Dynamic Texture Recognition With Video Set Based Collaborative Representation

Jin Xie[a,b], Yi Fang[a,b]

[a] *Department of Electrical and Computer Engineering, New York University Abu Dhabi, UAE*
[b] *New York University Multimedia and Visual Computing*

Abstract

Efficient feature description and classification of dynamic texture (DT) is an important problem in computer vision and pattern recognition. Recently, the local binary pattern (LBP) based dynamic texture descriptor has been proposed to classify DTs by extending the LBP operator used in static texture analysis to the temporal domain. However, the extended LBP operator cannot characterize the intrinsic motion of dynamic texture well. In this paper, we propose a novel video set based collaborative representation dynamic texture classification method. First, we divide the dynamic texture sequence into subsequences along the temporal axis to form the video set. For each DT, we extract the video set based LBP histogram to describe it. We then propose a regularized collaborative representation model to code the LBP histograms of the query video sets over the LBP histograms of the training video sets. Finally, with the coding coefficients, the distance between the query video set and the training video sets can be calculated for classification. Experimental results on the benchmark dynamic texture datasets demonstrate that the proposed method can yield good performance in terms of both classification accuracy and efficiency.

Keywords: Dynamic texture classification, local binary pattern, texture feature extraction, collaborative representation

1. Introduction

Dynamic textures (DT) are textures with motion [1]. They are video sequences of moving scenes, which vary not only on the spatial distributions of intensity, but also on the dynamics over time. There are such video sequences in the real word, for example, the sequences of forest fire, waterfall, swarm of birds and humans in crowds, etc. In recent years, the study of DT has been receiving considerable attention, including DT modeling, synthesis, segmentation and classification. Effective feature extraction is a key step for DT classification. It is desirable that effective DT feature can characterize texture appearance and motion well. Also, the extracted DT feature can also be applied to video understanding. For example, the DT feature can be used to characterize appearance and motion of human for human action recognition [2, 3, 4, 5].

Different from static textures, dynamic textures exhibit certain stationary properties in the temporal

Email addresses: jin.xie@nyu.edu (Jin Xie), yfang@nyu.edu (Yi Fang)

domain. One challenging problem in DT classification is how to discriminatively describe DT for classification, i.e., texture appearances in the spatial domain and dynamics in the temporal domain. Existing approaches characterize DT by either patterns of motion field in DT, or the linear dynamic system or the combination of texture appearance features and motion features. The patterns of motion field [6, 7] of DTs are discriminative and thus very useful for DT classification. In [8], the normal flow features are combined with periodicity features for DT, attempting to explicitly characterize motion magnitude, directionality and periodicity. Based on the velocity and acceleration fields of DT, Lu *et al.* [9] constructed the spatio-temporal multi-resolution histograms of the velocity and acceleration fields for DT classification, where image sequences at different spatio-temporal scales are estimated with the structure tensor method. In [10], the normal flow features and regularized complete flow features are compared, and it is concluded that normal flow contains information on both shape and dynamics of DT.

Doretto *et al.* [11] showed that the spatial appearance and dynamics of DT can be modeled by a linear dynamic system (LDS). Thus, DT sequences can be discriminated with different dynamic parameters of the LDS. Based on the LDS representation, several methods have been proposed, where different distances between the model parameters of two LDSs are defined. In [12], the Binet-Cauchy kernel is used to compare the parameters of the two LDSs. Chan and Vasconselos [13] employed kernel PCA to learn a non-linear kernel dynamic texture and used the Martin distance to measure the similarity between the kernel dynamic textures for DT classification. However, these methods cannot deal with DT sequences with viewpoint changes. In order to handle viewpoint changes of DTs, the dictionary of the parameters of the LDS [14, 15, 16] is learned with K-means clustering and the bag-of-feature representation is used for DT classification.

Since dynamic texture can be viewed as texture with motion, researchers combined the texture appearance features and the motion features to characterize DT. Some local operator based static texture feature extraction methods [17, 18, 19, 20, 21], therefore, are extended to DT. In [19, 20], the fractal dimension of the 2D slice of the 3D spatio-temporal volume was proposed to characterize the self-similarities of DTs for classification. In [21, 22, 23], the distributions of 3D Gaussian third derivative filter are used to characterize the dynamic structure of DT. The volume LBP (VLBP) [24, 25] is an extension of the LBP descriptor [17, 26, 27] widely used in static texture analysis by combining texture appearance and motion. The VLBP descriptor compares each pixel with its neighborhoods in the previous, current and posterior frames to encode each pixel as a sequence of binary codes. Then, three sequences of binary codes from the three frames are concatenated to form the VLBP descriptor to describe the local structure in DT. When the number of neighborhood points increases, the number of VLBP becomes very large. To make the VLBP computationally simple, LBP histograms from three orthogonal planes (LBP_TOP) [25], i.e., the XY plane, XT plane and YT plane, are extracted to characterize DT. Then the LBP histograms from the three planes are concatenated to form the LBP_TOP descriptor.

