Neurocomputing 171 (2016) 1230–1241

Contents lists available at ScienceDirect

Neurocomputing

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

Dynamic texture and scene classification by transferring deep image features

Xianbiao Qi [a,b], Chun-Guang Li [b,*], Guoying Zhao [a], Xiaopeng Hong [a], Matti Pietikäinen [a]

[a] Center for Machine Vision Research, University of Oulu, PO Box 4500, FIN-90014, Finland
[b] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

ARTICLE INFO

Article history:
Received 17 January 2015
Received in revised form
18 June 2015
Accepted 22 July 2015
Communicated by Qingshan Liu
Available online 5 August 2015

Keywords:
Dynamic texture classification
Dynamic scene classification
Transferred ConvNet feature
Convolutional neural network

ABSTRACT

Dynamic texture and scene classification are two fundamental problems in understanding natural video content. Extracting robust and effective features is a crucial step towards solving these problems. However, the existing approaches suffer from the sensitivity to either varying illumination, or viewpoint changes, or even camera motion, and/or the lack of spatial information. Inspired by the success of deep structures in image classification, we attempt to leverage a deep structure to extract features for dynamic texture and scene classification. To tackle with the challenges in training a deep structure, we propose to transfer some prior knowledge from image domain to video domain. To be more specific, we propose to apply a well-trained Convolutional Neural Network (ConvNet) as a feature extractor to extract mid-level features from each frame, and then form the video-level representation by concatenating the first and the second order statistics over the mid-level features. We term this two-level feature extraction scheme as a Transferred ConvNet Feature (TCoF). Moreover, we explore two different implementations of the TCoF scheme, i.e., the *spatial* TCoF and the *temporal* TCoF. In the spatial TCoF, the mean-removed frames are used as the inputs of the ConvNet; whereas in the temporal TCoF, the differences between two adjacent frames are used as the inputs of the ConvNet. We evaluate systematically the proposed spatial TCoF and the temporal TCoF schemes on three benchmark data sets, including DynTex, YUPENN, and Maryland, and demonstrate that the proposed approach yields superior performance.

1. Introduction

Dynamic texture and scene classification are two fundamental problems in understanding natural video content and have gained considerable research attention in the past decade [46,31,14,47,13,12,35,6,43,16,11,22,15,38]. Roughly, dynamic textures can be described as visual processes, which consist of a group of particles with random motions; whereas dynamic scenes can be considered as places where events occur. In Fig. 1, we show some sample images from a dynamic scene data set YUPENN [11]. The ability to automatically categorize dynamic textures or scenes is useful, since it can be used to recognize the presence of events, surfaces, actions, and phenomena in a video surveillance system.

However, categorizing automatically dynamic textures or dynamic scenes is a challenging problem, due to the existence of a wide range of naturally occurring variations in a short video, e.g., illumination variations, viewpoint changes, or even significant camera motions. It is commonly accepted that constructing a robust and effective representation of a video sequence is a crucial step towards solving these problems. In the past decade, a large number of methods for video representation have been proposed, e.g., Linear Dynamic System (LDS) based methods [14,35,1,6], GIST based method [29], Local Binary Pattern (LBP) based methods [28,47,32–34], and Wavelet based methods [12,16,19,46]. Unfortunately, the existing approaches suffer from the sensitivity to either varying illumination, or viewpoint changes, or even the camera motion, and/or the lack of spatial information.

Recently there is a surge of research interests in developing *deep structures* for solving real world applications. Deep structure based approaches set up numerous recognition records in image classification [2,37,39,42,17],object detection [36], face recognition and verification [41,40], speech recognition [10], and natural language processing [8,9]. Inspired by the great success of deep structures in image classification, in this paper, we attempt to leverage a deep structure to extract features for dynamic texture and scene classification. However, learning a deep structure needs huge amounts of training data and is quite expensive in computational demand. Unfortunately, as in other video classification tasks, dynamic texture and scene classification suffer from the small size of training data. As a result, the lack of training data is actually an obstacle to deploy a deep structure for dynamic texture and scene classification tasks.

* Corresponding author.
E-mail addresses: qixianbiao@gmail.com (X. Qi), lichunguang@bupt.edu.cn (C.-G. Li), gyzhao@ee.oulu.fi (G. Zhao), xhong@ee.oulu.fi (X. Hong), mkp@ee.oulu.fi (M. Pietikäinen).

http://dx.doi.org/10.1016/j.neucom.2015.07.071
0925-2312/

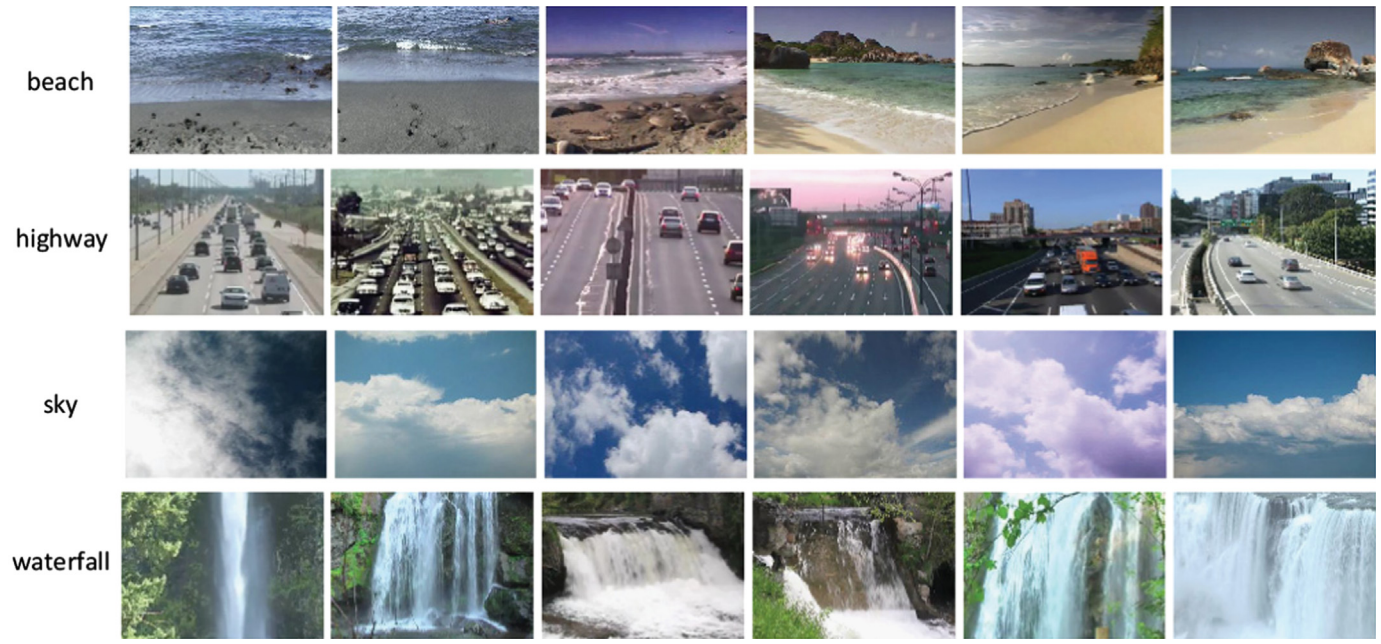


Fig. 1. Sample images from dynamic scene data set YUPENN. Each row corresponds a category.

