the algorithm has also improved multi-label crowded recognition performance under the guidance of manually defined rules. However, the above method does not consider the imbalance of samples, which is disadvantageous to the category of fewer samples in the process of training. To enable categories with fewer samples to perform better in classification, this paper has introduced the idea of subspace. We construct subspaces using category association, which can effectively solve the problem of imbalanced classification. The associative subspace principle is shown in Fig.1. When it comes to distinguish the main category, we expect to generate a suitable subspace in which it can be easily distinguished by using correlation information between the

The associate editor coordinating the review of this manuscript and approving it for publication was Fan-Hsun Tseng.

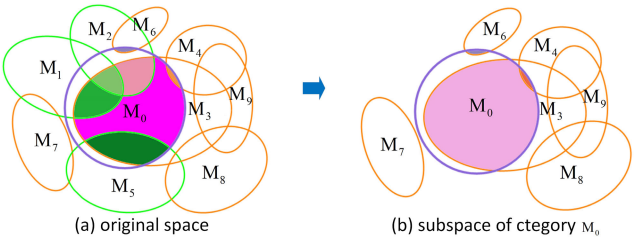

FIGURE 1. An example of subspace (The purple circle is the main category M_0, and ellipses represent other categories). Figure (a) shows the projections of other categories on M_0, and (b) is the corresponding subspace. The projection area of category M_3 is the largest in original space, indicating that there is a close relationship between category M_3 and M_0. Therefore, the subspace contains category M_3. The projection areas of categories M_6 and M_7 are very small (or zero), showing that there are strong distinctions between M_0 and these categories. Thus, category M_6 and M_7 also serve as a part of the subspace. However, for categories M_1 and M_2 and M_5, the association information with the main category during training is not easy to distinguish, so they are not included in the subspace.

main category and the other categories. For each category with a small number of samples, this study needs to generate a corresponding subspace so that the main category can better use the subordinate relationship between the categories in the corresponding subspace.

Regarding the task of imbalanced classification, we construct subspaces with category information so that categories with fewer samples can achieve satisfying performance. Meanwhile, we design a classifier for categories with fewer samples. The classifier is utilized to optimize the current category, which is different from the globally optimized classifiers [3], [5], [6] utilizing the relationship between categories. Specifically, the classifier can weaken the main category and reduce the dependence of the main category on the number of samples on the one hand, and enhance the relationship between the main category and other categories for the indirect classification of the main category on the other hand. In a nut shell, the classifier designed for categories with fewer samples can optimize the current subspace by weakening the main category and weighting the association relationship between categories.

Feature representation plays a pivotal role in visual classification tasks [2], [3]. At present, the mainstream video classification features are obtained through two streams (static and dynamic). Most existing video features [1]–[3], [7]–[9], however, are easily affected by appearance and motion noise, due to the substantial differences in crowded scenes and great variety of motion information. Thus, to tackle the recognition task in crowded scenes, we choose to combine motion trend features with dynamic evolution features. First, multi-label information is integrated into 3D dynamic features by a Graph Convolution Network (GCN) [10] to capture the global motion trend. Then, along with the motion trend, a Long Short-Term Memory (LSTM) network with memory function is used for collecting important evolution features and erasing dynamic and appearance noise.

The main contributions of this paper include: (1) the idea of subspace for categories of fewer samples; (2) the classifier based on subspace correlation and designed to address the problem of imbalanced samples during classification; (3) feature representation that combines motion trend features with dynamic evolution features to enhance the description of video change trend.

II. RELATED WORK

In this section, we will introduce and discuss the related work towards multi-label classification in terms of subspace construction, subspace classifier design and feature representation, respectively.

Subspace clustering seeks to partition the original space into multiple subspaces within a dataset, and clustering algorithms have been widely used for determining the subspaces [11]–[13]. Among the existing subspace clustering methods, the spectral clustering methods [14]–[17] have become increasingly popular because of easy implementation, as well as high probability to converge to a global optimum compared with conventional clustering algorithms. However, it still remains a drawback that the mixed-signed result given by eigenvalue decomposition of Laplacian may degrade the clustering performance [18]. Thus, to address this problem, literature [19] established an equivalence with spectral clustering and proposed two non-negative spectral clustering algorithms. Although spectral clustering has been achieved in many applications, the relationship between the affinity matrix and the labels of the data is not fully exploited, thus there is no guarantee for an overall optimal performance. To overcome the challenge, a new unified optimization framework is proposed, which enforces the coherence and discrimination of the affinity matrix as well as the labels [20]. However, it should be noted that these clustering approaches are inherently unsupervised learning algorithms which tend to ignore the category information. The label information in the crowded scene appears in pairs, and category information plays an important role in the recognition process. The spectral clustering methods fail to notice the association between categories, and are therefore not suitable for multi-label crowd behavior recognition. In view of this, we utilize dependencies among categories to construct subspaces in this paper. For a certain category, its subspace is generated on the basis of the association with other categories.

In this paper, crowd behavior recognition is deemed as a multi-label classification task, and each object is represented by a single instance when associated with multiple labels. Multi-label learning has been extensively studied during the past decades, and many algorithms have been proposed. For example, the simplest method is to decompose a multi-label task into a series of binary classification problems [21]. However, the method is essentially limited by overlooking the label correlations. In this connection, it stimulates research for coming up with approaches to capture and explore the label correlations in various ways. Some approaches, based on graph representation learning [6], [10], are proposed to capture the label correlations for multi-label recognition. Besides, a novel approach multi-instance multi-label fast

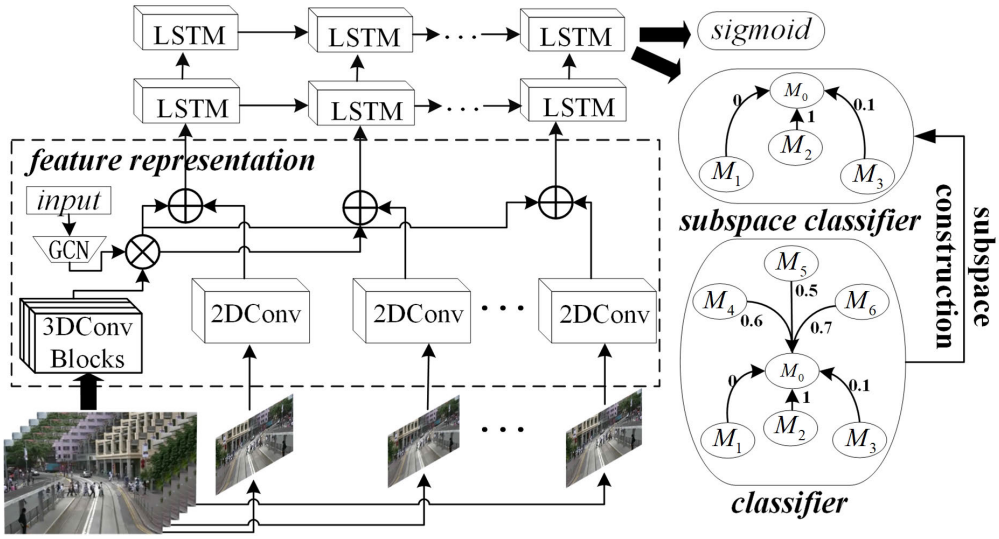

FIGURE 2. An overall framework of crowd behavior recognition. Feature representation is the fusion of static features and dynamic features with category information. The subspace classifier design exploits category association, where M_0 is main category.

learning (MIMLfast) is proposed in literature [5] to utilize the relations among multiple labels. However, it has to be admitted that given the imbalance of sample distribution in our work, the categories with a small number of samples obtain worse results in the classification process. In addition, the aforementioned methods optimize all categories globally, which is adverse to categories with fewer samples. Accordingly, we construct subspaces for these categories by utilizing the category association. Meanwhile, corresponding classifiers are designed for each category subspace.

