Neurocomputing 173 (2016) 1310–1321

Contents lists available at ScienceDirect

Neurocomputing

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

Dynamic texture recognition by aggregating spatial and temporal features via ensemble SVMs

Feng Yang [a,b], Gui-Song Xia [a,b,*], Gang Liu [c], Liangpei Zhang [a,b], Xin Huang [a,b]

[a] State Key Lab. LIESMARS, Wuhan University, Wuhan 430079, China
[b] Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China
[c] CNRS LTCI, Telecom ParisTech, Paris 75013, France

ARTICLE INFO

Article history:
Received 22 April 2015
Received in revised form
8 August 2015
Accepted 1 September 2015
Communicated by X. Li
Available online 10 September 2015

Keywords:
Dynamic textures
Spatial texture features
LDS
Ensemble SVM
Deep scattering transformation
Shape co-occurrence patterns

ABSTRACT

A dynamic texture (DT) refers to a sequence of images that exhibit spatial and temporal regularities. The modeling of DTs plays an important role in many video-related vision tasks, where the main difficulty lies in fact how to simultaneously depict the spatial and temporal aspects of DTs. While unlike the modeling of DTs, tremendous achievements have been recently reported on static texture modeling.

This paper addresses the problem of dynamic texture recognition by aggregating spatial and temporal texture features via an ensemble SVM scheme, and bypassing the difficulties of simultaneously spatio-temporal description of DTs. More precisely, firstly, by considering a 3-dimensional DT video as a stack 2-dimensional static textures, we exploit the spatial texture features of single frame to combine different aspects of spatial structures, followed by randomly selecting several frames of the DT video in the time augmentation process. Secondly, in order to incorporate temporal information, the naive linear dynamic system (LDS) model is used to extract dynamics of DTs in temporal domain. Finally, we aggregate these spatial and temporal cues via an ensemble SVM architecture. We have experimented not only on several common dynamic texture datasets, but also on two challenging dynamic scene datasets. The results show that the proposed scheme achieves the state-of-the-art performances on the recognition of dynamic textures and dynamic scenes. Moreover, our approach offers a simple and general way to aggregate any spatial and temporal features into the task of dynamic texture recognition.

1. Introduction

A dynamic texture (DT) is considered as a sequence of images of moving scenes that exhibit certain stochastic stationary properties, by the first introduction of the concept in [1]. Traditionally, textures are spatially repetitive visual patterns with stochastic mechanism in the microscopic view but presenting periodicity at the macroscopic scale, though it is difficult to make a strict definition of textures. When referring to dynamic or temporal texture, the notion of self-similarity consisting in conventional image texture of 2-dimensional (2D) plane is extended to the spatio-temporal domain [2] which is viewed as 3-dimensional (3D) volume. DTs present spatial and temporal regularity, some type of homogeneity both in space and time [3], and they appear in many natural physical phenomena such as falling rain, flowing rivers, and blowing foliage. Dynamic texture recognition (DTR) amounts to interpreting and understanding what dynamic scene elements displayed in video contents and classifying dynamic textures into meaningful semantic categories. The significance of dynamic textures description and recognition is relevant to a wide range of complex video processing and understanding tasks, such as animation scene modeling, content based video retrieval and anomaly detection in video surveillance.

1.1. Problem statement

The research of dynamic texture primitively stems from physics, where the technique of *particle image velocimetry* (PIV) [4] generates the sequences of visible flow by injecting particles to show the fluid motion that can be considered as dynamic textures. Hereafter, the analysis of dynamic texture is inspired a lot by physics to build the rigid mathematical model for motion patterns. However, the tasks in computer vision and physics are quite different from each other. The goal of dynamic texture recognition is

* Corresponding author at: State Key Lab. LIESMARS, Wuhan University, Wuhan 430079, China.
E-mail addresses: fengyang@whu.edu.cn (F. Yang), guisong.xia@whu.edu.cn (G.-S. Xia), gang.liu@telecom-paristech.fr (G. Liu), zlp62@whu.edu.cn (L. Zhang), xhuang@whu.edu.cn (X. Huang).

http://dx.doi.org/10.1016/j.neucom.2015.09.004
0925-2312/

to assign category labels to the given dynamic texture examples. By far, most of existing investigations have concentrated on simultaneously modeling the spatial and temporal patterns to construct unified spatio-temporal descriptions for DTR. However, underlying physics of temporal or motion patterns are too complicated or very little is understood of them. Roughly speaking, there are three main obstacles in constructing a unified spatio-temporal approach by treating a DT sequence as a 3D volume:

- Difficulties arise when simultaneously modeling the spatial and temporal patterns to form the unified spatio-temporal description, subject to the requirement of rigorous mathematical or physical derivation. Motivated by the motion-based features, optical flow [5–10] computes frame-to-frame motion estimation, but the assumptions of brightness constancy and local smoothness are sometimes undesired for dynamic textures, not mentioning the chaotic dynamics in DT.
- Data-dependent feature extraction limits the generalization to wider dynamic texture classes. Many physics-based spatio-temporal approaches, *e.g.* [11], derive models from the generating process of specific dynamic textures, which leads to data-dependent feature extraction.
- How the spatial appearance and underlying dynamics jointly perform in the recognition is still an open question. Whether much of the recognition performance is highly tied to the spatial appearance or the underlying dynamics remains to be investigated.

Toward this end, the necessity of modeling the integrated spatio-temporal descriptors by simultaneously incorporating spatial and temporal patterns remains questioned in DTR. *Can we aggregate complementary information from separated spatial and temporal features to accomplish the task of DTR?* In other words, we prefer to use a simple scheme which merely needs to combine spatial and temporal features regardless of the complicated mechanism in the interaction of appearance and dynamics.

1.2. Motivation and objective

Primarily, we note the fact that: *Given an example of dynamic textures, one can easily classify it into certain class, even without knowing the dynamics but just with several single frames*. For instance, it is easy to discriminate sea waves, bubbling fountain and flapping foliage by virtue of single image frame, as illustrated in Fig. 1. It can be explained that in such cases the spatial appearance conveys sufficient discriminative information, which contributes more than motion patterns in recognition. It is significant to construct the feature space of research object in recognition tasks. Intuitively, if spatial texture features can be used to well recognize DTs, they should be separable in the feature space. To roughly verify this point, we select 100 successive frames from each DT video sample, then calculate the 2D Gabor features of each frame, plotting each frame as a dot in the first 3 dimensions of the feature space. As shown in Fig. 2, different colors correspond to different DT examples and the color of these dots

Fig. 1. Given examples of dynamic textures, one can easily classify them into certain classes, by just with single frame even without knowing the dynamics.

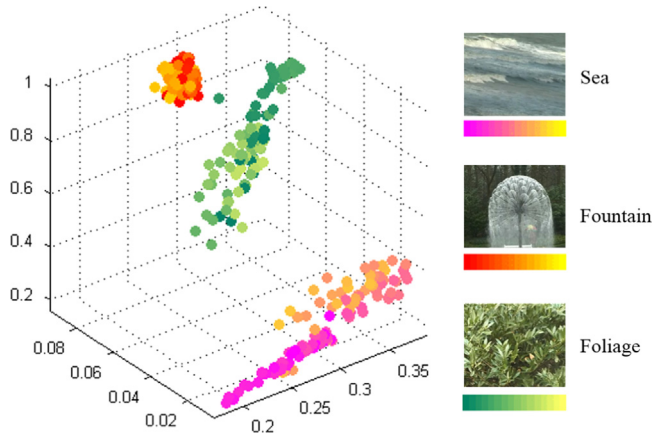


Fig. 2. The first 3 dimensions of spatial Gabor texture feature of DTs, which forms good shape clusters inter-class. (For interpretation of the references to color in this figure, the reader is referred to the web version of this paper.)

varying from dark to shallow indicates different frames in time sequence. As expected, spatial texture features form distinguishable good shape clusters and demonstrate a strong power for recognition.

Then it is natural to quest *how to make use of spatial texture analysis for dynamic texture recognition*, especially, when there are tremendous achievements on spatial texture analysis in the past decades. For example, descriptors such as LBP [12], Gabor [13], Gist [14], ScatNet [15] and SCOP [16,17] are available to handle problems of texture or scene recognition. Since there are so many off-the-shelf powerful algorithms for texture images, considering the difficulties in developing unified spatio-temporal descriptors of DTs, we turn to incorporating these spatial features in our DTR approach. According to [18], motion provides informative cues about the scene. For instance, chaotic traffic and smooth traffic are much similar in their static appearance, where the motion patterns can be utilized to distinguish them between disordered movement and regular flow.