By noticing that there is a lot of work in learning deep structures for classifying images [2,37,39,42,17], in this paper, we attempt to transfer the knowledge in image domain to compensate the deficiency of training data in training a deep structure to represent dynamic textures and scenes. Concretely, we propose to apply a well-trained Convolutional Neural Network (ConvNet) as a feature extractor to extract mid-level features from each frame in a video, and then form the video-level representation by concatenating the first and the second order statistics over the mid-level features. We term this two-level feature extraction scheme as a Transferred ConvNet Feature (TCoF).

Our aim in this paper is to explore a robust and effective way to capture the spatial and temporal information in dynamic textures and scenes. Briefly, our contributions can be highlighted as follows:

- We propose a two-level feature extraction scheme to represent dynamic textures and scenes, which applies a trained Convolutional Neural Network (ConvNet) as a feature extractor to extract mid-level features from each frame in a video and then computes the first and the second order statistics over the mid-level features to form the video-level representation. To the best of our knowledge, this is the first investigation of using a deep network with transferred knowledge to represent dynamic texture and dynamic scenes.
- We systematically investigate the performance of the video-level representation which is formed by using the spatial and/or the temporal mid-level features on three benchmark data sets. Experimental results show that: (a) the spatial feature is more effective for categorizing the dynamic textures and dynamic scenes and (b) when the video is stabilized the temporal feature could provide some complementary information to the spatial feature, which captures the intrinsic variation of motion patterns.

The remainder of the paper is organized as follows. We review the related studies in Section 2 and present our proposals in Section 3. We evaluate the proposed spatial and temporal TCoF in Section 4 and finally we conclude this paper with a discussion in Section 5.

2. Related work

In the literature, there are numerous approaches for dynamic texture and scene classification. While being closely related, dynamic texture classification [46,31,14,47,13,12,35,6] and dynamic scene classification [43,16,11,22,15,38] are usually considered separately as two different problems so far.

The research history of dynamic texture classification is much longer than that of the dynamic scene. The later, as far as we know, started since two dynamic scene data sets—Maryland Dynamic Scene data set "in the wild" [38] and York stabilized Dynamic Scene data set [31]—were released. Although there might not be a clear distinction in nature, the slight difference of dynamic texture from dynamic scene is that the frames in a video of dynamic texture consist of images with richer texture; whereas the frames in a video of dynamic scene are a natural scene evolving over time. In addition, compared to dynamic textures which are usually stabilized videos, the dynamic scene data usually include some significant camera motions.

The critical challenges in categorizing the dynamic textures or scenes come from the wide range of variations around the naturally occurring phenomena. To overcome the difficulties, numerous methods for video representation have been proposed. Among them, Linear Dynamic System (LDS) based methods [14,35,1,6], GIST based method [29], Local Binary Pattern (LBP) based methods [28,47,32–34], and wavelet based methods [12,16,19,46] are the most widely used. LDS is a statistical generative model which captures the spatial appearance and dynamics in a video [14]. While LDS yields promising performance on viewpoint-invariant sequences, it performs poorly on viewpoint-variant sequences [35,1,6]. Besides, it is also sensitive to illumination variations. GIST [29] represents the spatial envelope of an image (or a frame in video) holistically by Gabor filter. However, GIST suffers from scale and rotation variations. Among LBP based methods, Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) [47] is the most widely used. LBP-TOP describes a video by computing LBPs from three orthogonal planes (*xy*, *xt* and *yt*) *only*. After LBP-TOP, several variants have been proposed, e.g., Local Ternary Pattern on Three Orthogonal Planes (LTP-TOP) [34], Weber Local Descriptor on Three Orthogonal Planes (WLD-TOP) [7], Local

Phase Quantization on Three Orthogonal Planes (LQP-TOP) [34]. While LBP-TOP and its variants are effective at capturing spatial and temporal information and robust to illumination variations, they are suffering from camera motions. Recently, wavelet based methods are also proposed, e.g., Spatiotemporal Oriented Energy (SOE) [16], Wavelet Domain Multifractal Analysis (WDMA) [19], and Bag-of-Spacetime-Energy (BoSE) [16]. Combined with the Improved Fisher Vector (IFV) encoding strategy [30,4], BoSE leads to the state-of-the-art performance on dynamic scene classification. However, the computational cost of BoSE is expensive due to slow feature extraction and quantization.

The aforementioned methods can be roughly divided into two categories: the *global* approaches and the *local* approaches. The *global* approaches extract features from each frame in a video sequence by treating each frame as a whole, e.g., LDS [14] and GIST [29]. While the global approaches describe the spatial layout information well, they suffer from the sensitivity to illumination variations, viewpoint changes, or scale and rotation variations. The *local* approaches construct a statistics (e.g., histogram) on a bunch of features extracted from local patches in each frame or local volumes in a video sequence, including LBP-TOP [47], LQP-TOP [34], BoSE [16], Bag of LDS [35]. While the local approaches are robust against transformations (e.g., rotation, illumination), they suffer from the lack of spatial layout information which is important to represent dynamic texture or dynamic scene.

In this paper, we attempt to leverage a deep structure with transferred knowledge from image domain to construct a robust and effective representation for dynamic textures and scenes. To be more specific, we propose to use a pre-trained ConvNet—which has been trained on the large-scale dataset ImageNet [23,37,2,39,42,17]—as transferred prior knowledge, and then fine-tune the ConvNet with the frames in the videos of training set. Equipped with a trained ConvNet, we extract mid-level features from each frame in a video and represent a video by the concatenation of the first and the second order statistics over the mid-level features.

Compared to the previous studies, our approach possesses the following advantages:

- Our approach represents a video with a two-level strategy. The deep structure used in the frame level is easier to train or even train-free, since we are able to adopt prior knowledge from image domain.
- The extracted mid-level (i.e., frame-level) features are robust to translations, small scale variations, partial rotations, and illumination variations.
- Our approach represents a video sequence by a concatenation of the first and the second order statistics of the frame-level features. This process is fast and effective.

In the next section, we will present the framework and two different implementations of our proposal.

3. Our proposal: Transferred ConvNet feature (TCoF)

Our TCoF scheme consists of three stages:

- Constructing a ConvNet with transferred knowledge from image domain.
- Extracting the mid-level feature with the ConvNet from each frame in a video.
- Forming the video-level representation by concatenating the calculated first and the second order statistics over the mid-level features.