Feature representation is an important factor for classification. Traditional manual features [22], [23] are gradually replaced by deep learning features. Recently, deep neural networks have been successfully applied to action recognition [24]–[26]. Previously, 2D convolutional neural networks [27], [28] trained by ImageNet [29] were usually exploited for RGB image classification. However, for the task of video classification, appearance information is not enough, and dynamic features representation play a vital role in the process of recognition [9], [30]. To simulate motion information, K. Simonyan *et al.* proposed a two-stream ConvNet architecture which incorporates spatial and temporal networks [8], where the temporal stream is trained to recognize actions from motion in the form of dense optical flow. Literature [7], based on two-stream architecture, made further improvement by introducing residual connections. Meanwhile, there are also some other works trying to adapt existing techniques to solve the action recognition task in videos [31]–[35]. In a nut shell, obtaining effective spatio-temporal feature representation is essential for action recognition. However, in our work, because even the same crowd behavior may have different scenes, the appearance noise will be relatively large; Meanwhile, crowd behavior is usually accompanied by a variety of motion information resulting in relatively large dynamic noise. Existing methods for fusing appearance and dynamic features can be influenced by noise. To effectively describe motion information in the video, we combine motion trend features with dynamic evolution features. To be specific, under the guidance of the motion trend, LSTM is used for collecting important evolution features, and discarding dynamic and appearance noise. Due to the limited expressive power of optical flow in complex motion scenes, we exploit a 3D convolution network [36] to obtain dynamic information. For the task of multi-label recognition, category information can enhance semantic representation. Hence, we integrate dependency relationship between categories into dynamic information to obtain the motion trend with semantic association. Then, dynamic features with strong semantic correlation is fused into frame-level static features. Besides, extensive research has shown that LSTM, as a variant of Recurrent Neural Network (RNN), demonstrates a strong ability to model long-term time dependency in sequence modeling. The LSTM network is also used for video action recognition. In this paper, characteristics of LSTM are used for filtering out the appearance and dynamic noise.

III. METHOD

The architecture of crowd behavior recognition is illustrated in Fig.2. It primarily involves three stages: construction of subspace, subspace classifier design and video feature representation. We construct a subspace for each special category before designing the subspace classifier to optimize the category subspace. Afterward, during representation of video features, we combine motion trend features with dynamic evolution features. Coupled with the motion trend, LSTM serves as a tool of filtering out dynamic and appearance noise. In the following sections, we will refine the process of subspace construction, subspace classifier design and feature representation. In order to make this paper easy to understand, we add a table of symbols (shown in Table 1).

TABLE 1. Nomenclature.

Symbol	Description	Symbol	Description
c	Category number of training set.	$\mathbf{A}$	$A \in [0,1]^{c\times c}$ Incidence matrix of categories.
n	Capacity of training set.	$\mathbf{X}$	$\mathbf{X} \in R^{d\times n}$ is the characteristic matrix of training samples.
e	Category number of small samples.	$\mathbf{U}$	$\mathbf{U} \in R^{d\times c}$ is the mapping matrix from feature to category.
ε	A threshold parameter.	$\mathbf{G}$	$\mathbf{G} \in \{0,1\}^{n\times c}$ is a sample label matrix.
d	The dimension of each sample feature.	$\mathbf{I}$	Identity matrix.
r_i	The capacity of category i in training set.	$\mathbf{U}^{(j)}$.	$\mathbf{U}^{(j)} \in R^{d\times c^{(j)}}$ is the mapping matrix of the j-th subspace.
r_{ij}	The capacity of samples belonging to both category i and j.	$\mathbf{G}^{(j)}$	$\mathbf{G}^{(j)}$ is the label matrix of the j-th subspace.
α	Weight parameter of subspace classifier.	$\mathbf{S}^{(j)}$	$\mathbf{S}^{(j)} \in R^{c^{(j)}\times c^{(j)}}$ is a subspace category incidence matrix.
Ω^j	The j-th subspace	$\Phi^{(j)}$	Regularization term of j-th subspace.
β_1, β_2	Parameters of generating a subspace.	$\mathbf{W}^{(j)}$	The weight matrix of j-th subspace.
$c^{(j)}$	Category capacity of j-th subspace.	$\mathbf{P}_1^{(j)}$	The prediction matrix of the main category of j-th subspace.
γ, λ	Parameters of regularization terms.	$\{R_i\}$	Space of original training data set.
$\{R_i^{min}\}$	Small sample set.	M_0	Main category.
$\mathbf{X}_t^L$	Matrix of input feature of LSTM in frame t.	m	CNN feature dimension of an image.
$\mathbf{X}^M$	Matrix of action trend features.	q	Word dimension in dictionary.
$\mathbf{X}_t^S$	Matrix of static features in frame t.	z	Dimension of 3D features.
$\mathbf{X}^C$	Matrix of dictionary features.	η	Models in neural networks.
$\mathbf{X}^D$	Matrix of 3D network features.	η_{cnn}	CNN neural network model.
$\mathbf{X}_t^I$	Original image of frame. t.	η_{gcn}	GCN neural network model.
$\mathbf{X}^G$	Matrix of GCN network features.	η_{3d}	3D neural network model.
f	Operation functions related to neural networks.	$\mathcal{L}$	Loss function.
f_{ap}	Average pooling operation in neural networks.	$\tilde{\mathbf{G}}$	Prediction matrix of training samples.
f_{cnn}	Convolutional neural network function.	f_{3d}	3D convolution neural network function.
f_{act}	Activation function operation in neural networks.	l	The number of frames selected from a video.

A. CONSTRUCTION OF SUBSPACE

In the real world, the probability of some crowd behaviors is small, and it is difficult to obtain sufficient samples. The distribution of video data samples of crowd is thus extremely imbalanced. These categories with insufficient samples cannot achieve a good classification effect in the training process in spite of strong feature representation. For instance, experiments [6] on multi-label dataset (VOC 2007 [37]) show that the performance (mean average precision (mAP) is 92%) of categories with fewer samples (the number of samples is less than the average on training set) is inferior to that (mAP is 95.1%) of categories with more samples (the number of samples is larger than the average on training set). Simultaneously, the dataset WWW [1] is also imbalanced. According to literature [3], the average value (mAUC) of the area under Receiver Operating Characteristic curve is 89.3% (categories with fewer samples) and 92.4% (categories with more samples) respectively. In order to address the problem of imbalanced samples classification, the idea of subspace is proposed for crowd video categories with fewer samples.

The construction of subspace can be divided into two steps. Firstly, some rules and conditions are applied in determining categories with few samples according to the training set. Secondly, a screening mechanism is adopted to screen a series of categories from all categories and construct a subspace for each small sample category. The specific operation of the aforementioned two processes is as follows.

In order to clearly describe the construction of the subspaces, we first define the relevant symbols. Symbol c and n is the number of categories and the number of samples in the training set respectively. $\{R_i\}$ is the original category set, and R_i is the i-th category. $\{R_j^{min}\}$ refers to the set of small sample categories and contains $e(e < c)$ categories, and r_i is the number of each category samples in the entire training set, and r_{ij} is the number of co-occurrence sample between the i-th category and the j-th small sample category.