Inspired by the success of collaborative representation in face recognition [28, 29], in this paper we develop a novel collaborative representation based DT classification method. Our proposed method is

based on the LBP based DT descriptor. In the proposed method, in order to extract robust features from DT with complex motions, we first model the DT sequence by the video set, which is formed by dividing the sequence into subsequences in the temporal domain. And each DT sequence can be represented by the video set based LBP descriptors. Considering the similarities between motion patterns of different DTs, we then propose a regularized collaborative representation model to represent the query DT sequence over all training DT sequences with the video set based descriptor. Finally, we classify the query DT sequence by the minimal representation residual to the training DT sequences. Experimental evaluations on the benchmark dynamic texture datasets demonstrate the effectiveness of the proposed method. The main difference from the previous studies lies in that we model DT with the video set and use the video sets to collaboratively represent DT in order to capture the complex motion.

The rest of the paper is organized as follows. Section 2 presents in detail the proposed collaborative representation based DT classification scheme. Section 3 conducts experiments on the benchmark dynamic texture datasets and Section 4 concludes the paper.

2. Proposed Method

In this section, we describe in detail the proposed DT classification approach. There are three major components in our framework: video set based DT descriptor, video set based collaborative representation and DT classification. In Section 2.1, we present video set based DT descriptor. Moreover, we point out that the original LBP based DT descriptor is a special case of the video set based DT descriptor. The video set based collaborative representation and classification are presented in Section 2.2, respectively.

2.1. Video Set Based DT Descriptor

In this subsection, we propose a video set based DT descriptor to extract DT features in order for capturing the intrinsic motions of DT. Let $\boldsymbol{I}$ denote the DT sequence. We divide the DT sequence $\boldsymbol{I}$ into a sequence of homogeneous videos in the temporal domain to form a video set, $\boldsymbol{I}_1, \boldsymbol{I}_2, \cdots, \boldsymbol{I}_n$, where n is the number of the subsequence videos. Then, the LBP based DT descriptor $\boldsymbol{H}_i$ is employed to represent the subsequence $\boldsymbol{I}_i$, $i = 1, 2, \cdots, n$. Finally, we combine the LBP based DT descriptors of the subsequence videos to form the video set based DT descriptor $\boldsymbol{W}$, $\boldsymbol{W} = \sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{H}_i$, where $\boldsymbol{a}_i$ is the weight of the LBP histogram $\boldsymbol{H}_i$.

Suppose that there are c_i pixels in the homogeneous video $\boldsymbol{I}_i$. The LBP based DT descriptor $\boldsymbol{H}_i$ of the subsequence $\boldsymbol{I}_i$ is denoted, $\boldsymbol{H}_i = [o_{i,1}, o_{i,2}, \cdots, o_{i,s}]$, where $o_{i,r}$ denotes the normalized number of the r^{th} binary pattern, $r = 1, 2, \cdots, s$, and s is the number of binary patterns in the subsequence $\boldsymbol{I}_i$. Let $\boldsymbol{S}$ denote the original LBP based DT descriptor of the whole DT sequence, $\boldsymbol{S} = [p_1, p_2, \cdots, p_s]$, where p_r denotes the normalized number of the r^{th} binary pattern in the whole sequence, $r = 1, 2, \cdots, s$. For each component of the original LBP based DT descriptor, then we have the following equation:

$$p_r = \frac{\sum_{i=1}^{n} c_i o_{i,r}}{\sum_{i=1}^{n} c_i}. \tag{1}$$

where $r = 1, 2, \cdots, s$. Therefore, the original LBP based DT descriptor $\boldsymbol{S}$ can be viewed as a special case of the video set based DT descriptor:

$$\boldsymbol{S} = \frac{\sum_{i=1}^{n} c_i \boldsymbol{H}_i}{\sum_{i=1}^{n} c_i}. \tag{2}$$

The original VLBP descriptor[25] encodes the texture appearance in the consecutive three frames to describe DT. The LBP_TOP descriptor [25] is a simplified version of the VLBP descriptor. It uses LBP histograms obtained independently from three orthogonal planes (i.e., the XY plane, the YT plane and the XT plane) to characterize the texture appearance and dynamics of DT. Then the histograms from the three planes are concatenated into a single histogram to form the LBP_TOP descriptor. From Eq. (2), one can see that in VLBP and LBP_TOP, both LBP histograms are summed with the weight $\boldsymbol{a}_i$, $\boldsymbol{a}_i = \frac{c_i}{\sum_{i=1}^{n} c_i}$, which is highly relative to the pixels in the subsequence. The weight cannot characterize the motion change well. Thus, the original VLBP descriptor cannot describe DT with the complex motions well. For example, for a DT sequence of burning fire changing gradually from a spark to a large fire, it is homogeneous at each time instant but has varying statistics over time. Although the LBP_TOP descriptor can describe the motion of the burning fire at each time instant well, it is difficult to capture the whole dynamic statistics over time. Different from the original LBP based DT descriptor, the weight of the video set based DT descriptor is learned by collaborative representation. In the next subsection, we present the details of employing collaborative representation to learn the weights.