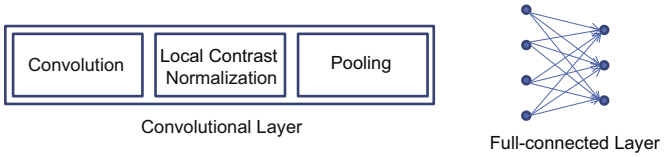


Fig. 2. The typical structure of a ConvNet.

3.1. Convolutional Neural Network with transferred knowledge for extracting frame-level features

There are a lot of work in learning deep structures for classifying images. Among them, Convolutional Neural Networks (ConvNets) have been demonstrated to be extremely successful in computer vision [24,23,36,20,5,2,39,42,17].

We show a typical structure of a ConvNet in Fig. 2. The ConvNet consists of two types of layers: convolutional layers and full-connected layers. The convolutional part, as shown in the left panel of Fig. 2, consists of three components – convolutions, Local Contrast Normalization (LCN), and pooling. Among the three components, the convolution block is compulsory, and LCN and the pooling are optional. The convolution components capture complex image structures. The LCN achieves robustness to illumination variations. The pooling component can *not only* yield partial invariance to scale variations and translations, but *also* reduce the complexity for the downstream layers. Due to sharing parameters which is motivated by the *local reception field* in biological vision system, the number of free parameters in the convolutional layer is significantly reduced. The full-connected layer, as shown in the right panel of Fig. 2, is the same as a multi-layer perception neural network.

In our TCoF framework, we use a ConvNet with five convolutional layers and two full-connected layers as shown in Fig. 3, which is the same as the most successful ConvNet implementation introduced in [5,23], to extract the mid-level features from each frame in a video. Note that we remove the final full-connected layer in the ConvNet introduced in [5,23].

As mentioned previously, training well a deep network like that in Fig. 3 needs huge amounts of training data and is quite expensive in computational demand. In our case, for dynamic texture or scene, the training data is limited. Instead of training a deep network from scratch, which is quite time-consuming, we propose to use the pre-trained ConvNet [5,23] as the initialization, and fine-tune the ConvNet with the frames in videos from training data if necessary. By using a good initialization, we virtually transfer miscellaneous prior knowledge from image domain (e.g., data set ImageNet) to the dynamic texture and scene tasks.

3.2. Construct video-level representation

Given a video sequence containing N frames, the ConvNet yields N ConvNet features. Note that we use each frame in the video subtracted from *an averaged image* as the input to the ConvNet in TCoF.

Denote X as a set of the ConvNet features $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ where $\mathbf{x}_i \in R^d$ is the ConvNet feature extracted from i-th frame. We extract the *first* and the *second order* statistics on feature set X.

The first-order statistics of X is the *mean* vector which is defined as follows:

$$\mathbf{u} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i, \tag{1}$$

where $\mathbf{u}$ captures the average behaviors of the N ConvNet features which reflect the average characteristics in the video sequence.

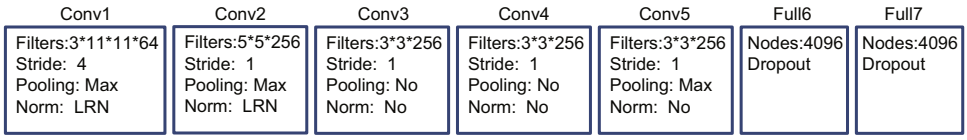


Fig. 3. The architecture of the ConvNet used in our TCoF scheme.

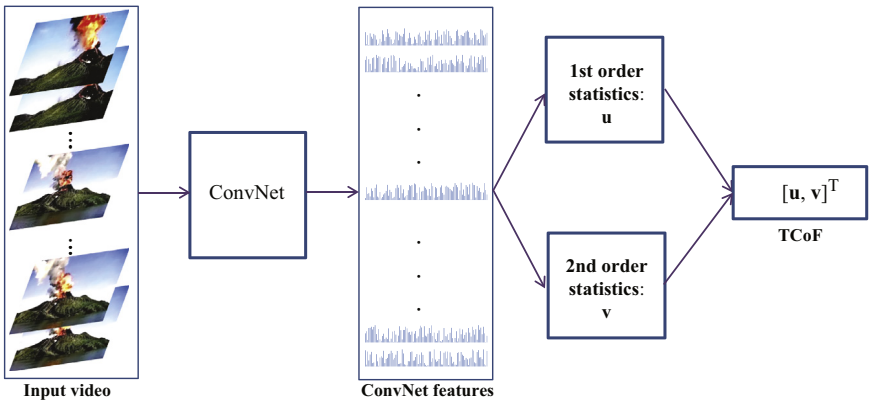


Fig. 4. An illustration of our TCoF scheme.

The second-order statistics is the *covariance* matrix which is defined as follows:

$$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{u})(\mathbf{x}_i - \mathbf{u})^{\top}, \tag{2}$$

where $\mathbf{S}$ describes the variation of the N ConvNet features from the mean vector $\mathbf{u}$ and the correlation among different dimensions. The dimension of covariance feature is $\frac{d(d+1)}{2}$. In our case, we use $d=4096$. If we use all the entries in S, the dimension would be $\frac{d(d+1)}{2} = 8,390,656$ (since S is symmetric). We notice of that the correlation between two different dimensions is relatively low and less informative. Therefore, by considering of the tradeoff in the curse of dimensionality and the power of representation, we propose to extract *only* the diagonal entries in $\mathbf{S}$ as the second-order feature, that is,

$$\mathbf{v} = \mathrm{diag}(\mathbf{S}), \tag{3}$$

where $\mathrm{diag}(\cdot)$ means to extract the diagonal entries of a matrix as a vector. The vector $\mathbf{v}$ is d-dimensional and captures the variations along each dimension in the ConvNet features.

Having calculated the first and the second order statistics, we form the video-level representation, TCoF, by concatenating $\mathbf{u}$ and $\mathbf{v}$, i.e.,

$$\mathbf{f} = \begin{bmatrix}\mathbf{u}\\ \mathbf{v}\end{bmatrix}, \tag{4}$$

where the dimension of a TCoF representation is $2d$.

For clarity, we illustrate the flowchart of constructing the TCoF representation for a video sequence in Fig. 4.

Remarks 1. Our proposed TCoF is a global approach. Since the spatial layout information can be captured well, we term the TCoF scheme described above as the *spatial* TCoF. Our proposed TCoF possesses robustness to translations, small scale variations, partial rotations, and illumination variations owing to the ConvNet component. In addition, the process of extracting a TCoF vector is extremely fast since the ConvNet adopts a so-called *stride* tactics and the second step in TCoF is to calculate the two statistics.

3.3. Modeling temporal information

While it is well accepted that dynamic information can enrich our understanding of the textures or scenes, modeling the dynamic information is difficult. Unlike the motion of rigid objects, dynamic textures and scenes are usually involving of non-rigid objects and thus the optical flow information seems relatively random.