Categories with fewer samples are determined by the following expression

$$R^{min} = \{R_i | r_i < \frac{n}{c}\varepsilon\}, \quad i = 1, 2, \ldots, c, \tag{1}$$

where the value of ε is set 0.8.

A subspace Ω^j based on the j-th category in the small sample set $\{R_j^{min}\}$ is formalized as

$$\Omega^j = \{R_i | \frac{r_{ij}}{r_j} > \beta_1 or \frac{r_{ij}}{r_j} < \beta_2\}, \quad i = 1, 2, \ldots, c; \quad j = 1, 2, \ldots, e, \tag{2}$$

where β_1 and β_2 are hyperparameters.

B. SUBSPACE CLASSIFIER

In this paper, since the crowd dataset WWW [1] is base on multi-label, we first utilize the sigmoid activation function to classify all categories. In the classification process, a probability value between 0 and 1 will be assigned to each category, and the categories are independent from each other. Categories with relatively sufficient samples can achieve decent performances by using the sigmoid activation function whilst categories with fewer samples obtain worse results under

the same circumstance. Thus, to overcome this problem, we design special classifier (called subspace classifier) for these categories via introducing the idea of subspace.

The design process of the subspace classifier is shown in Fig.2. Firstly, the main category (M_0) obtains corresponding subspace through category association. The subspace includes categories with a close relationship or strong distinction with M_0, with association relationship between categories falling in the range [0, 1]. After the classification of the main category (M_0) by the subspace classifier, larger the number is, the closer the relationship is.

The subspace category association classifier is inspired by [3] attribute assignment (AA) and [38]. AA harnesses the subordinate relationship between categories for multi-label crowd behavior classification. Suppose that $\mathbf{X} \in R^{d*n}$ represents a feature matrix comprising n training samples and d-dimensional feature vectors. Meanwhile, the mapping matrix $\mathbf{U} \in R^{d\times c}$ bridges low-level features with categories, and $\mathbf{G} \in \{0, 1\}^{n\times c}$ is employed to indicate the label matrix of the entire training set. According to [3], AA employs a convex optimization method with closed solutions

$$\min_{\mathbf{U}} ||\mathbf{X}^T\mathbf{U}\mathbf{A} - \mathbf{G}||_F^2 + \gamma||\mathbf{U}\mathbf{A}||_F^2 + \lambda||\mathbf{X}^T\mathbf{U}||_F^2 + \gamma\lambda||\mathbf{U}||_F^2, \tag{3}$$

where $||\bullet||_F^2$ is Frobenius (F-norm), γ and λ are hyperpara-meters. $\mathbf{A} \in [0, 1]^{c\times c}$ is the proposed dependency matrix which captures the interrelationship among categories. According to [3], the closed solution of the above convex optimization problem is formalized as

$$\mathbf{U} = (\mathbf{X}\mathbf{X}^T + \gamma\mathbf{I})^{-1}\mathbf{X}\mathbf{G}\mathbf{A}(\mathbf{A}\mathbf{A}^T + \lambda\mathbf{I})^{-1}, \tag{4}$$

where I is the identity matrix. Given the closed-form solution to optimization problems, the solution efficiency is very high.

However, AA optimizes globally, which is not conductive to categories with few samples. To end this, we construct a subspace to distinguish categories with fewer samples and design a classifier based on the idea of subspace and called subspace classifier. *In our framework, each category with fewer samples is configured with a special subspace, which is different from AA algorithm. For each category with few samples, the subspace classifier is used for optimizing the corresponding subspace.* In a subspace, the category to be distinguished is deemed as the main category.

According to [3], [38] the optimization function of the subspace is defined as

$$\min_{\mathbf{U}^{(j)}} ||\mathbf{X}^T\mathbf{U}^{(j)}\mathbf{S}^{(j)}\mathbf{W}^{(j)} - \mathbf{G}^{(j)}||_F^2 + \Phi^{(j)}, \tag{5}$$

where $\mathbf{U}^{(j)} \in R^{d\times c^{(j)}}$ is the mapping matrix of the main category (j) on the subspace, which establishes a bridge from features to category information. The matrix $\mathbf{G}^{(j)} \in R^{n\times c^{(j)}}$ is the label matrix of main category (j) on the subspace. Symbol $c^{(j)}$ represents the number of categories on the subspace for the main category (j). The symbol $\mathbf{S}^{(j)} \in R^{c^{(j)}\times c^{(j)}}$ represents correlation matrix of main category (j) on the subspace, and $\mathbf{W}^{(j)} \in R^{c^{(j)}\times c^{(j)}}$ is weight matrix, and $\Phi^{(j)}$ is regularization term. Then, we will introduce the matrices $\mathbf{S}^{(j)}, \mathbf{W}^{(j)}$ and regularization term $\Phi^{(j)}$ in turn. According to [3],the element of matrix $\mathbf{S}^{(j)}$ is formalized as

$$\mathbf{S}_{ik}^{(j)} = \frac{(\mathbf{G}_i^{(j)})^T\mathbf{G}_k^{(j)}}{\sum_{l=1}^{c^{(j)}}\mathbf{G}_{li}^{(j)}}, \quad (c^{(j)} < c). \tag{6}$$

In order to facilitate training, we adjust the order for categories to make the main category the first column, while $\Omega_1^{(j)}$ (formula (2)) is correspondingly the main category. The proportion of the main category is adjusted by the category correlation matrix $\mathbf{S}^{(j)}$ of $\Omega^{(j)}$. The matrix $\mathbf{W}^{(j)}$ weakens the main category and weights the association relationship between categories. The element of $\mathbf{W}^{(j)}$ is formalized as

$$\mathbf{W}_{ik}^{(j)} = \begin{cases} \alpha, & i = 1 \\ 1-\alpha, & i = k \text{ and } i \neq 1, \\ 0, & else \end{cases}$$
$$(j = 1, 2, \ldots, e; \quad k = 1, 2, \ldots, c^{(j)}) \tag{7}$$

where the value range of α is between 0 and 0.3. Symbol e represents the number of subspaces. The matrix $\mathbf{W}^{(j)}$, according to formula (7), is expressed as

$$\mathbf{W}^{(j)} = \begin{bmatrix} \alpha & \alpha & \ldots & \alpha \\ 0 & 1-\alpha & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1-\alpha \end{bmatrix}_{c^{(j)}\times c^{(j)}}, \tag{8}$$

where each element on the main diagonal represents the proportion of each category on the subspace. The value of $\mathbf{W}_{11}^{(j)}$ is α, indicating that the main category (j) is weakened because the value of α is less than 0.3. Other element values on the main diagonal are $1-\alpha$, implying that the proportion of other categories is strengthened on the subspace. The relationship between the main category and other categories is established by setting the value of the element to α (the value of α in the matrix $\mathbf{W}$ is the same). Specifically, the element $\mathbf{W}_{1k}^{(j)}$ is α. The element value is 0 in the matrix, suggesting that there is no dependency between two categories. *If $\mathbf{W}^{(j)}$ is an identity matrix, following formula (3), the classifier will be attribute assignment (AA). If not, according to formula (8), the classifier will be the subspace classifier.* In addition, the regularization term $\Phi^{(j)}$ is defined according to [3]

$$\Phi^{(j)} = \gamma^{(j)}||\mathbf{U}^{(j)}\mathbf{S}^{(j)}\mathbf{W}^{(j)}||_F^2 + \lambda^{(j)}||\mathbf{X}^T\mathbf{U}^{(j)}||_F^2 + \gamma^{(j)}\lambda^{(j)}||\mathbf{U}^{(j)}||_F^2, \tag{9}$$

and $\mathbf{U}^{(j)}$ is formulated as

$$\mathbf{U}^{(j)} = (\mathbf{X}\mathbf{X}^T + \gamma^{(j)}\mathbf{I})^{-1}\mathbf{X}\mathbf{G}^{(j)}\mathbf{S}^{(j)}\mathbf{W}^{(j)} \times (\mathbf{S}^{(j)}\mathbf{W}^{(j)}\mathbf{W}^{(j)T}\mathbf{S}^{(j)T} + \lambda^{(j)}\mathbf{I})^{-1}. \tag{10}$$

In test phase, the inference of the main category is similar to [3]

$$\mathbf{P}_1^{(j)} = \mathbf{X}^T\mathbf{U}^{(j)}\mathbf{S}_1^{(j)}, \tag{11}$$

where $\mathbf{S}_1^{(j)}$ represents the first column in $\mathbf{S}^{(j)}$.