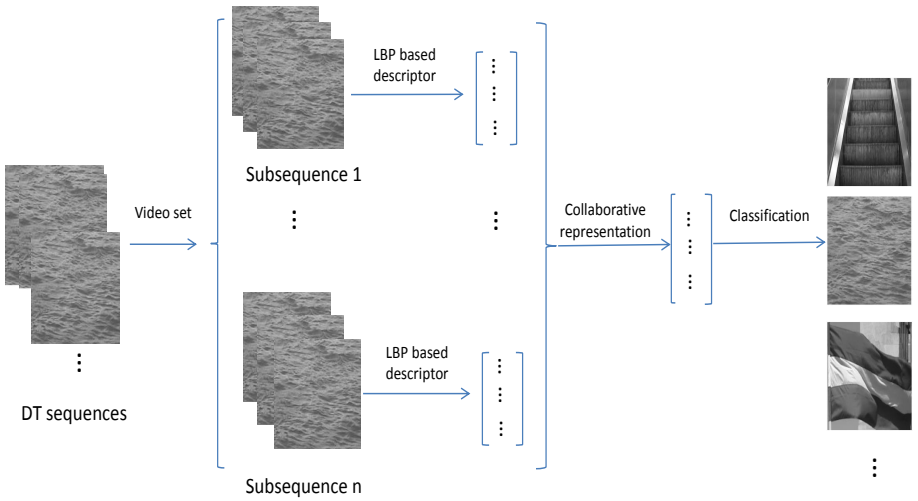


Figure 1: Flow chart of the video set based DT descriptor.

2.2. Video Set Based Collaborative Representation and Classification

Suppose there are C classes of DTs and q training samples (sequences) for each class. Each sequence can be divided into n subsequences to extract VLBP feature to represent it. Thus, we can construct a training set $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_C]$, $\boldsymbol{X}_k = [\boldsymbol{W}_{k,1}, \boldsymbol{W}_{k,2}, \cdots, \boldsymbol{W}_{k,l}] \in \mathcal{R}^{s \times l}$, where $\boldsymbol{W}_{k,j}$ is the VLBP based DT descriptor of the j^{th} homogeneous video from class k, s is the dimension of $\boldsymbol{W}_{k,j}$, $l = nq$ is the number

of the homogeneous videos of class k, $j = 1, 2, \cdots, l$, $k = 1, 2, \cdots, C$. Let $\boldsymbol{Y} = [\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_n]$ denote the video set based DT descriptor of the query DT sequence, where $\boldsymbol{Y}_j$ is the LBP based DT descriptor of the j^{th} subsequence in the video set, $j = 1, 2, \cdots, n$. One fact in DT analysis is that motion patterns from different DTs still have much similarities. Therefore, especially when there are few training samples, it is helpful to use other classes of DT sequences to collaboratively represent the query DT sequence for classification. In addition, for DT with the complex motions, it is difficult to use the single LDS to model the dynamics of the DT sequence. It is desirable to represent the complex dynamics of the query DT with the combination of dynamics from all classes of DTs.

In order to achieve high representation accuracy, we propose the following collaborative representation model:

$$\begin{aligned}(\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}) = argmin_{\boldsymbol{a},\boldsymbol{b}} \|\boldsymbol{Y}\boldsymbol{a} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda_1\|\boldsymbol{a}\|_2^2 + \lambda_2\|\boldsymbol{b}\|_2^2 \\ +\lambda_3 \sum\nolimits_{k=1}^{C} \|\boldsymbol{b}_k - \boldsymbol{e}_1\mu_k\|_2^2 + \lambda_4\|\boldsymbol{e}_2\boldsymbol{a} - \boldsymbol{1}\|_2^2\end{aligned} \tag{3}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are the representation vectors, i.e., the weight vectors of the DT descriptors of the query DT and training DT sequences, $\boldsymbol{a} = [a_1, a_2, \cdots, a_n]$, $\boldsymbol{b} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_C]$; $\boldsymbol{b}_k$ is the representation vector associated with training samples $\boldsymbol{X}_k$; μ_k is the mean of all representations $\boldsymbol{b}_{k,j}$ from class k, i.e., $\mu_k = \frac{1}{l}\sum_{j=1}^{l} b_{k,j}$; a_i is the i^{th} coefficient in $\boldsymbol{a}$; λ_1, λ_2, λ_3 and λ_4 are positive scalars to balance the representation residual and the regularizer; $\boldsymbol{e}_1 \in \mathcal{R}^{n\times 1}$ is a column vector whose elements are one and $\boldsymbol{e}_2 \in \mathcal{R}^{1\times n}$ is a row vector whose elements are one . In the proposed model, the video set based DT descriptor $\boldsymbol{Y}\boldsymbol{a}$ is collaboratively represented over the training DT set $\boldsymbol{X}$. In order to make the solution stable, the regularization terms of l_2-norm of the representation coefficient vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are imposed. To avoid the trivial solution $\boldsymbol{a} = \boldsymbol{b} = \boldsymbol{0}$, the constraint $\sum a_i = 1$ is imposed. In addition, since the training DT sequences from the same class share similarities in texture appearances and dynamics, their representation coefficients should be similar. Therefore, in the proposed model, we enforce the representation coefficient vector $\boldsymbol{b}_k$ to approach to their mean μ_k, i.e., minimizing the regularization term $\sum_{k=1}^{C} \|\boldsymbol{b}_k - \boldsymbol{e}_1\mu_k\|_2^2$. This term is basically to reduce the intra-class variation for more accurate classification.