Motivated by these, we intend to integrate static appearance features with motion patterns to form complementary description for dynamic textures. More precisely, we separately process spatial and temporal clues of dynamic textures and aggregate these components in an overall framework. Spatial texture features extracted from dynamic textures frame-by-frame are first combined via an ensemble SVM algorithm to effectively fuse different kinds of individual features, then are incorporated with simple dynamic feature to capture the motion patterns. For augmentation in temporal domain, we randomly select several independent frames during the period of video to pick up temporal alignment. We want to achieve two simple goals:

- First, we intend to propose a simple scheme which is very easy or even naive to adapt to different data, and try to maintain consistency of recognition work between static and dynamic textures.
- Second, such a basic process via aggregation in spatial features and augmentation in time can serve as a good baseline to justify the use of more advanced or more sophisticated spatio-temporal architectures for dynamic texture recognition.

We have also explored an extension of attempt on dynamic scene recognition for further verification. “Dynamic scene” refers to a place where an action or event occurs with time, *i.e.*, scene elements exhibit spatial displacement over time, consisting of objects that motion elements as foreground and non-motion patterns as background (*e.g.* burning fire in forest and rotating

windmill on the farm). Dynamic texture is restricted to a single relatively uniformly structured region, whereas several inner-related regions of different types appear in dynamic scenes. Thus, dynamic scene recognition encounters bigger challenges than that dynamic texture recognition does.

1.3. Related work

In recent years, tremendous investigations in the literature have been devoted to the characterization and recognition of dynamic texture.

A popular trend of research focus on simultaneously modeling the spatial and temporal patterns to construct efficient spatio-temporal descriptions for DTR. Early investigations on dynamic texture modeling include the work of Doretto et al. [19], which jointly exploited the spatial and temporal regularities of dynamic textures by using a linear dynamical system (LDS) model. In the past decade, following this work, many variants have been proposed, see *e.g.* [20–23]. Xia et al. [23] model textures with Gaussian processes for either static or dynamic, which is restricted to space–time stationary textures yet. Such model-based methods [24–28] often explicitly model the statistical generative process and then classify different DTs based on the values of the associated model parameters. The LDS is a statistical generative model which jointly learns the appearance and dynamics of dynamic textures, but the restriction of first-order Markov property and linearity assumption in LDS model makes it powerless to describe complex dynamic textures with cluttered background.

Observing that most of these existing investigations have concentrated on simultaneously modeling the spatial and temporal patterns to construct spatio-temporal descriptions for DTR, a few work studied the dynamic or spatial patterns of dynamic textures separately, except Crivelli et al. [3] who mainly reported the modeling of motion aspects by mixed-state Markov random field. Shroff et al. [18] characterized motion at a global level by using dynamic attributes with chaotic invariants, not requiring localization or tracking of motion elements, but without specific assumptions on the underlying mapping function compared to LDS models. They fused the global spatial GIST feature and the dynamic chaotic invariants into a single feature vector, named by Chaos+GIST, which ignored the different distribution of spatial and temporal cues.

Another type of methods simplifies the modeling of dynamics by equally treating DT in spatial and temporal domain, to extract the features of 2D cross profiles along three orthogonal x-, y- and t-axes in 3D DT volume. For instance, Zhao et al. [29–31] extended local binary pattern (LBP) to the 3D volume by computing LBP of a DT sequence in three orthogonal planes, constructing local binary patterns from three orthogonal planes (LBP-TOP). Wavelet-based multi-fractal spectrum (WMFS) [32] method uses the wavelet-based spatial-frequency analysis in multi-scale pyramids to build a descriptor for both static and dynamic textures. They extended the WMFS descriptor from 2D static texture to 3D dynamic texture by concatenating the WMFS for each 2D slice along x, y and t three axes. However, these methods may bring in redundant and noisy information due to the lack of logical explanation in extracting static texture features for 2D slice along the x- and y-axes.

Based on the spacetime oriented energies (SOE), each dynamic texture pattern is represented as a measurement of histogram that indicates the distribution of a particular set of 3D orientation structures in spacetime, captured by a bank of spatiotemporal filters [33–35]. In the later work, Feichtenhofer et al. [36] aggregated complementary information from separate spatial and temporal orientation measurements in spacetime pyramids via random forest classifier. Inspired by deep learning, the Bags of Spacetime Energies (BoSE) system [37] computes densely extracted local oriented spacetime energies using local linear coding (LLC) [38], subsequently pooled by adaptive dynmax-pooling. While SOE-based approaches have demonstrated promise on dynamic scene recognition, their application to classifying the dynamic textures' patterns has been shown to perform poorly [33].

1.4. Contributions

In this paper, we propose an aggregation-based approach which simplifies the description of DTs, seeking a way for static texture analysis methods to be utilized in DTR. The major contributions of this paper are as follows:

- First, we develop a DTR approach by aggregating spatial and temporal features based on the ensemble SVMs multiclassifier system. In this way, we bypass the difficulties in simultaneously considering spatial appearance and dynamics as unified spatio-temporal description, yet our approach achieves the state-of-the-art recognition performance.
- Second, we investigate the discriminative capacity of features aggregation to accurately capture the semantic categories information of dynamic texture. In other words, just aggregating off-the-shelf spatial texture features provides an alternative perspective to deal with DTR, which could be considered as a valuable baseline for more advanced or more sophisticated spatio-temporal approaches in dynamic texture recognition tasks.

To evaluate the proposed approach, we perform experiments on benchmark dynamic texture datasets and compare with the state-of-the-art methods. There is also an extensional experiment in the dynamic scene datasets to further verify the proposed method.

The remaining parts of the paper are organized as follows: Section 2 introduces the proposed spatial and temporal features aggregation approach, and features being adopted are also mentioned. Section 3 shows an experiment-based evaluation of the proposed approach in contrast with available DTR methods. Finally, conclusions are provided in Section 4.

2. Methodology

Given a set of N DT samples $V=\{V_1,\ldots,V_n,\ldots,V_N\}$, the task of DTR is to assign a class label $c\in\{1,\ldots,C\}$ to each sample V_n, where C is the number of dynamic texture classes. Moreover, each sample V_n is in fact a sequence of L images $V_n=\{I_1^{(n)},\ldots,I_\ell^{(n)},\ldots,I_L^{(n)}\}$. (Observing that different DT samples may contain a different number of frames of images, we let L be the smallest number of frames contained by each sample V_n in the whole V.) Let the ℓ-th frame, $I_\ell^{(n)}$, of the sample V_n be described by a set of K spatial texture features $F_\ell^{(n)}=[F_{\ell,k}^{(n)}]_{k=1}^K$, by

$$\forall 1\le k\le K,\quad F_{\ell,k}^{(n)}=f_k\circ I_\ell^{(n)}, \tag{1}$$

where f_k is a static texture feature extractor, *e.g.* LBP [12], Gabor [13], Gist [14], ScatNet [15] and SCOP [16,17]. Meanwhile, we use a naive dynamic feature $D^{(n)}$, *e.g.* LDS, to depict each sample V_n. Thus, DTR amounts to estimate the class posterior probability $\mathbf{P}(c|V_n)$ for each sample V_n, with respect to the features $\{F_1^{(n)},\ldots,F_\ell^{(n)},\ldots,F_L^{(n)},D^{(n)}\}$.

In this section, we first illustrate the overall framework for aggregating such features for estimating probability $\mathbf{P}(c|V_n)$ for DTR, and then briefly introduce the spatial features and linear dynamical system (LDS) model adopted in the proposed approach.

2.1. Ensemble scheme for aggregating spatial and temporal features

Given the L frames of a DT video V_n, one may analogically solve the problem of dynamic texture recognition by multi-frame static image recognition. Furthermore, due to the self-similarities among image sequences, there is high repeatability in overall slices of 2D images constituting a video displaying textures or scenes, which indicates that several frames may capture sufficient spatial appearance information for representing the whole sample for recognition. Hence, we first propose to use only M random frames of a DT sample V_n, with $1 \le M \le L$, to build a discriminative feature for V_n, by aggregating its spatial texture features $\{F_1^{(n)}, \ldots, F_m^{(n)}, \ldots, F_M^{(n)}\}$ while with regardless of the motion pattern.