In this paper, we propose to use the difference of adjacent two frames in a short-time to capture the random-like micro-motion patterns. To be specific, we take the difference between the $(i+\tau)$-th and i-th frames as the input of the ConvNet component in TCoF scheme, where $\tau \in \{1, \ldots, N-1\}$ is an integer which corresponds to the resolution in time to capture the random-like micro-motion patterns. In practice, we set τ as a small integer, e.g., 1, 2 or 3.

Given a video sequence containing N frames, the ConvNet produces $N-\tau$ *temporal* frame-level features. Then we extract the *first* and the *second order* statistics over the temporal ConvNet features to form a temporal TCoF for the input video, with the same way as for the spatial TCoF in Section 3.2.

Remarks 2. The temporal TCoF differs from the spatial TCoF in the input of the ConvNet. In the spatial TCoF, we take each frame in a video subtracted from a precalculated average image as input; whereas in the temporal TCoF we take the difference of two frames in a short-time and there is no need to subtract an average image.

Remarks 3. Note that t-TCoF is constructed based on the difference between frames at interval τ, where τ is a parameter to control the 'scale' or 'resolution' in temporal domain to characterize the pattern of motions. Roughly, when the video is stabilized,[1] the t-TCoF reflects the intrinsic information about the mot ion patterns; whereas the s-TCoF captures more appearance

[1] Otherwise, i.e., when huge camera motion exists, the t-TCoF features could not represent the true intrinsic motion of the dynamic textures (or scenes) but the motion of the camera.

information in each frame. This is the complementary property between t-TCoF and s-TCoF.

Remarks 4. In our proposed TCoF, we treat the extracted N ConvNet features as a set and ignore the sequential information in features. The rationale of this simplification comes from the property of dynamic textures and dynamic scenes. Note that the dynamic textures are visual processes of a group of particles with random motions, and dynamic scenes are places where natural events are occurring, thus the sequential information in these processes are relatively random and less critical. Experimental results in Section 4.3 support this point.

4. Experiments

In this section, we introduce the benchmark data sets, the baseline methods, and the implementation details, and then present the experimental evaluations of our approach.

4.1. Data sets description

DynTex [31] is a widely used dynamic texture data set, containing 656 videos with each sequence recorded in PAL format. The sequences in DynTex are divided into three subsets – "Alpha", "Beta", and "Gamma": (a) "Alpha" subset contains 60 sequences which are equally divided into 3 categories: "sea", "grass" and "trees'; (b) "Beta" subset consists of 162 sequences which are grouped into 10 categories: "sea", "grass", "trees", "flags", "calm water", "fountains", "smoke", "escalator", "traffic", and "rotation"; (c) "Gamma"' subset is composed of 264 sequences which are grouped into 10 categories: "flowers", "sea", "trees without foliage", "dense foliage", "escalator", "calm water", "flags", "grass", "traffic" and "fountains". Compared to "Alpha" and "Beta" subsets, "Gamma" subset contains more complex variations, e.g., scale, orientation, and etc. Sample frames from the three subsets are shown in Fig. 5.

YUPENN [11] is a "stabilized" dynamic scenes data set. This data set was introduced to emphasize scene-specific temporal information. YUPENN consists of fourteen dynamic scene categories with 30 color videos in each category. The sequences in YUPENN have significant variations, such as frame rate, scene appearance, scale, illumination, and camera viewpoint. Some sample frames are shown in Fig. 6.

Maryland [38] is a dynamic scene data set which was introduced earlier than the YUPENN [11]. It consists of 13 categories with 10 videos per category. The data set have large variations in illumination, frame rate, viewpoint, and scale. Besides, there are variations in resolution and camera dynamics. Some sample frames are shown in Fig. 7.

4.2. Baselines and implementation details

Baselines: We compare our proposed TCoF approach with the following state-of-the-art methods.[2]

- *GIST* [29]: Holistic representation of the spatial envelope which is widely used in 2D static scene classification.
- *Histogram of Flow* (*HOF*) [26]: The HOF is a well-known descriptor in action recognition.
- Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) [47].
- Chaotic Dynamic Features (Chaos) [38].
- Slow Feature Analysis (SFA) [43].
- Synchrony Autoencoder (SAE) [22].
- Synchrony K-means (SK-means) [22].
- Complementary Spacetime Orientation (CSO) [15]: In CSO, the complementary spatial and temporal features are fused in a random forest framework.
- Bag of Spacetime Energy (BoSE) [16].
- Bag of System Trees (BoST) [27].

Implementation details: In both the spatial TCoF (s-TCoF) and the temporal TCoF (t-TCoF), we resize the frames into 224×224 and normalize both s-TCoF and t-TCoF with L_2-norm, respectively. For the combination of both the spatial and temporal TCoF, we take the concatenation of the two normalized s-TCoF and t-TCoF and denote it as st-TCoF. We do not use any data augmentation method. For our t-TCoF, we use $\tau=3$. We use the MatConvNet [45] and CAFFE toolbox [20] to extract the proposed TCoFs. Note that we take the weights in each layer but the final full connection layer in the well-trained ConvNet [23] as the initialization. And then, the whole ConvNet could be fine-tuned with the training data. While the fine-tuning stage is easier than training a ConvNet from scratch with random initialization, we observed that the improvement by the extra fine-tuning was minor. Thus we use the ConvNet without a further fine-tuning to extract the mid-level features.[3] For LBP-TOP, we use the best performing setting of $\mathrm{LBP}-\mathrm{TOP}_{8,8,8,1,1,1}$, and the χ^2 kernel. To fairly compare with previous methods, we test our approach and other baselines with both the nearest neighbor (NN) classifier and SVM classifier separately. In SVM, we use a linear SVM with Libsvm toolbox [3], in which the tradeoff parameter C is fixed to 40 in all our experiments. Following the standard protocol, we use Leave-One-Out (LOO) cross-validation.

4.3. Evaluation of the spatial and temporal TCoFs

In this subsection, we evaluate the influence of different parameters and components when using the spatial and temporal TCoFs on DynTex, YUPENN, and Maryland data sets.

Evaluation of the effect of d: To evaluate the effect of the parameter d, which is the size of the fully connected layer, we prepare experiments on DynTex data set with *s*-TCoF to show classification results with different d. Experimental results are presented in Table 1.

As can be observed that the performance keeps consistently increasing when increasing d. When $d=4096$, *s*-TCoF could yield consistently good results. Therefore we use $d=4096$ for all experiments thereafter.

Effectiveness of the s-TCoF: Since the s-TCoF features are constructed by accumulating all features in all frames, it is interesting to investigate the effect on the final performance of with different number of frames. To this end, we evaluate the s-TCoF using seven different settings: (1) using only the *first* frame in a video, (2) the first $\frac{N}{8}$ frames, (3) the first $\frac{N}{4}$ frames, (4) the first $\frac{N}{2}$ frames, and (5) all N frames. Experimental results are shown in Table 2.