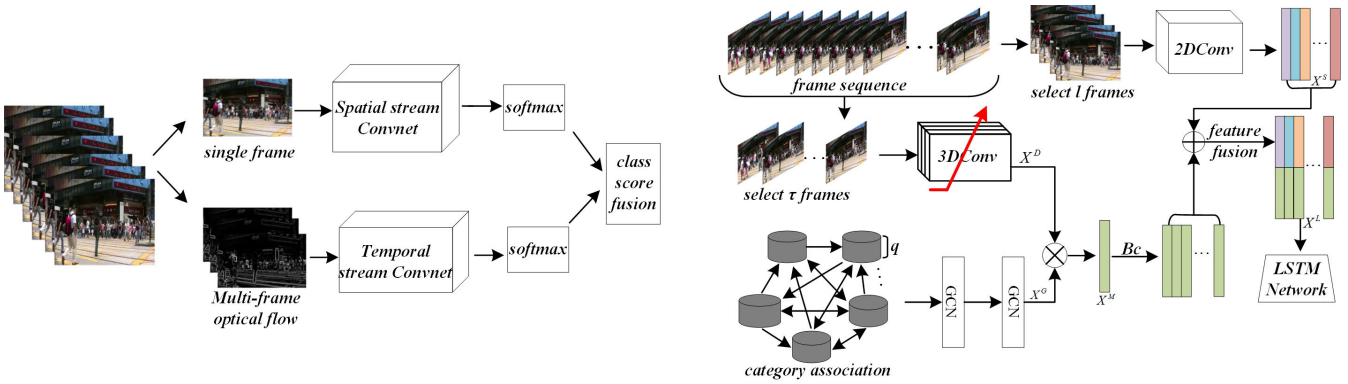

FIGURE 3. Feature representation. (a) Two-stream architecture for video classification according to literature [8]. (b) The process of fusing of static features and dynamic features with category information. Symbol τ and l represents the number of frames uniformly selected from the original frame sequence. Symbol Bc represents Broadcasting.

C. FEATURE REPRESENTATION

How to effectively capture the distinct spatio-temporal features to model the spatio-temporal evolution of different actions is crucial for video action recognition. As shown in Fig.3(a), in the traditional method [8], the spatial stream captures still frame-level features, whilst the temporal stream captures dynamic features in the form of dense optical flow. Finally, the class score is average of static score and dynamic score. However, joining static and dynamic features together could easily be subject to influence of noise. In order to obtain a discriminative feature representation in crowded scenes, we adopt the approach of combining motion trend features with dynamic evolution features rather than the traditional method. The overall framework of our approach is shown in Fig.3(b), which is composed of two main modules: dynamic evolution features and motion trend features.

1) DYNAMIC EVOLUTION FEATURES

With regard to the task of video action recognition, LSTM network is selected to capture context information, as shown Fig.2. The LSTM network is composed of two layers. In detail, the first layer 'lstm' unit obtains each state output of the video sequence, while the output of the second layer 'lstm' for the last time step is applied for classification. The input of the LSTM network is the frame-by-frame fusion of the motion trend features and static frame-level features

$$\mathbf{X}_t^L = \mathbf{X}^M \oplus \mathbf{X}_t^S \quad (t \in 1, 2, 3, \ldots, l), \tag{12}$$

where $\mathbf{X}_t^S$ represents the static features of the t-th frame, and $\mathbf{X}^M$ represents the motion trend features. The static frame is uniformly selected l frames from the original frame sequence, as is shown in Fig.3(b).

We can use any CNN basic model [27], [28] to capture the frame-level features of the video. In our experiments, following [1], [3], the ResNet50 model [28] pre-trained on ImageNet [29] is chosen as the backbone CNN. Therefore, if the input video frame $\mathbf{X}_t^I$ is a resolution of 224 × 224, we can obtain the 7 × 7 × 2048 feature map from the last 'conv' layer. Then, we adopt average pooling (AP) to obtain the frame-level feature $\mathbf{X}_t^S$

$$\mathbf{X}_t^S = f_{ap}(f_{cnn}(\mathbf{X}_t^I; \eta_{cnn})) \in R^m, \tag{13}$$

where η_{cnn} represents CNN neural network model, $m = 2048$.

2) MOTION TREND FEATURES REPRESENTATION

To effectively describe motion trend features ($\mathbf{X}^M$), the global correlation between labels is applied to the 3D dynamic feature map to figure out dynamic features with semantic association, with the overall framework depicted in Fig.3(b). First, we feed τ frames into a 3D convolution network for describing the overall motion trend. Then, category information is employed as the input of GCN to explore semantic association, ending up with the combination between semantic association and motion trend. In our work, in line with [10], [39], we have used stacked GCNs,

$$\mathbf{X}^G = f_{act}(f_{gcn}(\mathbf{X}^C, \mathbf{A}; \eta_{gcn})), \quad \mathbf{X}^G \in R^{c \times z}, \tag{14}$$

where η_{gcn} is a GCN neural network model, and f_{act} is a activation function (We employ sigmoid function in this research).

Category association graph contains $\mathbf{X}^C$ and $\mathbf{A}$ in Fig.3(b). $\mathbf{X}^C \in R^{c \times q}$ is the feature matrix of the category. Each category is presented using q dimension word vector, and c represents the number of categories. $\mathbf{A} \in R^{c \times c}$ is an asymmetric association matrix between categories. The symbol $\mathbf{X}^G$ is a matrix with category information, z is the dimension of the dynamic feature map.

In our experiments, the WWW [1] dataset has been trained using the MF-Net [36] model pre-trained on kinetics [40]. If the clip of each video input is 16 × 224 × 224, we can obtain the 8 × 7 × 7 × 768 feature map from the 'conv5' layer. Then, average pooling (AP) is employed to obtain 3D dynamic feature $\mathbf{X}^D$

$$\mathbf{X}^D = f_{ap}(f_{3D}(\mathbf{X}_{1:\tau}^I, \eta_{3d})) \in R^z, \tag{15}$$

where η_{3d} represents 3D neural network model, $z = 768$.

Thus, the motion trend feature ($\mathbf{X}^M$) is obtained

$$\mathbf{X}^M = \mathbf{X}^G \otimes \mathbf{X}^D. \tag{16}$$

We assume that the ground truth label of the video is represented as $\mathbf{G} \in R^{n\times c}(\mathbf{G}_{ij} \in \{0, 1\})$. The LSTM network is trained with the following loss function on the basis of X_t^L features

$$\mathcal{L} = \sum_{i=1}^{c}\sum_{j=1}^{n} \boldsymbol{G}_{ij} log(sig(\tilde{\boldsymbol{G}}_{ij})) + (1 - \boldsymbol{G}_{ij}) log(1 - sig(\tilde{\boldsymbol{G}}_{ij})), \tag{17}$$

where $sig(\bullet)$ is the sigmoid function [29], and $\tilde{\boldsymbol{G}}_{ij}$ represents the prediction value. In process of inference, general categories obtain probability information through the trained LSTM model. For small sample categories, the map of the previous layer of sigmoid function in LSTM network is taken as a video feature ($\mathbf{X}$), and then subspace classifier is used for class prediction.