2.1.1. Aggregating spatial features via ensemble SVMs

In order to aggregate different types of spatial features, we propose to use an ensemble method. Ensemble methods are learning algorithms that construct a set of classifiers and often used for efficiently combining classifiers [39]. Therefore, ensemble learning is also called committee-based learning or learning multiple classifier systems (MCS) [40]. Early researches in image recognition have shown that combining multiple descriptors is very useful to improve classification performance [41,42]. The naive solution to combination is that different descriptors are combined into a single vector. But a possible problem of creating one large input vector for a machine learning classifier, such as support vector machine (SVM), is that the input vector becomes of very large dimensionality, which may lead to overfitting and hinder generalization performance [43]. Furthermore, it is worth mentioning that, generally, the computational cost of constructing an ensemble is not much larger than creating a single learner [40]. Recently, ensemble methods have been used for efficiently combining classifiers. Support vector machine (SVM) as the so-called base classifier to construct multiple classifier systems (MCS), such a MCS based on SVM, is a very powerful way to combine multiple descriptors in ensemble methods [44]. SVM focuses on optimizing a single processing step, *i.e.*, the fitting of the presumably optimal separating hyperplane [45], while MCS relies on an ideally positive influence of a combined decision derived from several suboptimal yet sometimes computationally simple outputs. Toward this end, it seems desirable to combine SVM and MCS in a complementary approach.

Thus, we first learn the class posterior probability function $\mathbf{P}_{F_k}(c|\cdot)$ with respect to the k-th feature descriptor of the m-th frame of sample V_n by an SVM classifier. As the different SVM classifiers work in different feature spaces, we propose to use product rule to build on the final output decision in the late fusion ensemble architecture, according to [46,44]. Therefore, the output joint class posterior probability $\mathbf{P}(c|I_m^{(n)})$ given the K different spatial features of the m-th frame of sample V_n is

$$\forall c,\ \mathbf{P}(c|I_m^{(n)}) = \mathbf{P}(c|F_m^{(n)}) = \prod_{k=1}^{K} \mathbf{P}_{F_k}(c|F_{m,k}^{(n)}), \tag{2}$$

Then, given that image sequences of the same DT sample show a large extent of similarity, we average these class posterior probabilities of the randomly selected M image frames to get an overall evaluation of the class probability. Thus, the output class posterior probability $\mathbf{P}\left(c|I_1^{(n)}, \ldots, I_m^{(n)} \ldots, I_M^{(n)}\right)$ for the M selected frames of the DT sample V_n is estimated as

$$\forall c,\ \mathbf{P}\left(c|I_1^{(n)}, \ldots, I_m^{(n)} \ldots, I_M^{(n)}\right) = \frac{1}{M}\sum_{m=1}^{M} \mathbf{P}_I(c|I_m^{(n)}) \tag{3}$$

$$\forall c,\ \mathbf{P}\left(c|I_1^{(n)}, \ldots, I_m^{(n)} \ldots, I_M^{(n)}\right) = \frac{1}{M}\sum_{m=1}^{M}\prod_{k=1}^{K} \mathbf{P}_{F_k}(c|F_{m,k}^{(n)}). \tag{4}$$

where $\mathbf{P}_{F_k}(c|\cdot)$'s are learned by SVM classifiers.

2.1.2. Aggregating dynamic features

In addition to the spatial appearance, the naive linear dynamical system (LDS) model is also employed to extract temporal features in DTs. Herein we also learn the class probability function $\mathbf{P}_D(c|\cdot)$ over the LDS dynamic feature D with an SVM classifier.

In order to aggregate the complementary spatial and temporal information, we assume that the spatial and temporal features are independent and multiply the two terms:

$$\forall c,\ \mathbf{P}(c|V_n) = \mathbf{P}\left(c|I_1^{(n)}, \ldots, I_m^{(n)} \ldots, I_M^{(n)}\right) \cdot \mathbf{P}_D\left(c|D^{(n)}\right), \tag{5}$$

$$\forall c,\ \mathbf{P}(c|V_n) = \frac{1}{M}\left(\sum_{m=1}^{M}\prod_{k=1}^{K} \mathbf{P}_{F_k}(c|F_{m,k}^{(n)})\right) \cdot \mathbf{P}_D\left(c|D^{(n)}\right). \tag{6}$$

Then the final predicted class label of the input DT sample V_n is determined by the majority vote with the largest probability in the C classes:

$$\hat{c}(V_n) = \arg\max_{c\in\{1,\ldots,C\}}\left(\sum_{m=1}^{M}\prod_{k=1}^{K} \mathbf{P}_{F_k}(c|F_{m,k}^{(n)})\right) \cdot \mathbf{P}_D\left(c|D^{(n)}\right). \tag{7}$$

The overall flowchart of our approach is illustrated in Fig. 3.

2.2. Spatial features

A good feature for recognition should get the most efficient characteristics that preserve the intra-class invariance while capture the inter-class discriminative information of images. Although, so far, there is no single feature that can produce a universal solution for all images, a variety of features provide different aspects of discriminative information in images and may help to depict different structures of images. Combining multiple features that focus on extracting different types of patterns can make up complementary visual information of semantic description. Nevertheless, not all the arbitrary aggregation of features makes sense. In order to get a wise combination, the ensemble methods should comply with some criteria such as diversity, independence, decentralization and aggregation [47]. Based on these ensemble criteria, we select several kinds of features from different aspects of discriminative information description. Conventional efficient static texture features such as LBP [12] and Gabor [48] are used for simple and regular texture patterns extraction. For describing geometrical and high-order static texture information, shape-based texture descriptor SCOPs [17] and deep network-based feature ScatNet [15] are selected for complex or cluttered textures. For depicting scene-level information, we also select GIST descriptor as a holistic image feature. Moreover, in order to make use of chromatic information both for texture and scene, we propose to utilize the discriminative color descriptor [49].

2.2.1. Local Binary Patterns [12]

The basic Local Binary Patterns (LBP) [12] operator is a gray-scale invariant texture primitive statistic based on the measurement of local image contrasts. It has obtained good performance in the classification of various kinds of textures. For each pixel in an image, a binary code is calculated by thresholding its neighborhood with the value of the center pixel. LBP computes the joint distribution of the gray levels of $P(P>1)$ pixels in a local

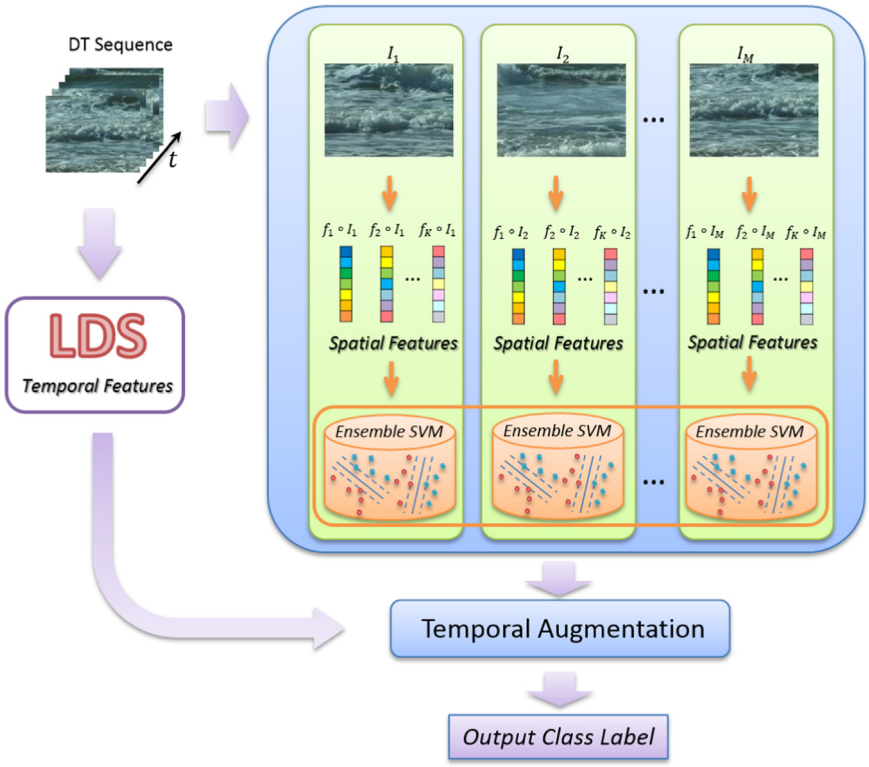


Fig. 3. The proposed dynamic texture recognition method by aggregating spatial and temporal features.

neighborhood of texture image:

$$LBP_{P,R} = \sum_{p=0}^{P-1} \text{sign}(g_p - g_c)2^p \quad \text{with } \text{sign}(g) = \begin{cases} 1, & g \geq 0 \\ 0, & g < 0 \end{cases} \tag{8}$$

where g_c corresponds to the gray value of the center pixel and g_p with $(p = 0, \ldots, P-1)$ corresponds to the gray values of its P local neighborhood pixels equally spaced on a circle of radius $R, (R > 0)$ that form a circularly symmetric neighbor set. The sign(x) is a sign function. Finally, a histogram is created to collect up the occurrences of different binary patterns.