The optimization of Eq. (3) can be easily conducted by alternatively optimizing $\boldsymbol{a}$ and $\boldsymbol{b}$. With some random initialization of $\boldsymbol{b}$, we could first fix $\boldsymbol{b}$ and update $\boldsymbol{a}$, and the problem in Eq. (4) is converted into a regularized least square problem:

$$\begin{aligned}\hat{\boldsymbol{a}} = argmin_{\boldsymbol{a}} \|\boldsymbol{Y}\boldsymbol{a} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \lambda_1\|\boldsymbol{a}\|_2^2 + \lambda_2\|\boldsymbol{b}\|_2^2 \\ +\lambda_3 \sum\nolimits_{k=1}^{C} \|\boldsymbol{b}_k - \boldsymbol{e}_1\mu_k\|_2^2 + \lambda_4\|\boldsymbol{e}_2\boldsymbol{a} - \boldsymbol{1}\|_2^2.\end{aligned} \tag{4}$$

Let the partial derivative of Eq. (5) with respect to $\boldsymbol{a}$ equal to zero, we have:

$$\hat{\boldsymbol{a}} = (\boldsymbol{Y}^T\boldsymbol{Y} + \lambda_1\boldsymbol{I} + \lambda_4\boldsymbol{e}_2^T\boldsymbol{e}_2)^{-1}(\boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{b} + \lambda_4\boldsymbol{e}_2^T). \tag{5}$$

Once the representation coefficient vector $\boldsymbol{a}$ is updated, we could fix $\boldsymbol{a}$ and update $\boldsymbol{b}$. We update $\boldsymbol{b}_k$ one by one. When updating $\boldsymbol{b}_k$, the other sub-vectors of representation coefficients associated with the

LBP based DT descriptors, $\boldsymbol{b}_j (j \neq k)$, are fixed. The objective function in Eq. (3) is then reduced to the following problem:

$$\hat{\boldsymbol{b}}_k = argmin_{\boldsymbol{b}_k} \|\boldsymbol{Y}\boldsymbol{a} - \sum\nolimits_{j \neq k} \boldsymbol{X}_j \boldsymbol{b}_j - \boldsymbol{X}_k \boldsymbol{b}_k\|_2^2 + \lambda_2 \|\boldsymbol{b}_k\|_2^2 + \lambda_3 \|\boldsymbol{b}_k - \boldsymbol{e}_1 \mu_k\|_2^2. \tag{6}$$

Since $\mu_k = \frac{\boldsymbol{e}_1^T \boldsymbol{b}_k}{n}$, Eq. (6) can be re-written as:

$$\hat{\boldsymbol{b}}_k = argmin_{\boldsymbol{b}} \|\boldsymbol{Y}\boldsymbol{a} - \sum\nolimits_{j \neq k} \boldsymbol{X}_j \boldsymbol{b}_j - \boldsymbol{X}_k \boldsymbol{b}_k\|_2^2 + \lambda_2 \|\boldsymbol{b}_k\|_2^2 + \lambda_3 \|\boldsymbol{b}_k - \frac{\boldsymbol{e}_1 \boldsymbol{e}_1^T \boldsymbol{k}}{n}\|_2^2. \tag{7}$$

Let the partial derivative of Eq. (7) with respect to $\boldsymbol{b}_k$ equal to zero, we have:

$$\hat{\boldsymbol{b}}_k = (\boldsymbol{X}_k^T + \lambda_2 \boldsymbol{I} + \lambda_3 \boldsymbol{I})^{-1} (\boldsymbol{X}_k^T \boldsymbol{Y}\boldsymbol{a} - \sum\nolimits_{j \neq k} \boldsymbol{X}_j \boldsymbol{b}_j - \lambda_3 \boldsymbol{I}). \tag{8}$$

Once all representation coefficient vectors $\boldsymbol{b}_k$ are updated, we then fix $\boldsymbol{b}$ and update the representation coefficient vector $\boldsymbol{a}$ with Eq. (5). Since in each iteration the objective function in Eq. (3) can reduce, the alternative optimization process will converge to a local minimum and the representation coefficient vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are learned.