From Table 2, we can see that the spatial TCoF performs well even when using the first frame *only*, on DynTex and YUPENN data sets. This confirmed the effectiveness of the spatial TCoF scheme. Note that intuitively, the performance should tend to increase as more frames were used. In practice, however, the dynamics in the video vary significantly and quickly. In some part of the video, the

[2] For the LBP-TOP, we report the results with our own implementation and for other methods we cite the results from their papers.

[3] Note that we use the Leave-One-Out cross-validation to evaluate the performance. The training data are changed from each trial. If we chose to fine-tune the ConvNet, we should fine-tune for each trial. Since the improvements were minor, we report the results without a fine-tuning to keep all experimental results are repeatable.

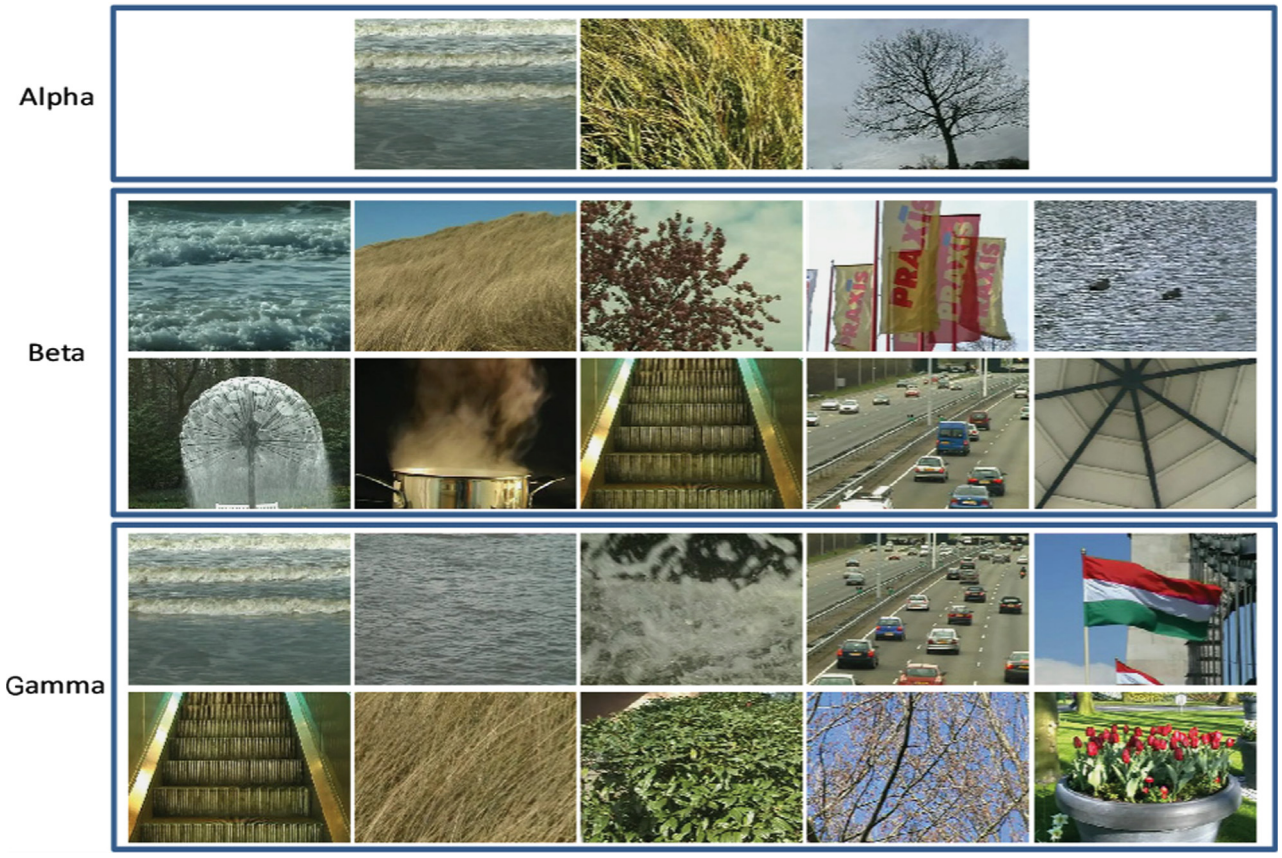


Fig. 5. Sample frames from DynTex data set. The "Alpha" , "Beta", and "Gamma" show the sample frames from each category in the data set.

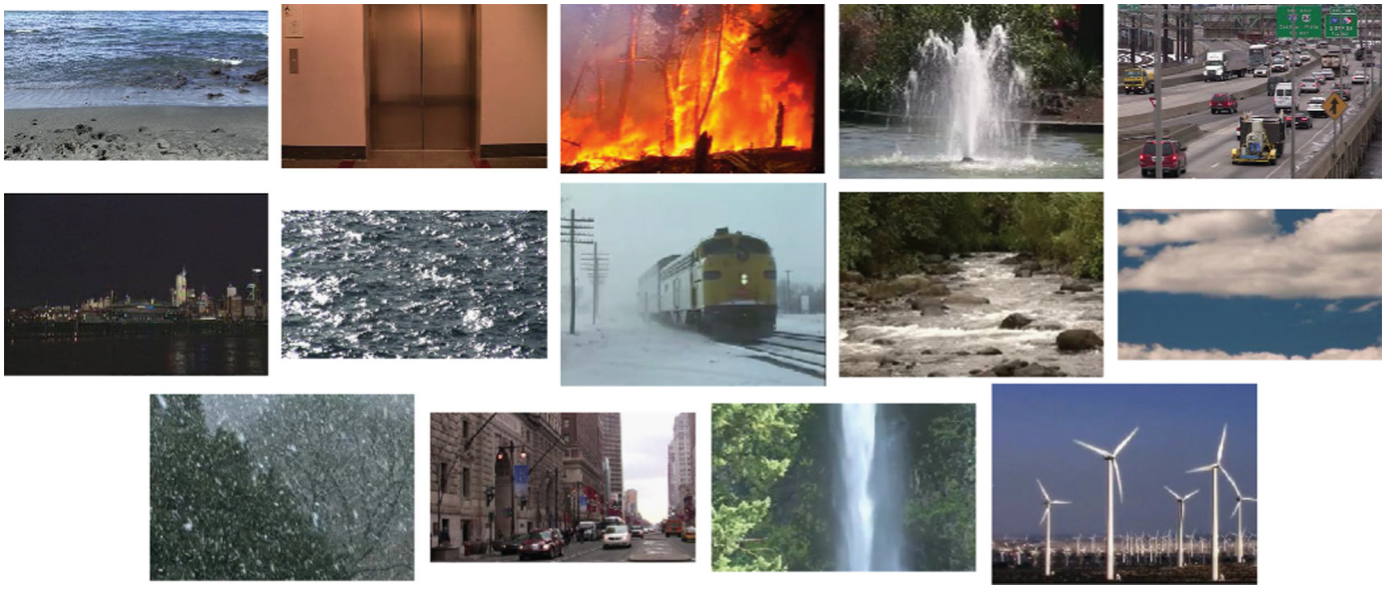


Fig. 6. Sample frames from dynamic scene data set YUPENN. Each image corresponds to a category of video sequence.

dynamic textures or scenes might be relatively stable and thus have small variations which lead to slightly higher classification accuracy; whereas in some parts the dynamic textures or scenes might have significant variations which thus lead to a slightly lower accuracy. This accounts for the fluctuations in performance. It should be mentioned that although there were some fluctuations in performance the results on average are still improved.