2.2.2. Gabor filter responses

Based on the important discovery by Hubel and Wiesel in the early 1960s [50], the neurons of the primary visual cortex respond to lines or edges of a certain orientation in different positions of the visual field. The simple cells of the visual cortex of mammalian brains are best modeled as a family of self-similar 2D Gabor filters. As a local band-pass filter with the conjoint space–frequency domain [48] for image analysis, 2D Gabor filters have both the multi-resolution and multi-orientation properties to measure texture features. Generally speaking, one texture image is convolved with a set of Gabor filters of different preferred orientations and spatial frequencies, resulting in filter responses to form a feature vector field that is for further applications [13].

A family of Gabor functions [13,51] can be defined as a product of two terms:

$$\forall (x,y) \in \Omega, \quad g_{\lambda,\theta,\varphi}(x,y) = e^{-(x'^2+\gamma^2 y'^2)/\sigma^2} \cdot \cos\left(2\pi \frac{x'}{\lambda} + \varphi\right) \tag{9}$$

where $x' = x\cos\theta + y\sin\theta, y' = -x\sin\theta + y\cos\theta$. The first term is a Gaussian envelope function that restricts the filter in the spatial domain. A cosine carrier function varies the frequency of the modulating wave and the angle parameter $\theta \in [0, 2\pi)$ describes the spatial orientation of the filter.

2.2.3. GIST scene descriptors [14]

GIST is a computational model of the recognition of real world scenes that bypasses the segmentation and the processing of individual objects or regions [14]. As a holistic image feature of scene recognition, GIST is based on a very low dimensional representation of the scene, which can be reliably estimated using spectral and coarsely localized information via frequency analysis. Like Gabor filters, scene descriptors have also proven very useful for texture analysis [52], so features of this kind are also included here.

2.2.4. Shape-based Co-occurrence Patterns [17]

Shape Co-occurrence Patterns (SCOPs) [17] proposed a flexible shape-based texture analysis framework by investigating the co-occurrence patterns of shapes. It follows the work of the shape-based invariant texture analysis (SITA) method [16], which relies on morphological operations to obtain a tree of explicit shapes. More precisely, a texture is decomposed into a tree of shapes (the topographical map) relying on a fast level set transformation (FLST) [53], where each shape is associated with some attributes. The shape-based elements of the topographical map provide a local representation of the image, with which we can analyze the local textons of images. SITA [16] collects the shapes and describes the texture by a histogram of individual attributes. In [17], a set of co-occurrence patterns of shapes is learned from texture images by clustering algorithm (*e.g. K*-means). Taking the learnt co-occurrence patterns of shapes as visual words, a bag-of-words model is finally established to describe a texture. SCOP demonstrated superior performance both on the multiple texture dataset and the complex scene dataset.

2.2.5. Deep network-based features [15]

Motivated by the deep neural networks (DNNs), Bruna and Mallat [15] proposed a prefixed cascaded wavelet transform convolutions with non-linear modulus and averaging operators, which is called wavelet scattering networks (ScatNet). ScatNet

constructed a cascade of invariants about rotation, translation and scaling, which have demonstrated superior performance on static texture recognition. ScatNet is implemented by a deep convolution network with wavelets filters and modulus non-linearities [54,55].

Given an image, locally invariant translation and rotation coefficients are first computed by averaging the image x with a rotation invariant low pass filter. The high frequencies covariant to the action of a group $G \in \mathbb{R}^2$ are recovered by convolution with high pass wavelet filters ψ, followed by a modulus operator to make it more insensitive to translations.

2.2.6. Chromatic information [49]

Chromatic information also contains rich information for texture or scene recognition [56]. Discriminative color descriptors [49] partition color values into clusters based on their discriminative power in a classification problem while preserving the photometric invariance. By taking an information theoretic approach named as Divisive Information-Theoretic Clustering (DITC) algorithm [57], the clustering of color description has the objective to minimize the decrease of mutual information of the final representation. A universal color vocabulary to represent the real-world is built by joining several objects and scene training sets together. As a consequence of universality, there is no need to learn a new color representation for every new dataset and one can just apply the universal color representation to our problem. We follow the standard bag-of-words pipeline to construct color description by using the Fisher vectors (FV) coding method [58]. Finally, to represent an image we form a color descriptor of length of 500 dimensions by means of PCA dimension reduction after FV coding.

2.3. Temporal features

In addition to the spatial appearance, the linear dynamical system (LDS) model is also employed to extract temporal features in DTs, which depicts dynamic systems as second-order stationary stochastic processes [19]. LDS is a statistical generative model that captures the input and output of dynamical system by a set of model parameters. The model can be written as

$$x(t+1) = Ax(t) + w(t), \quad w(t) \sim \mathcal{N}(0, R), \tag{10}$$

$$I(t) = Cx(t) + v(t), \quad v(t) \sim \mathcal{N}(0, Q), \tag{11}$$

where $x(t)$ is the hidden state vector, $I(t)$ is the observation vector, *i.e.*, the image frame at each instant of time t; $w(t)$ and $v(t)$ are independent and identically distributed (IID) noise components with normal distribution of zero mean and covariance matrix R and Q respectively. The parameter A is the state-transition matrix while C models the observation matrix, learned from the input and output of the dynamical system. The model parameters can be estimated by using the singular value decomposition (SVD) of the data matrix. The LDS features lie in a non-Euclidean space so that subspace angles are usually used as a similarity metric [59].

Given two DTs, modeled with $M_i = (A_i, C_i)$ and $M_j = (A_j, C_j)$, the pairwise element k_{ij} of the similarity matrix K is calculated as

$$k_{ij} = e^{-d_M^2(M_i, M_j)} \tag{12}$$

where $d_M(\cdot,\cdot)$ is the Martin distance [60] for subspace angles. This similarity matrix can be used for the SVM classifier with pre-computed kernel mode.

3. Experiments and discussions

In this section, we conduct a series of experiments to verify the performance of the proposed approach by aggregating spatial and temporal features. We use two types of databases related to dynamic texture and dynamic scene. In contrary to the variety of established image datasets, there is currently lack of video classification benchmarks because videos are significantly more difficult to collect, annotate and store [61]. With regard to the specific kind of videos, constructing a comprehensive database of dynamic scene is not an easy task, especially for dynamic textures. In this paper, for dynamic scene recognition, we evaluate the proposed approach on the benchmark datasets: Maryland "In-The-Wild" [18] dataset and the YUPENN Dynamic Scenes [34] dataset. UCLA [1] and DynTex [62] from dynamic texture community are surely two main dynamic texture databases that are used by almost every recognition method in the state-of-the-art literatures.

For a fair comparison and being consistent with previous studies [20,21,29,32,34,36,18], the same experimental setup, leave-one-out classification procedure, was employed in all our experiments and the average classification accuracy was used for evaluations. We utilized the LIBSVM [63] in our experiments, and the SVM scores are directly used as estimations of the class posterior probabilities. More precisely, in the LIBSVM option mode, "-*b*" is set up to do probability estimation. SVM with RBF kernel is used by the cross-validation to determine the optimal parameters of C and g. With regard to the augmentation in time, we randomly select 1, 5, 10, 15 and 20 frames of images from a given DT video sample to extract features for each video instance, then use the ensemble SVM framework to aggregate features for each single frame as described in Section 2, and these selected frames are averaged for each DT video. The complementary temporal information is combined with spatial features finally to determine the output class label of DTs. We have tested the influence of randomness in choice of frames on the final results, which demonstrated that the randomness affected the recognition rates at an acceptable level, thus we made no further analysis about this in the experiments.

3.1. Experiments on dynamic texture datasets

3.1.1. Results on UCLA-50 dataset

The UCLA dynamic texture dataset was first introduced to test the recognition performance of LDS-based methods in [1] and then widely used for evaluating dynamic texture analysis. It includes 50 dynamic texture categories, called UCLA-50, each class with 4 grayscale instances at a frame rate of 15 fps. The original sequences are carefully cropped as a spatiotemporal volume of size of $48 \times 48 \times 75$, *i.e.*, 75 frames with 48×48 pixels. The 50 classes are formed by artificially separating semantically equivalent classes of video shots at different viewpoints or scales into different categories. Fig. 4 shows several sample frames of DTs from the UCLA dataset.