IV. EXPERIMENT

In this section, we first describe evaluation metrics and implementation details. Secondly, we report the experimental results of the multi-label crowd dataset WWW. Then, subspace classification results are analyzed before we further conduct ablation study to evaluate the key aspects of the proposed approach.

A. EVALUATION METRICS AND IMPLEMENTATION DETAILS

In all experiments, we use the area under receiver operating characteristic (ROC) curve (AUC) and average precision (AP) as the evaluation indicators. AUC is a popular classification indicator for measuring the classifier performance. AP can be effectively used for measuring classification results of each category. To fairly compare with existing methods, we also adopt AP and AUC on each category.

The parameters β_1, β_2 in formula (2), and γ, λ in formula (9) of the subspace generated by the special category are set as follows during the training process: $\gamma \in (0, 5000)$, $\beta_2 < \beta_1$ and $\beta_1, \beta_2 \in (0, 1)$, $\lambda \in \{0.3, 0.5, 5 \times 10^{-3}, 5 \times 10^{-9}\}$.

In the video feature representation branch, ResNet50 [28] is adopted as the static feature extraction backbone, which is pre-trained on ImageNet [29]. During training, we select 75 frames from each video (the same frame length as [3]), with the resolution of the input video frame as 224×224. In terms of dynamic feature representation, the 3D network MF-Net [36] is used, which is pre-trained on kinetics [40]. The 3D network is trained for 100 epochs, wherein input is $\tau = 16$ frames.

We use 3D dynamic information, category feature representation and association matrix as input to train the GCN network. According to the trained model, the product of the 3D dynamic feature and the output of the last layer of GCN is used as feature with semantic information. In the process of training, according to literature [10], we used two GCN layers with the output dimensionality of 384 and 768, respectively. As is depicted in Fig.3(b), the input of GCN includes the category feature matrix and association matrix $\mathbf{A}$. Each category feature is represented by a q dimensional word vector ($q = 300$). We get a word vector model by training the Wikipedia dataset [41] and categories of WWW together. For network optimization, SGD is used as the optimizer. The momentum is set to be 0.9. Weight decay is 10^{-4}. The learning rate is 0.001.

Our LSTM network (in Fig.2) consists of two LSTM layers with output dimensionality of 512 and 768, respectively. In the process of training, the regular dropout is set to 0.4 in LSTM network, thus speeding up the convergence in experiments. Adam is used as the optimizer for network optimization, while the binary cross entropy is adopted as the loss function.

B. EXPERIMENTAL RESULTS

In this part, we present a comparison of our proposed method with state-of-the-art methods on dataset WWW first. Then, quantitative evaluation results of dataset WWW are reported.

1) A COMPARISON OF OUR METHOD WITH STATE-OF-THE-ARTS

WWW [1] is a multi-label large-scale crowded dataset, which contains 10,000 video clips and 94 different categories. According to [1], the dataset is split into training, validation, and test sets at a ratio of 7 : 1 : 2. We use cross-validation on the training and validation sets according to a ratio of 9 : 1. Finally, we evaluate the accuracy of the model on the test sets.

We conduct a comparison of our proposed method with state-of-the-art methods, including DLSF+DLMF [1], DLF+ DLFO+AA [3], S-CNN [2], MIML [5] and CLDF [42](shown in Table 2). In this paper, a model based on the LSTM network (DGSF-LSTM) is proposed, and the feature representation with category association is used as the input of the network.

1) DLSF+DLMF [1]. A deep model is used for learning the features of each category from the appearance and action information of each video, and the learned model is used for identifying unknown categories in the crowded video.
2) DLF+DLFO+AA [3]. The dependence of category information is used for obtaining a mapping relationship between categories and features to achieve better scene classification effect, and a low-level feature extraction mechanism is also used for obtaining more descriptive feature information.
3) S-CNN [2]. A new sliced convolutional neural network is proposed, which exploits 2D filters. Spatial filters obtain appearance information, and time slices capture dynamic clues. This method shows a strong ability to capture spatio-temporal features.
4) MIML [5]. A simple linear model is adopted. By using the relationship between multiple labels, the model can learn a shared space for all labels from the original features, and then trains a label-specific linear model

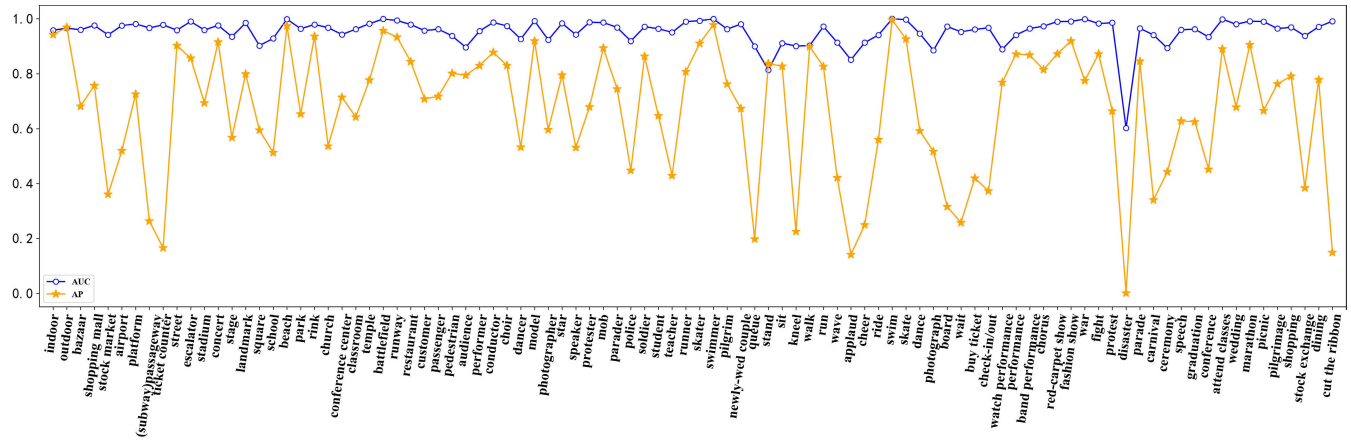

FIGURE 4. Performance of each category in terms of AUC and AP on the WWW dataset. The circle represents the value of AUC, and the five-pointed star represents the value of AP.

TABLE 2. Comparison between our method and other algorithms on the WWW dataset.

Methods		CNN Backbone	mAUC	mAP
DLSF+DLMF [1]		———	0.88	0.412
AA [3]		VGG-16	0.92	0.552
S-CNN [2]		S-CNN	0.940	0.625
(DSF)MIML [5]		R50	0.908	0.634
(DGSF)MIML [5]		R50	0.919	0.638
CLDF [42]		R50	0.957	0.698
Ours	CLDF [42]+subspace	R50	**0.961**	**0.699**
	SF-LSTM	VGG-16	0.928	0.587
	DSF-LSTM	VGG-16	0.937	0.599
	DGSF-LSTM	VGG-16	0.948	0.656
	SF-LSTM	R50	0.941	0.635
	DSF-LSTM	R50	0.950	0.668
	DGSF-LSTM	R50	0.954	0.673
	DGSF-LSTM+ subspace	R50	0.958	0.675

from this space. This linear model reduces the number of parameters and speeds up training. This very linear model to multi-label crowd behavior recognition is applied in the current paper.