Effectiveness of the t-TCoF: Recall that t-TCoF is constructed based on the difference between adjacent two frames at interval τ, where τ is a parameter to control the "resolution" in temporal domain to characterize the pattern of motions (when the video is stabilized). To evaluate the influence of the parameter τ, we conduct experiments on all data sets with t-TCoF for τ varying from 1 to 5.

Experimental results are shown in Table 3. As could be observed, that t-TCoF is not sensitive to the choice of τ. Note that almost all the results of s-TCoF in Table 2 outperform that of t-TCoF in Table 3. This suggests that s-TCoF is more effective than t-TCoF,

Fig. 7. Sample frames from Maryland scenes data set.

Table 1
Evaluation of the effect of d on DynTex dynamic texture data set.

$d=$	128	1024	2048	4096
Alpha(NN)	98.33	**100**	**100**	**100**
Alpha(SVM)	**100**	**100**	**100**	**100**
Beta(NN)	96.91	97.53	98.77	**99.38**
Beta(SVM)	96.30	98.15	98.77	**100**
Gamma(NN)	95.45	96.21	96.33	**96.59**
Gamma(SVM)	95.83	97.73	97.95	**98.11**

Table 2
Evaluation of the spatial TCoF (s-TCoF) by using different number of frames.

Datasets	1st	$\frac{N}{8}$	$\frac{N}{4}$	$\frac{N}{2}$	N
Alpha (NN)	100	100	100	100	100
Alpha (SVM)	100	100	100	100	100
Beta (NN)	98.77	99.38	99.38	98.77	99.38
Beta (SVM)	99.38	100	100	99.38	100
Gamma (NN)	97.73	97.35	96.97	96.97	96.59
Gamma (SVM)	97.73	98.11	98.11	98.86	98.11
YUPENN (NN)	95.71	96.43	96.19	96.43	95.48
YUPENN (SVM)	96.90	96.90	96.90	**97.14**	96.90
Maryland (NN)	72.31	**80.00**	75.38	77.69	76.92
Maryland (SVM)	80.00	83.85	80.77	83.08	**88.46**

since the randomness of the micro-motions in dynamic texture or natural dynamic scene makes the temporal information less critical. Nevertheless, t-TCoF could provide complementary information to s-TCoF in some cases that will be shown later.

4.4. Comparisons with the State-of-the-Art Methods

Dynamic texture classification on DynTex: We conduct a set of experiments to compare our methods with LBP-TOP. Experimental results are shown in Table 4.

We observe from Table 4 that:

Table 3
Evaluating the performance of temporal TCoF (t-TCoF) as a function of parameter τ.

Datasets	$\tau=1$	$\tau=2$	$\tau=3$	$\tau=4$	$\tau=5$
Alpha (NN)	98.33	96.67	96.67	96.67	96.67
Alpha (SVM)	98.33	98.33	98.33	98.33	98.33
Beta (NN)	97.53	96.91	97.53	97.53	97.53
Beta (SVM)	96.30	97.53	97.53	97.53	97.53
Gamma (NN)	93.56	94.32	93.18	93.18	93.94
Gamma (SVM)	95.83	95.83	95.45	94.70	95.45
YUPENN (NN)	90.24	91.19	92.38	93.57	93.10
YUPENN (SVM)	94.52	96.19	96.67	96.90	97.14
Maryland (NN)	55.38	56.92	57.69	61.54	63.85
Maryland (SVM)	66.92	62.31	61.54	63.85	63.85

Table 4
Classification results on DynTex dynamic texture data set. The performance of LBP-TOP is based on our implementation.

Datasets	LBP-TOP	s-TCoF	t-TCoF	st-TCoF
Alpha (NN)	96.67	**100**	96.67	98.33
Alpha (SVM)	98.33	**100**	98.33	**100**
Beta (NN)	85.80	**99.38**	97.53	98.15
Beta (SVM)	88.89	**100**	97.53	**100**
Gamma (NN)	84.85	95.83	93.56	**98.11**
Gamma (SVM)	94.18	**98.11**	95.45	**98.11**

1. The s-TCoF and st-TCoF perform the best on data subsets Alpha and Beta. This results confirm that s-TCoF and st-TCoF are both effective for dynamic texture classification.
2. On Gamma subset, s-TCoF and t-TCoF significantly outperform LBP-TOP. Moreover by combining s-TCoF with t-TCoF, st-TCoF achieves the best result. This result implies that t-TCoF might provide complementary information to s-TCoF.

Dynamic scene classification on YUPENN: We compare our methods with the state-of-the-art methods, including CSO, GIST, SFA, SOE, SAE, BOSE, SK-means, and LBP-TOP, and the experimental results are presented in Tables 5 and 6.

Table 5
Classification results on dynamic scene data set YUPENN. The results of are taken from the corresponding papers. The performance of LBP-TOP is based on our implementation.

Methods	GIST	SOE	LBP-TOP	SFA	BoST	CSO	SK-means	SAE	BoSE	s-TCoF	t-TCoF	st-TCoF
NN	56	74	75.95	–	–	–	–	80.7	–	96.43	93.10	**98.81**
SVM	–	80.71	85.29	85.48	85.47	85.95	95.2	96.0	96.19	97.14	97.86	**99.05**

Table 6
Category-wise accuracy (%) for different methods on dynamic scene data set YUPENN. Our methods and LBP-TOP are based on our implementation and use linear SVM classifier. The results of BoST are taken from [27]. The other results are taken from [16].

Categories	Chaos+ GIST	HOF+ GIST	SOE	LBP-TOP	BoST	SFA	CSO	BoSE	s-TCoF	t-TCoF	st-TCoF
Beach	30	87	93	87	83	93	100	100	97	97	97
Elevator	47	87	100	97	100	97	100	97	100	100	100
Forest fire	17	63	67	87	100	70	83	93	100	97	100
Fountain	3	43	43	37	67	57	47	87	100	97	100
Highway	23	47	70	77	87	93	73	100	97	100	100
Light storm	37	63	77	93	100	87	93	97	90	100	100
Ocean	43	97	100	97	90	100	90	100	100	100	100
Railway	7	83	80	80	80	93	93	100	97	100	100
Rush river	10	77	93	100	80	87	97	97	97	97	97
Sky-clouds	47	87	83	93	93	93	100	97	100	97	100
Snowing	47	10	87	83	83	70	57	97	90	97	97
Street	17	77	90	93	90	97	97	100	100	97	100
Waterfall	10	47	63	90	67	73	77	83	93	93	97
Wind. farm	17	53	83	67	77	87	93	100	100	100	100
Overall	22.86	68.33	80.71	84.29	85.47	85.48	85.95	96.19	97.14	97.86	**99.05**

Table 7
Classification results on dynamic scene data set Maryland. The results of are taken from the corresponding papers. The performance of LBP-TOP is based on our implementation.