The recognition rates of different individual spatial texture features are shown in Fig. 5. The absence of color descriptor is due to the fact that the DTs are in grayscale. One can observe that when the number of randomly selected frames is up to 5, the recognition rate of using a single spatial feature almost comes up to 99.0%, which indicates that just several frames can capture enough discriminative information. To aggregate spatial and temporal features, we choose the spatial LBP and SCOP together with LDS dynamics, which make up the best combination in ensemble SVMs and display superiority to individual feature. To further specify the ensemble results, we demonstrate them in Table 1, where the performances of using individual LBP, SCOP and LDS are also presented. Notice that the process of LDS is performed on the whole DT volume so that it has nothing to do with the number of frames and the same accuracy value is shown about LDS. The highest recognition rate achieved is 100.00% when using 20 frames, which is better than the state-of-the-art result 99.75% [32].

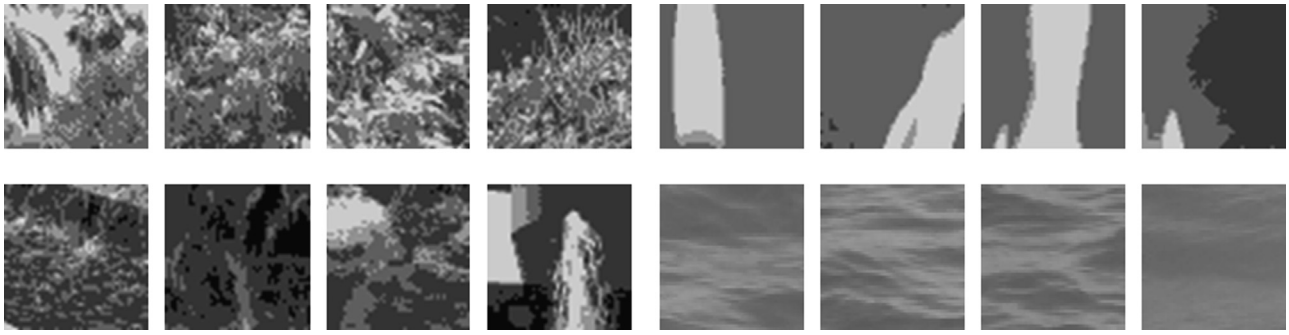


Fig. 4. Samples of video sequences from the UCLA dynamic texture database, including windblown plants, fountains, fire and rippling water (from up to down and left to right).

In comparison with the previous studies on this UCLA dataset, Table 2 presents the classification rates of the state-of-the-art methods reported in the previous works.

The *spatiotemporal oriented energies* (SOE) [33] achieved a classification accuracy of 81.0% on the UCLA dynamic texture datasets, based on matching histograms distribution of spacetime orientation structure. The reported best recognition performance based on LDS model is 97.50% by using the kernel dynamic texture LDS system (KDT-LDS) [20], while our method shows stronger discriminative power by integrating spatial features into the LDS. The maximum margin distance learning (MMDL) [21] learned the weights to different spatiotemporal dimensions and used a linear combination of three elementary distances representing DT space where the dimension of dynamics was also depicted by LDS, which got a classification rate of 99.0%. Wavelet-based multi-fractal spectrum (WMFS) [32] used the wavelet-based spatial frequency analysis in multi-scale pyramids to build a descriptor for both static and dynamic textures, their reported test on UCLA datasets reached a recognition rate of 99.75%. In our experiments, we directly compare our method with the results reported in their papers, as using the same experimental protocols.

Table 1
Ensemble results of the combined LBP, SCOP, LDS features and their individual recognition results on the UCLA dataset.

Number of frames	LBP	SCOP	LDS	Ensemble SVMs
1	72.96	35.42	90.70	77.32
5	96.76	97.54	90.70	99.00
10	98.00	99.10	90.70	99.46
15	**100.00**	99.38	90.70	99.90
20	99.74	**100.00**	90.70	**100.00**

Table 2
Comparison of performances for different methods on the UCLA dataset.

Methods	Rate (%)
SOE [33]	81.00
KDT-LDS [20]	97.50
MMDL [21]	99.00
WMFS [32]	99.75
Ours	**100.0**

3.1.2. Results on DynTex dataset

Another widely used dataset is the DynTex dynamic texture database [62]. It is a diverse collection of high-quality and colored dynamic texture videos where more than 650 sequences are available. The video sequences have a spatial size of 352×288 and consist of at least 250 frames with 25 fps. In order to provide DynTex for use of recognition, different sub-datasets with manual annotation for each DT have been compiled and labeled. One of the largest sub-datasets is the Gamma dataset which is composed of 264 dynamic textures divided into 10 classes: *Flowers(29), Sea (38), Naked trees(25), Foliage(35), Escalator(7), Calm water(30), Flags(31), Grass(23), Traffic(9), and Fountains(37)*. Here the number in parenthesis represents the amount of DT samples in each class. Examples of DynTex are shown in Fig. 6. The content of videos in the dataset contains not only dynamic texture regions but also surrounding background which makes the DTs less homogeneous. In addition, the amount of samples in different classes has number bias. Several classes have few samples, *e.g.* 7–9 video sequences in the classes of *Escalator* and *Traffic*, but the number of samples in the largest class reaches 38. These issues lead to challenges in adopting DynTex for testing.

Due to the disturbance and ambiguity of DynTex database, there are multiple subversions of this mother database and it has been reorganized into different smaller and customized datasets. The 10-class Gamma dataset of DynTex, named DynTex-10 for short, is the largest one with a diversity of classes and samples, whose annotations are specified by the data provider. To the best of our knowledge, DynTex-10 has not been tested by any DT analysis method so far.

Different reorganizations of the original DynTex dataset by previous methods prevent a direct quantitative comparison between our method and the state-of-the-arts. The classification accuracy of a 3-class DynTex-3 by joint segmentation and categorization proposed by Ravichandran et al. [64] is 72.5%. Motion textures represented by MRF model [3] provided an overall classification rate of 90.7% while testing on a 10-class subset of DynTex with only 3 samples in each class. Zhao et al. [29] also tested LBP-TOP on a subset of DynTex, which includes only 4 classes and each class is derived from a DT by dividing it into 10 sub-sequences. To some extent, these manipulations on data decreased the difficulty of recognition and they obtained an accuracy of 97.14%.

In this paper, we propose to test our approach on the whole DynTex, and the recognition results of different individual spatial texture features including color descriptor are shown in Fig. 7. For Dyntex, we combine LBP, SCOP and color features together with LDS dynamics in ensemble SVM. As presented in Table 3, LBP, SCOP and color achieved the recognition rate 77.50%, 90.30% and 78.40% respectively for 1 frame, as well as an accuracy of 86.80% for LDS, while the ensemble SVMs can reach 99.50%. As reported by [65], 3D Gabor filters with low speeds achieved better results than high speeds for the DynTex database, giving the evidence that the DynTex is indeed composed of DTs with low motion patterns. From the perspective about appearance of the DT, low motion patterns mean great similarities between image frames. In agreement with this fact, our experimental result also shows that there

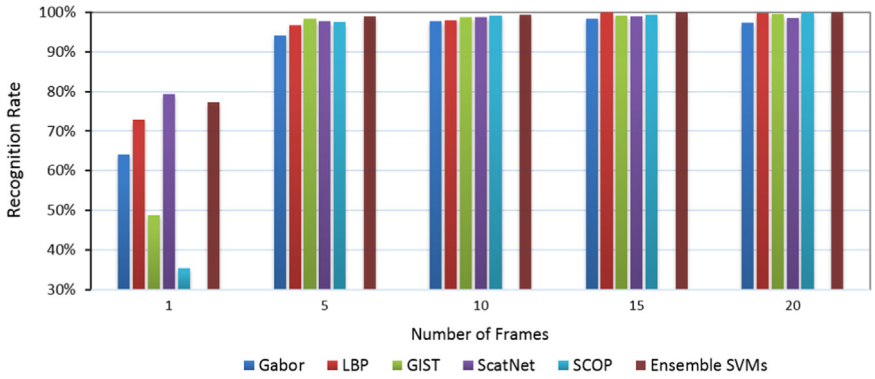


Fig. 5. The recognition rates against number of frames on the UCLA dynamic texture dataset. *Ensemble SVMs* represent the result of combined LBP, SCOP and LDS features.