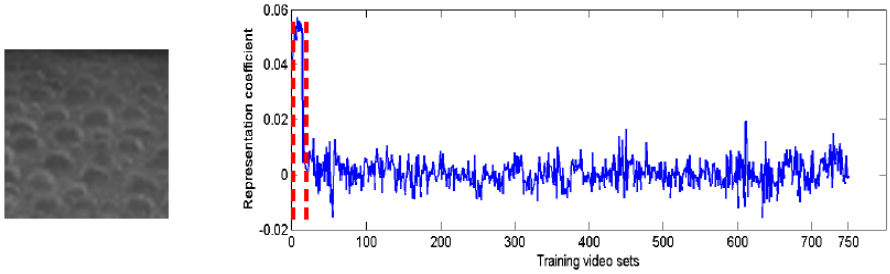
Suppose that the coefficient vectors $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ are obtained by Eqs. (5) and (8), we can write $\hat{\boldsymbol{b}}$ as $\hat{\boldsymbol{b}} = [\hat{\boldsymbol{b}}_1, \hat{\boldsymbol{b}}_2, \cdots, \hat{\boldsymbol{b}}_C]$, where $\hat{\boldsymbol{b}}_k$ is the sub-vector of coefficients associated with training set $\boldsymbol{X}_k$. We use the representation residual err_k of $\boldsymbol{Y}\hat{\boldsymbol{a}}$ over the video set $\boldsymbol{X}_k$ to determine the class label of $\boldsymbol{Y}$:

$$identity(\boldsymbol{Y}) = argmin_k err_k, k = 1, 2, \cdots, C \tag{9}$$

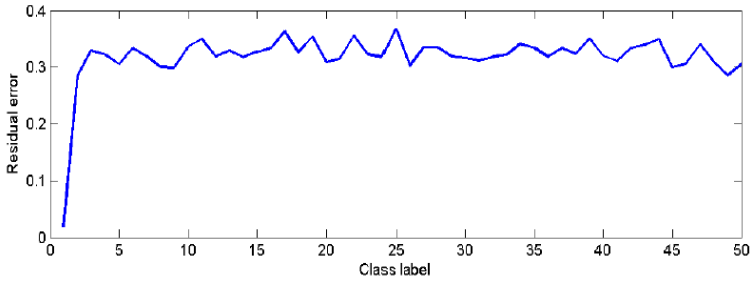
where $err_k = \|\boldsymbol{Y}\hat{\boldsymbol{a}} - \boldsymbol{X}_k \hat{\boldsymbol{b}}_k\|_2^2$.

In our objective function, fixing one variable to optimize the other variable is a regularized least square problem (please refer to Eqs. (4) and (6)). Thus, by fixing one variable, optimizing the other variable is a convex optimization problem and the global minimum can be obtained. Therefore, the convergence of the solution of the objective function Eq. (3) can be guaranteed.

Figs. 2 and 3 illustrate the representation coefficients and residual errors for classification on the UCLA DT dataset [11] with our proposed video set based collaborative representation scheme. For each class, one DT sequence is chosen as the training sample and other three samples are chosen as the query samples. Each DT is divided into 15 homogeneous video sequences to form the video set. Here we choose the boiling water DT from class 1 and the waterfall DT from class 50 as the query DT sequences. The representation coefficient vectors $\boldsymbol{b}$ are plotted in Fig. 2. (a) and Fig. (3).(a), respectively. The highlighted coefficients (marked as the red dot line) in Figs. 2 and 3 are associated with the training DT samples $\boldsymbol{X}_1$ and $\boldsymbol{X}_{50}$, respectively, which are consistent with the class labels of the query DT sequences. It is observed that these coefficients are much more greater than the coefficients associated with other classes. Moreover, err_1 and err_{50} in Fig. 2. (b) and Fig. (3).(b) are the minimal errors among all the training DT sequences, corresponding to the class labels of the query DT sets, respectively.

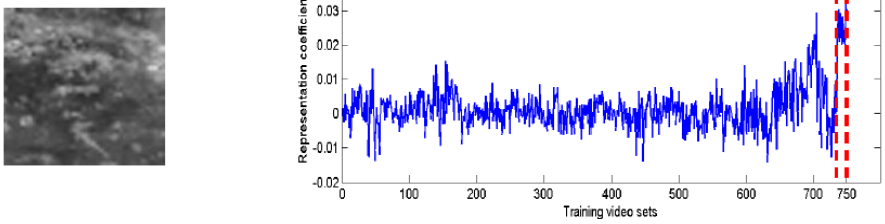


(a) The boiling water DT and the representation coefficients over the training video sets.

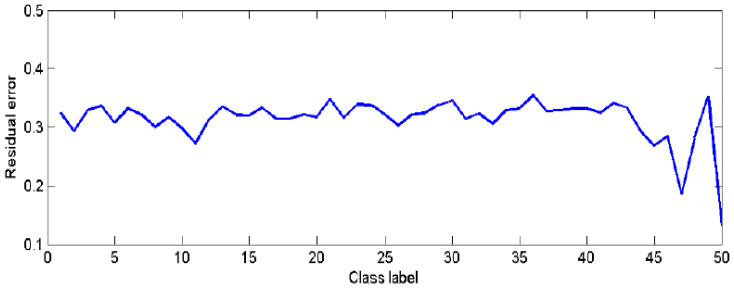


(b) The residual error for the boiling water DT.

Figure 2: The representation coefficients and residual errors for the boiling water DT.



(a) The waterfall DT and the representation coefficients over the training video sets.