Methods	LBP-TOP	SOE	SFA	CSO	BoSE	s-TCoF	t-TCoF	st-TCoF
NN	31.54	–	–	–	–	**74.62**	58.46	**74.62**
SVM	39.23	43.1	60	67.69	77.69	**88.46**	66.15	**88.46**

We observe from Table 5 that:

1. The s-TCoF and t-TCoF both outperform the state-of-the-art methods. Recall that YUPENN consist of stabilized videos. Hence these results confirm that both s-TCoF and t-TCoF are effective for dynamic scene data in a stabilized setting.
2. The combination of s-TCoF and t-TCoF, i.e., st-TCoF, performs the best. As shown in Table 6, s-TCoF and t-TCoF are complementary to each other on some categories, e.g., "Light Storm", "Railway", "Snowing", and "Wind. Farm". These results confirm the complementary property between s-TCoF and t-TCoF under a stabilized setting.

Dynamic scene classification on Maryland: We present the comparison of our methods with the state-of-the-art methods in Table 7 and the category-wise accuracy in Table 8.

As could be observed from Tables 7 and 8 that:

1. The s-TCoF outperforms the other methods significantly. This implies that the spatial information is extremely important for scene classification.
2. The results of t-TCoF are much worse than those of s-TCoF. This might be due to the significant camera motions in this data set. When huge camera motion exists, the t-TCoF features could not represent the true intrinsic motion of the dynamic textures or scenes but the motion of the camera. While the category-wise results in Table 8 still show a complementary property between s-TCoF and t-TCoF in same cases, however, the combination of s-TCoF and t-TCoF did not boost the overall performance due to the wrong motion information encoded in t-TCoF.

Remarks 5. Through Table 4,5 and 7, we observe that:

- The st-TCoF with SVM classifier yields consistently the best or at least matched overall performance, compared to the results of using s-TCoF or t-TCoF only. We account this superior performance to the feature selection property in SVM.
- The performance of st-TCoF with NN classifier is not consistently the best. Notice that the st-TCoF with NN outmatched or at least matched the results of using the single s-TCoF for three cases (i.e. YUPENN, Gamma, and Maryland), but degenerated on data sets Alpha and Beta. We account these two inferior results to the sensitivity of NN classifier to feature dimensionality. In other words, concatenating t-TCoF to s-TCoF did not bring enough discriminative information, but did doubled the dimensionality, and thus degenerated the results.

4.5. Further investigations and remarks

Data visualization: To show the discriminative power of the proposed approach, we use t-Stochastic Neighbor Embedding

Table 8
Category-wise accuracy (%) for different methods on dynamic scene data set Maryland. All methods use linear SVM classifier. Our methods and LBP-TOP are based on our implementation. The other results are taken from [16].

Categories	HOF+ GIST	Chaos+ GIST	SOE	SFA	CSO	BoSE	LBP-TOP	t-TCoF	s-TCoF	st-TCoF
Avalanche	20	60	40	60	60	60	10	30	**90**	80
Boiling water	50	60	50	70	80	70	60	60	80	**90**
Chaotic traffic	30	70	60	80	90	90	50	60	90	**100**
Forest fire	50	60	10	10	80	90	50	70	80	80
Fountain	20	60	50	50	80	70	70	80	**90**	**90**
Iceberg collapse	20	50	40	60	60	60	40	80	**90**	**90**
Landslide	20	30	20	60	30	60	30	90	**100**	**100**
Smooth traffic	30	50	30	50	50	70	30	60	80	**90**
Tornado	40	80	70	70	80	90	10	50	**100**	**100**
Volcanic eruption	20	70	10	80	70	80	70	90	**100**	**100**
Waterfall	20	40	60	50	50	100	60	10	90	90
Waves	80	80	50	60	80	90	10	50	80	70
Whirlpool	30	50	70	80	70	80	20	40	**80**	70
Overall	33.08	58.46	43.08	60.00	67.69	77.69	39.23	66.15	**88.46**	**88.46**