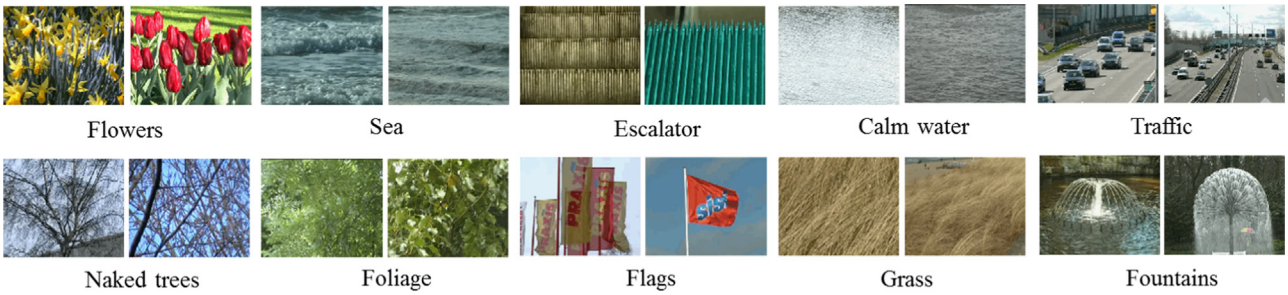


Fig. 6. Examples of video sequences from the DynTex dynamic texture database.

is no promotion accompanied by the increase of number of frames, as shown in Fig. 7 and Table 3. Another explanation is that dynamic texture in DynTex exhibit huge intra-class variations in the background and only a small amount of common foreground area, *i.e.*, the dynamic texture itself.

3.2. Experiments on dynamic scenes

3.2.1. Dynamic scene datasets

- *Maryland "In-The-Wild" scene dataset*: It consists of 13 classes of dynamic scenes with 10 color videos per class. The average size of these videos is $308 \times 417 \times 617$, *i.e.*, average spatial size of 308×417 pixels with a length about 617 frames. The videos were collected from Internet websites, *e.g.* YouTube, which have large variations in illuminations, frame rates, viewpoints, scales and resolutions, as well as various degrees of camera-induced motion (*e.g.* panning and jitter), taking large intra-class variations to this dataset. Classes included in this dataset are *Avalanche, Boiling Water, Chaotic Traffic, Forest Fire, Fountain, Iceberg Collapse, Landslide, Smooth Traffic, Tornado, Volcanic Eruption, Waterfall, Waves and Whirlpool*. Fig. 8 displays several sample frames from the Maryland dataset.
- *YUPENN dataset*: It contains 14 dynamic scene classes with 30 color videos for each class. Compared to the Maryland dataset, this dataset emphasizes scene specific temporal information over short time durations caused by dynamics of objects and surfaces without camera-induced motion [34]. Thus it is more concerned about the task of scene recognition. The average size of the dataset is $250 \times 370 \times 145$. The videos derive from a variety of sources such as websites and shoots by the suppliers. These videos also contain variations in image resolutions, frame rates, scene appearances, scales, illumination conditions and camera viewpoints, nonetheless most of them are obtained from a stationary camera. Fig. 9 shows the sample frames from the YUPENN dataset.

Table 3
Recognition performances achieved by using individual and ensemble results of the LBP, SCOP, LDS features on the DynTex-10 dataset.

Number of frames	LBP	SCOP	Color	LDS	Ensemble SVMs
1	77.50	90.30	78.40	86.80	99.50
5	84.80	90.50	80.40	86.80	98.00
10	85.50	92.30	80.90	86.80	97.80
15	81.00	90.70	80.90	86.80	97.20

3.2.2. Results on dynamic scene recognition

Based on our experiments, the optimized feature sets are GIST, SCOP and Color together with the LDS for both the dynamic scene datasets, where we use a scene descriptor GIST for scene recognition instead of LBP texture feature used in dynamic textures. Figs. 10 and 11 show that both the dynamic scene datasets achieve their best recognition results with 15 frames selected, which are merely small part of the frames in the whole video.

We compare our approach to several previous methods that have reported excellent performance: GIST with chaotic dynamic features (Chaos) [18], spatiotemporal oriented energies (SOE) [34], complementary spacetime orientation (CSO) features [36], and the Bags of Spacetime Energies (BoSE) system [37]. These studies almost cover the recently remarkable development on the topic of dynamic scene recognition. As the comparison presented in Table 4, for both datasets, the proposed approach outperforms the previous state-of-the-art methods. Here, our approach obtains an average accuracy of 78.77% and 96.43% for Maryland and YUPENN respectively, which makes a further improvement better than BoSE method [37] that employed trivially detailed dynmax-pooling strategy and got good accuracies both on the two datasets. More than the accuracy, allowing for the simplicity for video processing, our approach also demonstrates superior performance on the computation complexity. Chaos+GIST [18] concatenated GIST feature and the dynamic chaotic invariants into a single feature vector, and fed them into a classifier, which ignored the different distribution of spatial and temporal cues. CSO [36]

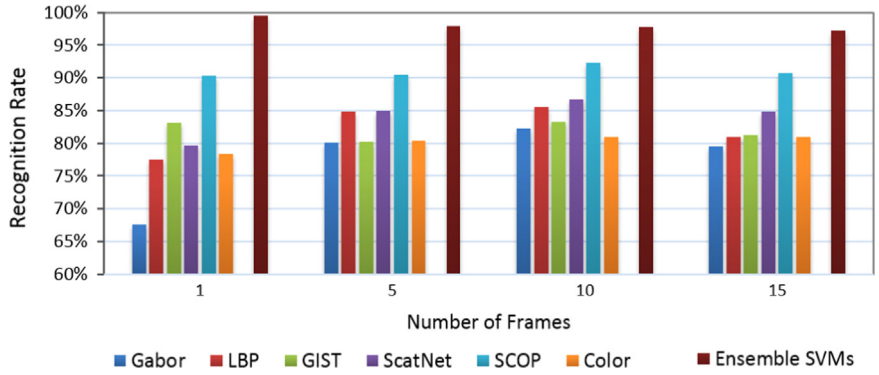


Fig. 7. The recognition rate against number of frames on the DynTex dynamic scenes dataset. *Ensemble SVMs* represent the result of combined LBP, SCOP, Color and LDS features.

Fig. 8. Maryland "in-the-wild" dynamic scenes dataset.

Fig. 9. Sample frames of YUPENN dynamic scenes dataset.

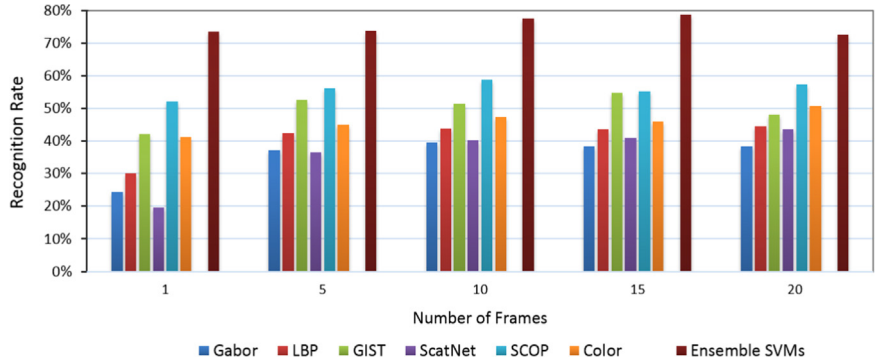


Fig. 10. The recognition rates against number of frames on the Maryland dynamic scenes dataset. *Ensemble SVMs* represent the result of combined GIST, SCOP, Color and LDS features.

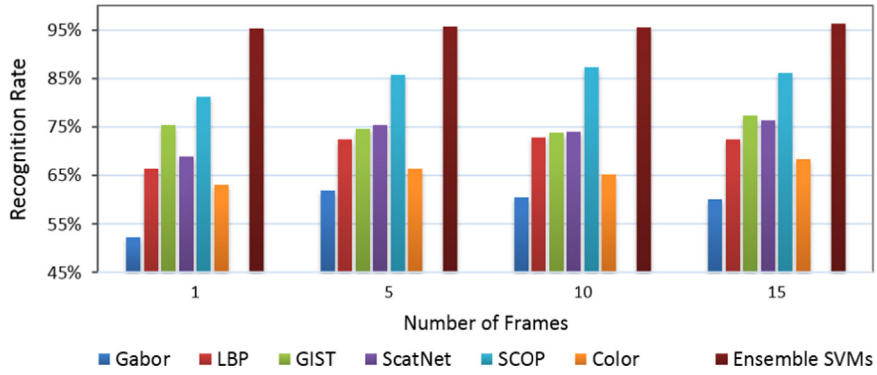


Fig. 11. The recognition rates against number of frames on the YUPENN dynamic scenes dataset. *Ensemble SVMs* represent the result of combined GIST, SCOP, Color and LDS features.

Table 4
Comparison of performances for different methods on the Maryland and YUPENN dataset.