(b) The residual error for the waterfall DT.

Figure 3: The representation coefficients and residual errors for the waterfall DT.

Table 1: Algorithm of the proposed video set based collaborative representation.

Input: query set $\boldsymbol{Y}$; training set $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_C]$; λ_1, λ_2, λ_3, λ_4.

Output: the label of query DT sequence $\boldsymbol{Y}$.

Initialize $\boldsymbol{b}$ and $t = 0$.

While $t < max\_num$ do

Update $\boldsymbol{a}$ with Eq. (5);

Update $\boldsymbol{b}$ with Eq. (8);

Set $t = t + 1$.

End while

Compute $\boldsymbol{err}_k = \|\boldsymbol{Y}\hat{\boldsymbol{a}} - \boldsymbol{X}_k\hat{\boldsymbol{b}}_k\|_2^2$.

$Identity(\boldsymbol{Y}) = argmin_k \boldsymbol{err}_k$.

3. Experimental Evaluation

3.1. Dynamic Texture Datasets

To demonstrate the effectiveness of our proposed video set based collaborative representation DT classification method, two representative and benchmark DT datasets are used in our experiments: UCLA and Dyntex. The UCLA dynamic texture dataset [11] contains 50 classes of various texture, including boiling water, fountains, fire, waterfalls and flowers swaying in the wind, etc. Each class of DT contains four grayscale video sequences with 75 frames of 160×110 pixels. We crop each sequence to a 48×48 window that contains the representative motion. The sample DTs in the UCLA dataset are shown in Fig. 4.

The DynTex dataset [30] is a large DT dataset. It contains 35 classes of DTs, including various kinds of DT video sequences, such as escalator, smoke and fountains, etc. The sequences in the DynTex dataset are taken with the scale and rotation changes. Each sequence is a color video with 250 frames with 400×300 pixels and de-interlaced with a spatial-temporal median filter. In this experiment, following the setting in [25], each DT sequence is divided into eight non-overlapping subsequences at different positions, which are randomly selected. In addition, two subsequences are generated from the original sequence randomly cut along the temporal axis. Thus, each original DT sequence generates ten subsequences with different dimensions, which have the same class label of the original DT sequence. Fig. 5 shows some sample DTs in the DynTex dataset.

3.2. Parameter Selection and Implementation Details

The evaluation methodology on the two DT datasets is as follows: l DT sequences are randomly chosen per class for training and the remaining sequences are used for testing. On the UCLA DT dataset, 1, 2 and 3 DT samples per class are randomly chosen to form the training set; On the DynTex dataset, 1, 3 and 5 DT samples per class are randomly chosen for the training set. For each setting, the experiments

Figure 4: Example DT sequences in the UCLA DT dataset.

Figure 5: Example DT sequences in the DynTex DT dataset.

are run 100 times and the average classification accuracy is reported. In the video set based collaborative representation model, parameters λ_1, λ_2, λ_3 and λ_4 are set to 0.001. For each DT sequence in the UCLA and DynTex datasets, 15 and 25 subsequences are generated by dividing the DT sequence into a set of homogeneous videos in the temporal domain, respectively.

3.3. Experimental Results

Evaluation of the proposed method: We denote our proposed video set based collaborative representation method with the VLBP and LBP_TOP descriptors by VSCR_VLBP and VSCR_LBPTOP. In the case of different numbers of training DT samples, We compare the performance of VSCR_LBPTOP on the two DT datasets with the video set based DT descriptor in the case of different neighborhood points and radii, denoted by $VSCR\_LBPTOP_{P_{XY},P_{YT},P_{XT},R_X,R_Y,R_T}$, where P_{XY}, P_{YT} and P_{XT} are the numbers of the neighboring points in the XY, YT and XT planes, R_X, R_Y and R_T are the radii of the neighborhood in the XY, YT and XT planes, respectively. The experimental results on the UCLA and DynTex datasets are listed in Tables 2 and 3, respectively. From the tables, one can see that the number of the neighboring points is relatively large, such as $P_{XY} = P_{YT} = P_{XT} = 8$, different radii have little effects on the final classification accuracy. For example, when $P_{XY} = P_{YT} = P_{XT} = 8$, for one training sample, the proposed method VSCR_LBPTOP with $R_X = R_Y = R_T = 1$ can achieve classification accuracy of 96.35% while VSCR_LBPTOP with $R_X = R_Y = R_T = 3$ can achieve classification accuracy of 97.33% on the UCLA DT dataset. On the Dyntex dataset, when $P_{XY} = P_{YT} = P_{XT} = 8$, for one training sample, the proposed method VSCR_LBPTOP with $R_X = R_Y = R_T = 1$ can achieve classification accuracy of 83.49% while VSCR_LBPTOP with $R_X = R_Y = R_T = 3$ can achieve classification accuracy of 84.40%.

Table 2: Classification rates on the UCLA DT dataset by VSCR_LBPTOP with different neighborhood points and radii.