5) CLDF [42]. The class-level difficulty factors for multi-label classification are proposed in the literature. We reproduced experiments according to the idea of [42]. ResNet50 model [28] pre-trained on Image-Net [29] was adopted as backbone. Then, we retrained the network on basis of dataset WWW [1]. Finally, classification results are 95.7% and 69.8% respectively in terms of mAUC and mAP. Based on it, the subspace idea was used to predict small sample categories (CLDF [42] + subspace). Both mAUC and mAP results (mAUC, 0.961, mAP, 0.699) outperformed results of [42]. It could be concluded from experiments that the subspace idea proposed in this paper is effective against unbalanced sample data sets. Simultaneously, we compared CLDF [42]+subspace method with ours (DGSF-LSTM+subspace) on 94 categories of WWW-datasets. Experiments show that features obtained by LSTM perform better in categories of action information, such as stand, sit, walk, run, swim, dance, photography, dining, shopping, *et al*. However, CLDF [42] has more advantages in categories of scene information, such as indoor, outdoor, airport, street, stadium, concern, square, beach, school, *et al*.

2) QUANTITATIVE EVALUATION

Quantitative evaluation results are shown in Table 2, showing that our model is efficient for crowded scene classification. In particular, the performance of our subspace model (DGSF-LSTM + subspace) in terms of the mean AUC reaches the state-of-the-art level.

To evaluate our feature representation, other comparison experiments are also conducted, as demonstrated in Table 2. Under the condition that the pre-trained model is VGG-16 [27] (especially mAP), the proposed DGSF (the fusion of static features and dynamic features with category information) is 7% higher than SF (static features) and 6% higher than DSF (the fusion of static features and dynamic features), indicating that category association can facilitate the recognition of video action. Meanwhile, the experiments under ResNet50 and VGG-16 display that the ResNet50 network is superior to VGG-16.

We are also interested in the performance of each category. Fig.4 shows the AUC and AP values for all categories through the DGSF-LSTM method. Some categories (like ''indoor'', ''Outdoor'', ''Street'' and ''Performance'') shown in the line chart can obtain better classification results through the fusion of static features and dynamic features with category information (DGSF), which means categories with a relatively large number of samples can converge easily in the process of training. However, as for categories with only a few samples, such as ''police'', ''queue'' and ''disaster'', we cannot obtain good classification results using LSTM network and sigmoid activation function.

C. SUBSPACE CLASSIFICATION

In this part, the subspace classification results are mainly reported. To analyze and understand crowd behaviors, in the

TABLE 3. The scores of AP and AUC of few samples categories are under the subspace classifier on the WWW dataset.

AP / categories / methods	stockM	subway	ticketC	classR	police	queue	knell	board	disaster	attendC	picnic	stockE	cutTR	mAP
MIML [5]	0.139	0.206	0.125	**0.586**	**0.422**	**0.095**	0.207	0.391	0.001	**0.861**	0.683	**0.127**	**0.014**	0.297
MIML [5] + subspace	**0.239**	**0.228**	**0.208**	0.583	0.277	0.08	**0.221**	**0.465**	**0.008**	0.86	**0.69**	0.114	0.008	**0.307**
CLDF [42]	**0.343**	0.281	0.171	0.698	0.170	**0.310**	**0.320**	0.312	0.001	0.814	**0.924**	0.189	0.042	0.352
CLDF [42] + subspace	0.338	**0.285**	**0.207**	**0.699**	**0.177**	0.281	0.266	**0.318**	**0.009**	**0.816**	0.877	**0.333**	**0.049**	**0.358**
DGSF-LSTM(ours)	**0.36**	0.263	**0.165**	0.642	**0.447**	0.197	0.225	0.315	0.001	**0.89**	0.665	**0.384**	0.149	0.362
DGSF-LSTM+subspace	0.314	**0.376**	0.135	**0.653**	0.431	**0.202**	**0.234**	**0.429**	**0.009**	0.887	**0.748**	0.248	**0.214**	**0.375**

AUC / categories / methods	stockM	subway	ticketC	classR	police	queue	knell	board	disaster	attendC	picnic	stockE	cutTR	mAUC
MIML [5]	0.805	0.941	0.946	0.903	0.852	0.794	**0.839**	**0.988**	0.421	**0.962**	**0.980**	0.793	**0.854**	0.852
MIML [5] + subspace	**0.901**	**0.953**	**0.964**	**0.914**	**0.885**	**0.797**	0.817	0.974	**0.935**	0.939	0.979	**0.798**	0.707	**0.889**
CLDF [42]	0.9478	**0.960**	**0.974**	**0.981**	0.922	0.912	0.880	0.969	0.636	**0.997**	**1.0**	0.891	0.957	0.925
CLDF [42] + subspace	**0.955**	0.956	0.960	0.959	**0.925**	**0.917**	**0.915**	**0.977**	**0.938**	0.994	0.998	**0.942**	**0.979**	**0.955**
DGSF-LSTM(ours)	0.941	0.966	**0.977**	0.962	0.918	0.899	0.9	0.971	0.601	**0.998**	0.989	0.937	0.991	0.927
DGSF-LSTM+subspace	**0.943**	**0.97**	0.974	**0.964**	**0.963**	**0.914**	**0.914**	**0.988**	**0.945**	0.995	**0.996**	**0.938**	**0.996**	**0.961**

existing studies [1]–[3], only experiments on the dataset WWW are conducted, so we apply the idea of subspace to the dataset WWW in this paper. The research goal of this paper is the crowd behavior recognition, however, the existing crowd datasets WorldExpo'10 [43], UCF-CC-50 [44], UCF-QNRF [45] and Shanghaitech [46] being applied for crowd counting. Other crowd datasets, such as S-Hock [47] and Violent-Flows [48], do not refer to multi-label. From our perspective, the dataset WWW is not enough, and the image-based multi-label dataset VOC 2007 (an imbalanced image dataset) is selected to illustrate the effectiveness of the subspace idea [37]. In the following sections, we will specify subspace classification results on the datasets WWW and VOC 2007.

1) SUBSPACE CLASSIFIER ON THE DATASET WWW

Firstly, categories with few samples in the training set, shown in Table 3, are determined according to formula (1). Then, according to formula (2), a subspace for each small sample category is constructed. For example, the subspace of category 'knell' has 76 categories while the original space has 94. Finally, according to formula (5) − (11), the subspace classifier is designed to optimize the corresponding subspace of category 'knell'. As demonstrated in the Table 3, the performance of category 'knell' has been significantly improved.

In addition, we also apply the idea of subspace to the classifier MIML [5] (MIML + subspace) and report experiment results. MIML is a multi-label classifier with category association, which can directly to integrate subspace idea. Table 3 indicates that the performance of most categories has shown a significant improvement by adopting the idea of subspace. However, the experiment result of DGSF-LSTM + subspace is superior to that of MIML+subspace, which shows that MIML [5] has certain limitations on categories with fewer samples. The mAUC of these categories is increased by 3.4% overall while the mAP of those is increased by 1.3% through the DGSF-LSTM+subspace. However, the performance of category 'attend classes' has declined after exploiting subspace. Since the category has achieved decent performance on the DGSF-LSTM model (mAUC is 99.8%), other classifiers will not be able to play a role in improving performance. CLDF [42] uses a variety of skills to extract video features suitable for multi-label classification. For the prediction of small sample categories, we use subspace classifier based on the features of CLDF. Experimental results show that subspace classifier can also be applied to the basis of CLDF features.