Methods	Maryland	YUPENN
Chaos+GIST [18]	58.46	22.86
SOE [34]	43.08	80.71
CSO [36]	67.69	85.95
BoSE [37]	77.69	96.19
Ours	78.77	96.43

method used random forest classifiers to combine spatial and temporal orientation measurements in spacetime pyramids, and obtained acceptable recognition rates of 67.69% and 85.95% on Maryland and YUPENN respectively. But just extracting the oriented spacetime structures limited the generalization ability to wider unconstraint dynamic scenes. Our proposed approach increases the diversity of spatial and temporal features, leading to a more flexibility in the aggregation architecture, aggregating various types of features with better complementary information to achieve the state-of-the-art performance. Moreover, the proposed approach is more likely to work stably in the presence of camera motion that frequently occurs in Maryland dataset, because the SCOP feature is relied on the topographic map of spatial appearance.

4. Conclusion

In this paper, we have investigated the ultimate recognition capacity of aggregating spatial and temporal features in dynamic texture recognition and described a complementary appearance-and-motion based method, leading to performance surpassing the state-of-the-art results. The framework of the proposed method offers a simple way to aggregate any static image analysis method into the realization of dynamic texture recognition, readily applicable from the spatial domain to the spatiotemporal case. Such simplicity provides an alternative and yet innovative perspective to deal with DTR, which could be viewed as a valuable baseline for studying more advanced appearance and motion models or spatio-temporal descriptors for recognition tasks. The extensional experiments on the dynamic scene datasets further verify the efficiency of the proposed method. Since dynamic texture recognition is a developing and potential research field, where a lot of work is still to be done, it is of great interest to investigate more effective schemes for DTR by aggregating spatial and temporal features. Other aspects include studying the problem how the spatial appearance and underlying dynamics jointly perform in dynamic texture recognition, investigating the use of the these dynamic information in the visual tracking system [66–69].

Acknowledgments

This research is supported by the National Natural Science Foundation of China under the Contract nos. 91338113 and 41501462, the Chenguang Program of Wuhan Science and Technology under the Contract no. 2015070404010182, and the Open Projects Program of National Laboratory of Pattern Recognition (201306301).

References

[1] P. Saisan, G. Doretto, Y.N. Wu, S. Soatto, Dynamic texture recognition, in: Proceedings of Computer Vision and Pattern Recognition, vol. 2, 2001, p. II-58.
[2] D. Chetverikov, R. Péteri, A brief survey of dynamic texture description and recognition, in: Computer Recognition Systems, 2005, Springer, Wrocław, Poland, pp. 17–26.
[3] T. Crivelli, B. Cernuschi-Frias, P. Bouthemy, J. Yao, Motion textures: modeling, classification, and segmentation using mixed-state Markov random fields, SIAM J. Imaging Sci. 6 (4) (2013) 2484–2520.
[4] I. Grant, Particle image velocimetry: a review, Proc. Inst. Mech. Eng. 211 (1) (1997) 55–76.
[5] R.C. Nelson, R. Polana, Qualitative recognition of motion using temporal texture, CVGIP Image Underst. 56 (1) (1992) 78–89.
[6] P. Bouthemy, R. Fablet, Motion characterization from temporal cooccurrences of local motion-based measures for video indexing, in: Proceedings of International Conference on Pattern Recognition, vol. 1, 1998, pp. 905–908.
[7] R. Péteri, D. Chetverikov, Dynamic texture recognition using normal flow and texture regularity, Pattern Recognition and Image Analysis, Springer, Estoril, Portugal (2005), p. 223–230.
[8] R. Vidal, A. Ravichandran, Optical flow estimation & segmentation of multiple moving dynamic textures, in: Proceedings of Computer Vision and Pattern Recognition, vol. 2, 2005, pp. 516–521.
[9] S. Fazekas, D. Chetverikov, Analysis and performance evaluation of optical flow features for dynamic texture recognition, Signal Proces.: Image Commun. 22 (7) (2007) 680–691.
[10] R. Fablet, P. Bouthemy, Motion recognition using nonparametric image motion models estimated from temporal and multiscale co-occurrence statistics, IEEE Trans. Pattern Anal. Mach. Intell. 25 (12) (2003) 1619–1624.
[11] T. Kung, W. Richards, Inferring water from images, Nat. Comput. (1988) 224–233.
[12] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[13] S.E. Grigorescu, N. Petkov, P. Kruizinga, Comparison of texture features based on Gabor filters, IEEE Trans. Image Process. 11 (10) (2002) 1160–1167.
[14] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (3) (2001) 145–175.
[15] J. Bruna, S. Mallat, Invariant scattering convolution networks, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1872–1886.
[16] G.-S. Xia, J. Delon, Y. Gousseau, Shape-based invariant texture indexing, Int. J. Comput. Vis. 88 (3) (2010) 382–403.