Training DT samples	1	2	3
$VSCR\_LBPTOP_{4,4,4,1,1,1}$	81.33%	87.4%	90.29%
$VSCR\_LBPTOP_{8,8,8,1,1,1}$	96.35%	97.20%	99.43%
$VSCR\_LBPTOP_{8,8,8,3,3,3}$	97.33%	98.15%	99.45%

Table 3: Classification rates on the DynTex DT dataset by VSCR with different neighborhood points and radii.

Training DT samples	1	3	5
$VSCR\_LBPTOP_{4,4,4,1,1,1}$	81.90%	87.76%	88.57%
$VSCR\_LBPTOP_{8,8,8,1,1,1}$	83.49%	92.24%	95.43%
$VSCR\_LBPTOP_{8,8,8,3,3,3}$	84.40%	93.5%	96.25%

Comparison evaluation: We compare our proposed method with representative and recently developed DT classification methods [13, 21, 31, 20]. In the case of 3 training DT samples, on the UCLA DT dataset, we compare the proposed VSCR_VLBP and VSCR_LBPTOP to the state-of-the-art DT classification methods [21, 13, 20, 31]. Table 4 lists the classification accuracies of the methods. The state-of-the-art classification accuracies are 81.00% [21], 89.50% [13], 99.12% [20] with the nearest neighborhood classifier and 99.00% [31] with maximum margin learning. Our proposed VSCR_VLBP and VSCR_LBPTOP can achieve classification accuracies of 83.21% and 99.43%, respectively. From this table one can see that our proposed VSCR_LBPTOP method can achieve the best performance. Since in the cases of 1 and 2 training DT samples the classification accuracies of [21, 13, 31, 20] are not available, we only compare our proposed method to the LBP based DT methods [25] (VLBP and LBP_TOP). Tables 5 and 6 list the comparison results on the UCLA and DynTex DT datasets. In VLBP and VSRC_VLBP, $P_{XY} = P_{YT} = P_{XT} = 4$ and $R_X = R_Y = R_T = 1$. In LBP_TOP and VSCR_LBPTOP, $P_{XY} = P_{YT} = P_{XT} = 8$ and $R_X = R_Y = R_T = 1$. From these tables, we can see that the proposed VSCR method can achieve the best classification accuracy in almost every case. The VSCR_VLBP and VSCR_LBPTOP methods are superior to the VLBP[25] method and the LBP_TOP[25] methods, respectively. This also implies that collaborative representation from other classes of DTs is helpful to classification. Particularly, on the UCLA DT dataset, compared to the LBP_TOP method, the proposed VSCR_LBPTOP method can significantly improve the classification accuracy. For example, in the cases of 1,2 and 3 training samples, the proposed VSCR_VLBP method can obtain the classification accuracies of 96.35%, 97.20% and 99.43% while the LBP_TOP method can obtain the classification accuracies of 84.19%, 89.40% and 90.29%. In addition, because the number of the neighboring points is small (if the number of the neighboring points is large, the dimension of the LBP histogram is very high), the VSCR_VLBP method cannot obtain good performance on the UCLA and DynTex DT datasets especially when there are limited training DT sequences.

Table 4: Classification rates on the UCLA DT dataset in the case of 3 training DT sequences.

Methods	[21]	[13]	[20]	[31]	$VSCR\_VLBP$	$VSCR\_LBPTOP$
Classification rates	81.00%	89.50%	99.12%	99.00%	83.21%	99.43%

Table 5: Classification rates on the UCLA DT dataset by different methods.

Training DT samples	1	2	3
$VLBP$[25]	68.67%	77.60%	81.00%
$VSCR\_VLBP$	71.33%	80.20%	83.21%
LBP_TOP[25]	84.19%	89.40%	90.29%
$VSCR\_LBPTOP$	96.35%	97.20%	99.43%

Table 6: Classification rates on the DynTex DT dataset by different methods.

Training DT samples	1	3	5
$VLBP$[25]	74.54%	82.31%	86.77%
$VSCR\_VLBP$	79.21%	85.32%	89.71%
LBP_TOP[25]	82.29%	89.40%	93.71%
$VSCR\_LBPTOP$	83.49%	92.24%	95.43%

4. Conclusions

We proposed a video set based collaborative representation method for dynamic texture classification, which is simple to implement and demonstrates promising performance. Each DT sequence was divided into subsequences to form the video set. The video set based DT descriptor with VLBP/LBP_TOP was extracted to represent the DT. Then, a collaborative representation model was proposed to use DT subsequences from all classes to represent the query DT sequence to ensure representation accuracy while reducing the intra-class variance. Finally, We used the representation residual associated with each class of training subsequences to classify DT. The proposed VSCR method was validated on two benchmark DT datasets: UCLA and DynTex. The experimental results showed that our proposed method can achieve good performance.

References

[1] M. Szummer, R. W. Picard, Temporal texture modeling, in: ICIP (3), 1996, pp. 823–826.

[2] L. Liu, L. Shao, X. Li, K. Lu, Learning spatio-temporal representations for action recognition: A genetic programming approach, IEEE T. Cybernetics 46 (1) (2016) 158–170.