2) SUBSPACE CLASSIFICATION RESULTS ON THE DATASET VOC 2007

PASCAL Visual Object Classes Challenge (VOC 2007) [37] is another popular dataset for multi-label recognition. It contains 9, 963 images from 20 object categories, which is divided into train, val and test sets. For fair comparisons, we use the trainval set to train our model, and evaluate the recognition performance on the test sets.

To evaluate the subspace idea, we conduct the same experiments on the dataset VOC 2007. Firstly, we select categories with fewer samples according to formula (1). These categories are shown in Table 4. Then, according to formula (2), we construct subspace for each category. Finally, we retrain the model SSGRL [6] by using the generated subspace of per category and obtain classification result. It is obvious in Table 4 that most of categories adopting subspace have better performance than the original model. Overall, the mAP of these categories is increased by 0.5%.

D. ABLATION STUDIES

In this section, we perform ablation study from three different aspects, including the number of video frames, the value of the parameter γ and λ and the replacement of the LSTM network structure with Gated Recurrent Unit (GRU).

1) VIDEO STATIC FRAMES

In order to fairly compare the test results, the number of static frames for the video in the test is the same as the number of frames in [1], [3], which are both 75. In this paper, we also

TABLE 4. Comparison of AP and mAP of the SSGRL and SSGRL+subspace on the PASCAL VOC 2007 dataset.

Methods	sofa	sheep	plant	motorbike	table	cow	bus	boat	bicycle	bottle	mAP
SSGRL[11]	0.806	0.97	0.834	**0.969**	**0.859**	**0.967**	0.94	0.975	0.973	0.83	0.912
SSGRL[11]+subspace	**0.828**	**0.977**	**0.838**	0.965	0.853	0.965	**0.952**	**0.978**	**0.977**	**0.837**	**0.917**

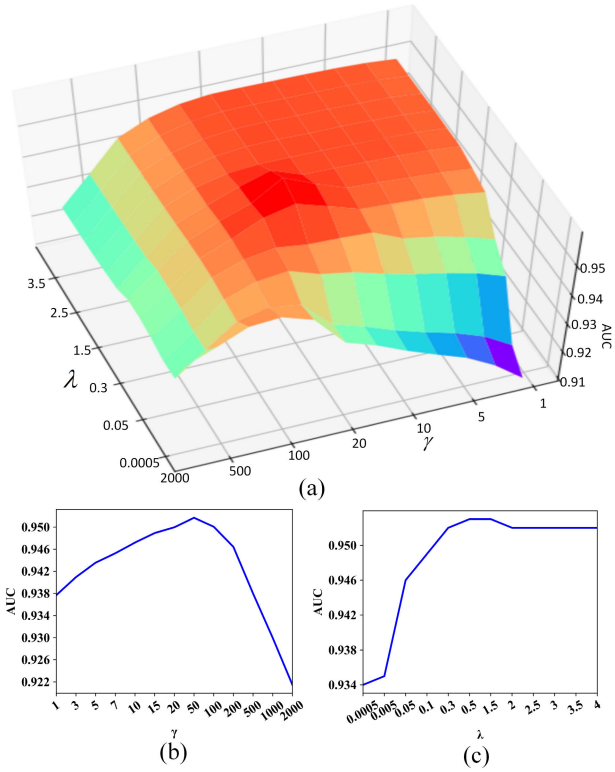

FIGURE 5. The score of AUC with different values of parameter γ and λ.

report the test results of other frames. The current results are based on the (DGSF-LSTM) model. Table 5 shows the mAUC and mAP of 5, 15, 25, 50 and 75 frames. However, the mAUC of 5 frames is only 0.12% less than the 95.4% of 75 frames. It shows that by reducing the number of frames and appropriately increasing the span of the video frames, relatively good classification results will be obtained.

However, the efficiency advantage of using fewer video frames is greater. In the case of 2080ti GPU configuration and 75 frames strategy, it takes 1224ms for a short video to complete category prediction. But it only takes 94ms for a video to complete category prediction with the strategy of 5 frames.

2) EFFECTS OF DIFFERENT PARAMETER VALUES

Results shown in Fig.5 are the average AUC of small sample categories(obtained according to formula (1)). We varied values of parameters γ and λ at the same time and tested them in training set, and plotted final results by a 3D mesh graph (shown in Fig.5(a)). In order to explore changing trend of the two parameters, we plotted cross-sections of the middle grid in two different directions in 3D grid (shown in Fig.5(b) and Fig.5(c)). In addition, we set empirically values of β_1 and β_2 (in formula (2)) to 0.05 and 0.01, respectively.

TABLE 5. The values of mAUC and mAP for different frames on the WWW dataset.

frame	CNN Backbone	RNN cell	mean AUC	mean AP
5	R50	LSTM	0.953	0.661
		GRU	0.955	**0.675**
15	R50	LSTM	0.9524	0.671
		GRU	0.955	0.669
25	R50	LSTM	0.9515	0.666
		GRU	0.953	0.668
50	R50	LSTM	0.95	0.654
		GRU	0.955	0.661
75	R50	LSTM	0.954	0.673
		GRU	**0.957**	0.673

3) GRU STRUCTURE

In our work, we also conduct some experiments under the GRU structure. GRU is a variant of LSTM, which combines the forget gate and the input gate into a single update gate. The parameter of GRU is less than LSTM. The final model is simpler than the LSTM model and is not easy to overfit. Experiments show that the results under the GRU structure (mAP is 67.3% and mAUC is 95.6%) are higher than the LSTM, when the number of frames is 75. It could be down to the fact that convergence is easier on the premise of GRU when the number of samples is insufficient.

V. CONCLUSION

Multi-label action recognition entails prediction of labels co-occurring in videos. Due to the imbalanced distribution of samples, some categories with fewer samples do not converge easily. In order to solve this problem, we introduced an idea of subspace. The subspace method not only harness association among categories, but also simplify the distribution of categories. Meanwhile, a classifier based on subspace is also designed for better classification results. In addition, in crowded scenes, to obtain discriminative feature representation, we injected dependence relationship among categories into dynamic information, strengthening the latter with a stronger semantic relationship. Then, dynamic features with strong semantic correlation are fused into frame-level static features.

In conclusion, the association information of categories is utilized to compensate for the imbalance of samples. Our method can be regarded as a fundamental technique that shows potentials of other related applications. For example, a new category is added in the recognition process, but the number of samples for this new category is insufficient. In this case, we can construct a subspace for the current category based on the idea of subspace, and address the problem of imbalance samples under the guidance of association relationship among categories.

REFERENCES

[1] J. Shao, K. Kang, C. C. Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 4657–4666, doi: 10.1109/CVPR.2015.7299097.

[2] J. Shao, C. C. Loy, K. Kang, and X. Wang, "Slicing convolutional neural network for crowd video understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 5620–5628, doi: 10.1109/CVPR.2016.606.

[3] C. Deng, Z. Cao, Y. Xiao, H. Lu, K. Xian, and Y. Chen, "Exploiting attribute dependency for attribute assignment in crowded scenes," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1325–1329, Oct. 2016.

[4] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 951–958, doi: 10.1109/CVPR.2009.5206594.

[5] S.-J. Huang, W. Gao, and Z.-H. Zhou, "Fast multi-instance multi-label learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2614–2627, Nov. 2019, doi: 10.1109/TPAMI.2018.2861732.

[6] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 522–531.