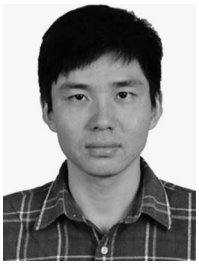
[17] G. Liu, G.-S. Xia, W. Yang, L. Zhang, Texture analysis with shape co-occurrence patterns, in: Proceedings of International Conference on Pattern Recognition, 2014, pp. 1627–1632.
[18] N. Shroff, P. Turaga, R. Chellappa, Moving vistas: exploiting motion for describing scenes, in: Proceedings of Computer Vision and Pattern Recognition, 2010, pp. 1911–1918.
[19] G. Doretto, A. Chiuso, Y.N. Wu, S. Soatto, Dynamic textures, Int. J. Comput. Vis. 51 (2) (2003) 91–109.
[20] A.B. Chan, N. Vasconcelos, Classifying video with kernel dynamic textures, in: Proceedings of Computer Vision and Pattern Recognition, 2007, pp. 1–6.
[21] B. Ghanem, N. Ahuja, Maximum margin distance learning for dynamic texture recognition, in: Proceedings of European Conference on Computer Vision, Springer, Crete, Greece, 2010, pp. 223–236.
[22] G.-S. Xia, S. Ferradans, G. Peyré, J.-F. Aujol, Compact representations of stationary dynamic textures, in: Proceedings of International Conference Image Processing, 2012, pp. 2993–2996.
[23] G.-S. Xia, S. Ferradans, G. Peyré, J.-F. Aujol, Synthesizing and mixing stationary gaussian texture models, SIAM J. Imaging Sci. 7 (1) (2014) 476–508.
[24] B. Abraham, O.I. Camps, M. Sznaier, Dynamic texture with fourier descriptors, in: Proceedings of the Fourth International Workshop on Texture Analysis and Synthesis, 2005, pp. 53–58.
[25] B.U. Toreyin, A.E. Cetin, Hmm based method for dynamic texture detection, in: Signal Processing and Communications Applications, 2007, pp. 1–5.
[26] A.B. Chan, N. Vasconcelos, Mixtures of dynamic textures, in: Proceedings of International Conference on Computer Vision, vol. 1, 2005, pp. 641–647.
[27] B. Ghanem, N. Ahuja, Phase based modelling of dynamic textures, in: Proceedings of International Conference on Computer Vision, 2007, pp. 1–8.
[28] M. Szummer, R.W. Picard, Temporal texture modeling, in: Proceedings of International Conference on Image Processing, vol. 3, 1996, pp. 823–826.
[29] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.
[30] G. Zhao, M. Pietikäinen, Dynamic texture recognition using volume local binary patterns, in: Dynamical Vision, Springer, Graz, Austria, 2007, pp. 165–177.
[31] G. Zhao, T. Ahonen, J. Matas, M. Pietikainen, Rotation-invariant image and video description with local binary pattern features, IEEE Trans. Image Process. 21 (4) (2012) 1465–1477.
[32] H. Ji, X. Yang, H. Ling, Y. Xu, Wavelet domain multifractal analysis for static and dynamic texture classification, IEEE Trans. Image Process. 22 (1) (2013) 286–299.
[33] K.G. Derpanis, R.P. Wildes, Dynamic texture recognition based on distributions of spacetime oriented structure, in: Proceedings of Computer Vision and Pattern Recognition, 2010, pp. 191–198.
[34] 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: Proceedings of Computer Vision and Pattern Recognition, 2012, pp. 1306–1313.
[35] 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.
[36] C. Feichtenhofer, A. Pinz, R.P. Wildes, Spacetime forests with complementary features for dynamic scene recognition, in: Proceedings of British Machine Vision Conference, 2013.
[37] C. Feichtenhofer, A.Pinz, R.P. Wildes, Bags of spacetime energies for dynamic scene recognition, in: Proceedings of Computer Vision and Pattern Recognition, 2014, pp. 2681–2688.
[38] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: Proceedings of Computer Vision and Pattern Recognition, 2010, pp. 3360–3367.
[39] T.G. Dietterich, Ensemble methods in machine learning, in: Multiple Classifier Systems, Springer, Cagliari, Italy, 2000, pp. 1–15.
[40] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, CRC Press, Taylor & Francis, Beijing, China, 2012.
[41] K. Grauman, T. Darrell, The pyramid match kernel: discriminative classification with sets of image features, in: Proceedings of International Conference on Computer Vision, vol. 2, 2005, pp. 1458–1465.
[42] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2169–2178.
[43] A. Abdullah, R.C. Veltkamp, M.A. Wiering, Spatial pyramids and two-layer stacking svm classifiers for image categorization: a comparative study, in: IJCNN, 2009, pp. 5–12.
[44] A. Abdullah, R.C. Veltkamp, M.A. Wiering, An ensemble of deep support vector machines for image categorization, in: Proceedings of Soft Computing and Pattern Recognition, pp. 301–306.
[45] H.-C. Kim, S. Pang, H.-M. Je, D. Kim, S. Yang Bang, Constructing support vector machine ensemble, Pattern Recognit. 36 (12) (2003) 2757–2767.
[46] L.A. Alexandre, A.C. Campilho, M. Kamel, On combining classifiers using sum and product rules, Pattern Recognit. Lett. 22 (12) (2001) 1283–1289.
[47] L. Rokach, Ensemble-based classifiers, Artif. Intell. Rev. 33 (1–2) (2010) 1–39.
[48] J.G. Daugman, Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, J. Opt. Soc. Am. A 2 (7) (1985) 1160–1169.
[49] R. Khan, J. Van de Weijer, F.S. Khan, D. Muselet, C. Ducottet, C. Barat, Discriminative color descriptors, in: Proceedings of Computer Vision and Pattern Recognition, 2013, pp. 2866–2873.
[50] D.H. Hubel, T.N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, J. Physiol. 160 (1) (1962) 106.
[51] N. Petkov, Biologically motivated computationally intensive approaches to image pattern recognition, Future Gener. Comput. Syst. 11 (4) (1995) 451–465.
[52] M. Crosier, L.D Griffin, Using basic image features for texture classification, Int. J. Comput. Vis. 88 (3) (2010) 447–460.
[53] P. Monasse, F. Guichard, Fast computation of a contrast-invariant image representation, IEEE Trans. Image Process. 9 (5) (2000) 860–872.
[54] S. Mallat, Group invariant scattering, Commun. Pure Appl. Math. 65 (10) (2012) 1331–1398.
[55] L. Sifre, S. Mallat, Rotation, scaling and deformation invariant scattering for texture discrimination, in: Proc. of Computer Vision and Pattern Recognition, 2013, pp. 1233–1240.
[56] K.E. Van de Sande, T. Gevers, C.G. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1582–1596.
[57] I.S. Dhillon, S. Mallela, R. Kumar, A divisive information theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res. 3 (2003) 1265–1287.
[58] F. Perronnin, C. Dance, Fisher kernels on visual vocabularies for image categorization, in: Proceedings of Computer Vision and Pattern Recognition, 2007, pp. 1–8.
[59] K. De Cock, B. De Moor, Subspace angles and distances between arma models, in: Proceedings of the International Symposium of Mathematical Theory of Networks and Systems, vol. 1, Citeseer, 2000.
[60] R.J. Martin, A metric for arma processes, IEEE Trans. Signal Process. 48 (4) (2000) 1164–1170.
[61] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: Proceedings of Computer Vision and Pattern Recognition, 2014, pp. 1725–1732.
[62] R. Péteri, S. Fazekas, M.J. Huiskes, Dyntex: a comprehensive database of dynamic textures, Pattern Recognit. Lett. 31 (12) (2010) 1627–1632.
[63] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 27.
[64] A. Ravichandran, P. Favaro, R. Vidal, A unified approach to segmentation and categorization of dynamic textures, in: Proceedings of Asian Conference on Computer Vision, Springer, Queenstown, New Zealand, 2011, pp. 425–438.
[65] W.N. Gonçalves, B.B. Machado, O.M. Bruno, Spatiotemporal gabor filters: a new method for dynamic texture recognition, arXiv preprint arXiv:1201.3612.
[66] X. Li, A. Dick, H. Wang, C. Shen, A. van den Hengel, Graph mode-based contextual kernels for robust svm tracking, in: Proceedings of International Conference on Computer Vision, 2011, pp. 1156–1163.
[67] X. Li, A. Dick, C. Shen, Z. Zhang, A. van den Hengel, H. Wang, Visual tracking with spatio-temporal Dempster–Shafer information fusion, IEEE Trans. Image Process. 22 (8) (2013) 3028–3040.
[68] X. Li, C. Shen, A. Dick, A. van den Hengel, Learning compact binary codes for visual tracking, in: Proceedings of Computer Vision and Pattern Recognition, 2013, pp. 2419–2426.
[69] X. Li, A. Dick, C. Shen, A. van den Hengel, H. Wang, Incremental learning of 3ddct compact representations for robust visual tracking, IEEE Trans. Pattern Anal. Mach. Intell. 35 (4) (2013) 863–881.

Feng Yang received the B.S. degree of photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2013. She is currently working toward the Ph. D. degree in the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing. Her research interests include dynamic texture analysis and pattern recognition.

Gui-Song Xia received the B.Sc. degree in electronic engineering and the M.Sc. degree in signal processing from Wuhan University, Wuhan, China, in 2005 and 2007 respectively, and the Ph.D. degree in image processing and computer vision from the CNRS LTCI, TELECOM ParisTech (also known as École Nationale Supérieure des Télécommunications), Paris, France, in 2011. Since March 2011, he has been a Postdoctoral Researcher with the Centre de Recherche en Mathmatiques de la Decision (CEREMADE), CNRS, Paris-Dauphine University, Paris, France, for one and a half years. Currently, he is an Associate Professor with the State Key Laboratory of Information Engineering, Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. His research interests include mathematical image modeling, texture synthesis, image indexing and content-based retrieval, structures from motions, perceptual grouping, and remote-sensing imaging.

Gang Liu received the B.S. in Communication Engineering and M.S. degree in Electrical Engineering from Wuhan University, China, in 2011 and 2013 respectively. He is currently working towards a Ph.D. degree in Computer Vision at Telecom ParisTech, France. His research involves texture analysis, including texture retrieving and classification. He also study remote sensing image processing and the application of Poisson point process in image analysis.

Liangpei Zhang received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Chinese Academy of Sciences, Xian, China, in 1988, and the Ph. D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998. He is currently the Head of the Remote Sensing Division, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. He is also a Chang-Jiang Scholar Chair Professor appointed by the Ministry of Education of China. He is currently a Principal Scientist with the China State Key Basic Research Project (2011–2016) appointed by the Ministry of National Science and Technology of China to lead the remote sensing program in China. He has authored more than 410 research papers. He is the holder of 15 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence. He is a Fellow of the Institution of Engineering and Technology, an Executive Member (Board of Governor) of the China National Committee of the International Geosphere–Biosphere Programme, and an Executive Member of the China Society of Image and Graphics. He regularly serves as a Cochair of the series SPIE Conferences on Multispectral Image Processing and Pattern Recognition, Conference on Asia Remote Sensing, and many other conferences. He edits several conference proceedings, issues, and Geoinformatics symposiums. He also serves as an Associate Editor of the International Journal of Ambient Computing and Intelligence, the International Journal of Image and Graphics, the International Journal of Digital Multimedia Broadcasting, the Journal of Geo-spatial Information Science, the Journal of Remote Sensing, and the IEEE Transactions on Geoscience and Remote Sensing.



Xin Huang received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2009. He is currently a Full Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan, China (LIESMARS). He has authored more than 50 peer-reviewed articles in international journals. His research interests include hyperspectral data analysis, high-resolution image processing, pattern recognition, and remote sensing applications. He was a recipient of the Top-Ten Academic Star of Wuhan University, Wuhan, China, in 2009; the Boeing Award for Best Paper in Image Analysis and Interpretation from the American Society for Photogrammetry and Remote Sensing, in 2010; the New Century Excellent Talents in University from the Ministry of Education of China, in 2011; and the National Excellent Doctoral Dissertation Award of China, in 2012. In 2011, he was recognized by the IEEE Geoscience and Remote Sensing Society (GRSS) as a Best Reviewer of the IEEE Geoscience and Remote Sensing Letters. He was the winner of the IEEE GRSS 2014 Data Fusion Contest. Since 2014, he has been an Associate Editor of the IEEE Geoscience and Remote Sensing Letters.