[3] L. Shao, X. Zhen, D. Tao, X. Li, Spatio-temporal laplacian pyramid coding for action recognition, IEEE T. Cybernetics 44 (6) (2014) 817–827.

[4] L. Shao, L. Liu, M. Yu, Kernelized multiview projection for robust action recognition, International Journal of Computer Vision 24 (7) (2015) 1–15.

[5] M. Yu, L. Liu, L. Shao, Structure-preserving binary representations for rgb-d action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 99 (7) (2015) 1–15.

[6] P. Bouthemy, R. Fablet, Motion characterization from temporal cooccurrences of local motion-based measures for video indexing, in: ICPR, 1998, pp. 905–908.

[7] R. Fablet, P. Bouthemy, Motion recognition using spatio-temporal random walks in sequence of 2d motion-related measurements, in: ICIP (3), 2001, pp. 652–655.

[8] R. Péteri, D. Chetverikov, Dynamic texture recognition using normal flow and texture regularity, in: IbPRIA (2), 2005, pp. 223–230.

[9] Z. Lu, W. Xie, J. Pei, J. Huang, Dynamic texture recognition by spatio-temporal multiresolution histograms, in: WACV/MOTION, 2005, pp. 241–246.

[10] S. Fazekas, D. Chetverikov, Normal versus complete flow in dynamic texture recognition: a comparative study, in: 4th International Workshop Texture Analysis and Synthesis, 2005, pp. 37–42.

[11] G. Doretto, A. Chiuso, Y. N. Wu, S. Soatto, Dynamic textures, International Journal of Computer Vision 51 (2) (2003) 91–109.

[12] S. V. N. Vishwanathan, A. J. Smola, R. Vidal, Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes, International Journal of Computer Vision 73 (1) (2007) 95–119.

[13] A. B. Chan, N. Vasconcelos, Classifying video with kernel dynamic textures, in: CVPR, 2007.

[14] A. Ravichandran, R. Chaudhry, R. Vidal, View-invariant dynamic texture recognition using a bag of dynamical systems, in: CVPR, 2009, pp. 1651–1657.

[15] A. Ravichandran, P. Favaro, R. Vidal, A unified approach to segmentation and categorization of dynamic textures, in: ACCV (1), 2010, pp. 425–438.

[16] A. Ravichandran, R. Chaudhry, R. Vidal, Categorizing dynamic textures using a bag of dynamical systems, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2) (2013) 342–353.

[17] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[18] Z. Guo, L. Zhang, D. Zhang, A completed modeling of local binary pattern operator for texture classification, IEEE Transactions on Image Processing 19 (6) (2010) 1657–1663.

[19] Y. Xu, Y. Quan, H. Ling, H. Ji, Dynamic texture classification using dynamic fractal analysis, in: ICCV, 2011, pp. 1219–1226.

[20] H. Ji, X. Yang, H. Ling, Y. Xu, Wavelet domain multifractal analysis for static and dynamic texture classification, IEEE Transactions on Image Processing 22 (1) (2013) 286–299.

[21] K. G. Derpanis, R. P. Wildes, Dynamic texture recognition based on distributions of spacetime oriented structure, in: CVPR, 2010, pp. 191–198.

[22] K. G. Derpanis, R. P. Wildes, Spacetime texture representation and recognition based on a spatiotemporal orientation analysis, IEEE Trans. Pattern Anal. Mach. Intell. 34 (6) (2012) 1193–1205.

[23] K. G. Derpanis, M. Lecce, K. Daniilidis, R. P. Wildes, Dynamic scene understanding: The role of orientation features in space and time in scene classification, in: CVPR, 2012, pp. 1306–1313.

[24] G. Zhao, M. Pietikäinen, Dynamic texture recognition using volume local binary patterns, in: Dynamical Vision, ECCV 2006 Workshop, Graz, Austria, May 13, 2006., 2006, pp. 165–177.

[25] G. Zhao, M. Pietikäinen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.

[26] G. Zhao, T. Ahonen, J. Matas, M. Pietikäinen, Rotation-invariant image and video description with local binary pattern features, IEEE Transactions on Image Processing 21 (4) (2012) 1465–1477.

[27] Y. Guo, G. Zhao, M. Pietikäinen, Local configuration features and discriminative learnt features for texture description, in: Local Binary Patterns: New Variants and Applications, 2013, pp. 113–129.

[28] L. Zhang, M. Yang, X. Feng, Sparse representation or collaborative representation: Which helps face recognition?, in: ICCV, 2011, pp. 471–478.

[29] P. Zhu, W. Zuo, L. Zhang, S. C. Shiu, D. Zhang, Image set-based collaborative representation for face recognition, IEEE Transactions on Information Forensics and Security 9 (7) (2014) 1120–1132.

[30] R. Péteri, S. Fazekas, M. J. Huiskes, Dyntex: A comprehensive database of dynamic textures, Pattern Recognition Letters 31 (12) (2010) 1627–1632.

[31] B. Ghanem, N. Ahuja, Maximum margin distance learning for dynamic texture recognition, in: ECCV (2), 2010, pp. 223–236.