[7] C. Feichtenhofer, A. Pinz and R. P. Wildes, "Spatio-temporal residual networks for video action recognition," in *Proc. NIPS*, 2016, pp. 3468–3476.

[8] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014, pp. 568–576.

[9] W. Huang, L. Fan, M. Harandi, L. Ma, H. Liu, W. Liu, and C. Gan, "Toward efficient action recognition: Principal backpropagation for training two-stream networks," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1773–1782, Apr. 2019, doi: 10.1109/TIP.2018.2877936.

[10] Z. Chen, X. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. CVPR*, Jun. 2019, pp. 5177–5186.

[11] P. Favaro, R. Vidal, and A. Ravichandran, "A closed form solution to robust subspace estimation and clustering," in *Proc. CVPR*, Providence, RI, USA, Jun. 2011, pp. 1801–1807, doi: 10.1109/CVPR.2011.5995365.

[12] M. Abavisani and V. M. Patel, "Multimodal sparse and low-rank subspace clustering," *Inf. Fusion*, vol. 39, pp. 168–177, Jan. 2018.

[13] Y. Chen, L. Zhang, and Z. Yi, "Subspace clustering using a low-rank constrained autoencoder," *Inf. Sci.*, vol. 424, pp. 27–38, Jan. 2018.

[14] Z. Xue, J. Du, D. Du, and S. Lyu, "Deep low-rank subspace ensemble for multi-view clustering," *Inf. Sci.*, vol. 482, pp. 210–227, May 2019.

[15] C. G. Li, C. You, and R. Vidal, "Structured sparse subspace clustering: A joint affinity learning and subspace clustering framework," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2988–3001, Jun. 2017.

[16] R. Heckel and H. Bolcskei, "Robust subspace clustering via thresholding," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 6320–6342, Nov. 2015.

[17] H. Hu, Z. Lin, J. Feng, and J. Zhou, "Smooth representation clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3834–3841.

[18] C. Ding, T. Li, and M. I. Jordan, "Nonnegative matrix factorization for combinatorial optimization: Spectral clustering, graph matching, and clique finding," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 183–192, doi: 10.1109/ICDM.2008.130.

[19] D. Tolić, N. Antulov-Fantulin, and I. Kopriva, "A nonlinear orthogonal non-negative matrix factorization approach to subspace clustering," *Pattern Recognit.*, vol. 82, pp. 40–55, Oct. 2018.

[20] H. Chen, W. Wang, X. Feng, and R. He, "Discriminative and coherent subspace clustering," *Neurocomputing*, vol. 284, pp. 177–186, Apr. 2018.

[21] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, and X. Geng, "Binary relevance for multi-label learning: An overview," *Frontiers Comput. Sci.*, vol. 12, no. 2, pp. 191–202, Apr. 2018.

[22] H. Zhang, W. Zhang, W. Liu, X. Xu, and H. Fan, "Multiple kernel visual-auditory representation learning for retrieval," *Multimedia Tools Appl.*, vol. 75, no. 15, pp. 9169–9184, Aug. 2016.

[23] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 428–441.

[24] J. Zhang and H. Hu, "Deep spatiotemporal relation learning with 3D multi-level dense fusion for video action recognition," *IEEE Access*, vol. 7, pp. 15222–15229, 2019, doi: 10.1109/ACCESS.2019.2895472.

[25] J. Wang, W. Wang, and W. Gao, "Multiscale deep alternative neural network for large-scale video classification," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2578–2592, Oct. 2018, doi: 10.1109/TMM.2018.2855081.

[26] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2018, doi: 10.1109/ACCESS.2017.2778011.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2014, pp. 1–14.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[30] Q. Liu, X. Che, and M. Bie, "R-STAN: Residual spatial-temporal attention network for action recognition," *IEEE Access*, vol. 7, pp. 82246–82255, 2019, doi: 10.1109/ACCESS.2019.2923651.

[31] Z. Liu and H. Hu, "Spatiotemporal relation networks for video action recognition," *IEEE Access*, vol. 7, pp. 14969–14976, 2019, doi: 10.1109/ACCESS.2019.2894025.

[32] Y. Xiao, J. Chen, Y. Wang, Z. Cao, J. Tianyi Zhou, and X. Bai, "Action recognition for depth video using multi-view dynamic images," *Inf. Sci.*, vol. 480, pp. 287–304, Apr. 2019.

[33] J. Zhang, Y. Han, J. Tang, Q. Hu, and J. Jiang, "Semi-supervised Image-to-Video adaptation for video action recognition," *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 960–973, Apr. 2017, doi: 10.1109/TCYB.2016.2535122.

[34] Z. Gao, T. T. Han, L. Zhu, H. Zhang, and Y. Wang, "Exploring the cross-domain action recognition problem by deep feature learning and cross-domain learning," *IEEE Access*, vol. 6, pp. 68989–69008, 2018, doi: 10.1109/ACCESS.2018.2878313.

[35] Y. Hou, S. Wang, P. Wang, Z. Gao, and W. Li, "Spatially and temporally structured global to local aggregation of dynamic depth information for action recognition," *IEEE Access*, vol. 6, pp. 2206–2219, 2018, doi: 10.1109/ACCESS.2017.2782258.

[36] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "Multi-fiber networks for video recognition," in *Proc. ECCV*, 2018, pp. 352–367.

[37] M. Everingham, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[38] B. Romera Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2152–2161.

[39] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. ICLR*, 2017, pp. 1–10.

[40] 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. 4724–4733.

[41] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[42] M. Marsden, K. McGuinness, J. Antony, H. Wei, M. Redzic, J. Tang, Z. Hu, A. Smeaton, and N. E. O'Connor, "Investigating class-level difficulty factors in multi-label classification problems," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, London, Jul. 2020, pp. 1–6, doi: 10.1109/ICME46284.2020.9102798.

[43] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 833–841.

[44] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2547–2554.

[45] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *ECCV*, 2018, pp. 532–546.

[46] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. CVPR*, Jun. 2016, pp. 589–597.

[47] D. Conigliaro, P. Rota, F. Setti, C. Bassetti, N. Conci, N. Sebe, and M. Cristani, "The S-HOCK dataset: Analyzing crowds at the stadium," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2039–2047.
[48] T. Hassner, Y. Itcher, and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 1–6.

CHUNHUA DENG received the Ph.D. degree in pattern recognition and intelligent systems from the School of Automation, Huazhong University of Science and Technology, Wuhan, China, in 2016. He is currently an Associate Professor with the School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan. His current research interests include computer vision, pattern recognition, and machine learning.

XIAOGE KANG received the B.S. degree from the Wuhan University of Science and Technology, Wuhan, China, in 2013, where she is currently pursuing the M.S. degree with the School of Computer Science and Technology. Her current research interests include computer vision and machine learning.

ZIQI ZHU (Member, IEEE) received the B.S. degree in computer science from Wuhan University, Wuhan, China, in 2005, and the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, in 2011. He is currently an Associate Professor with the School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan. His current research interests include scheduling, machine learning, pattern recognition, and computer vision.

SHIQIAN WU (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 1985 and 1988, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2001. From 1988 to 1997, he was an Assistant Professor, a Lecturer, and an Associate Professor with HUST. From 2000 to 2014, he was a Research Fellow or Research Scientist with the Agency for Science, Technology and Research, Singapore. He has been the Director of the Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System since 2019. He is currently a Professor with the School of Machinery and Automation, Wuhan University of Science and Technology, Wuhan. He has authored or coauthored two books and over 180 scientific publications (book chapters and journal/conference papers). His current research interests include image processing, pattern recognition, machine vision, and artificial intelligence.