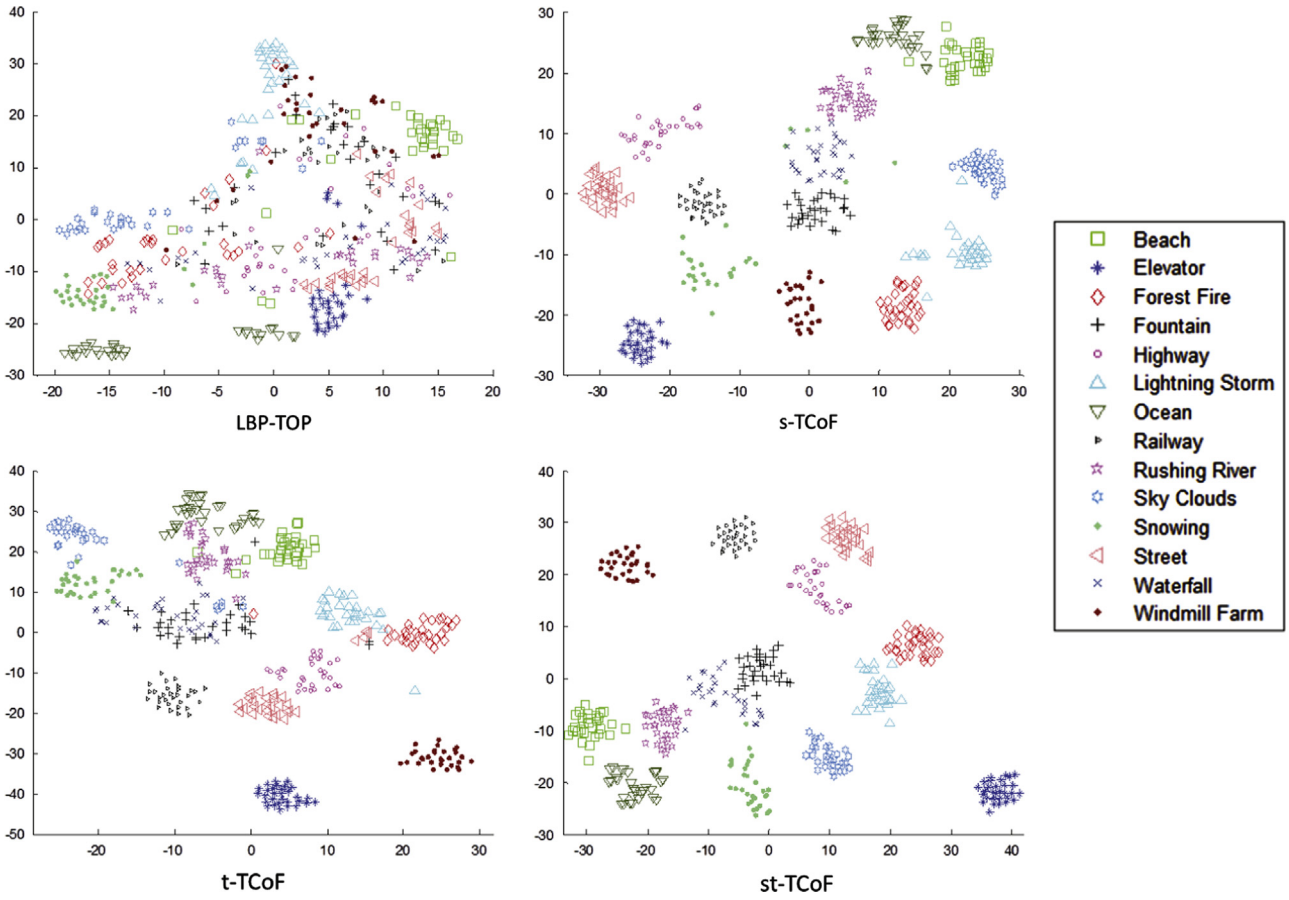


Fig. 8. Data visualization of LBP-TOP, s-TCoF, t-TCoF, and st-TCoF on dynamic scene data set YUPENN. Each point in the figure corresponds to a video sequence.

(t-SNE)[4] [44] to visualize the data distributions of the dynamic scene data sets YUPENN and Maryland. Results are shown in Figs. 8 and 9, respectively. We observe from Figs. 8 and 9 that s-TCoF, t-TCoF, and st-TCoF yield distinct separations between categories. These results reveal the effectiveness of our proposed TCoF approach vividly.

Remarks 6. Note that in our TCoF schemes, we treat the frames in a video as orderless images and extract mid-level features with a ConvNet. By doing so, the sequential information among features is ignored. The superior experimental results suggest that such a simplification is harmless. The sequential information in these processes contributes less discriminativeness, because of the fact that the dynamic textures can be viewed as visual processes of a group of particles with random motions, and the dynamic scenes are places where natural events are occurring. The effectiveness underlying our proposed TCoF approach for dynamic texture and scene classification is due to the following aspects:

- Rich filters' combination built on color channels in ConvNet describes richer structures and color information. The filters

[4] t-SNE is a (prize-winning) technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional data.

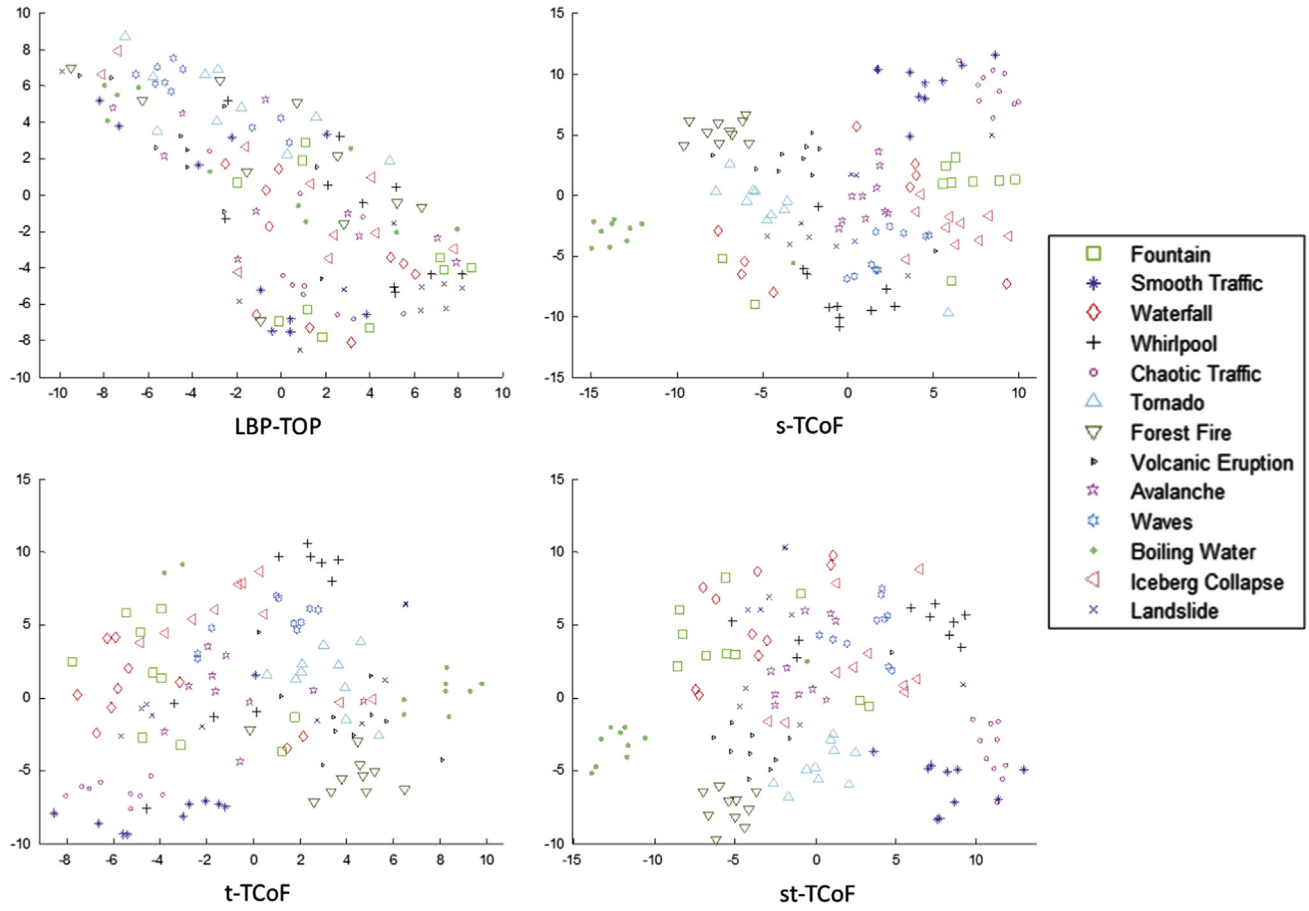


Fig. 9. Data visualization of LBP-TOP, s-TCoF, t-TCoF, and st-TCoF on dynamic scene data set Maryland. Each point in the figure corresponds to a video sequence.

that are built on different image patches capture stronger and richer structures compared to the hand-crafted features.

- ConvNet makes the extracted features robust to different sorts of image transformations due to the max-pooling and LCN components. Specifically, the max-pooling tactics makes ConvNet robust to translations, small scale variations, and partial rotations, and LCN makes ConvNet robust to illumination variations.
- The first and the second order statistics capture enough information over the mid-level features.
- When the video sequences are stabilized, the t-TCoF could provide complementary information to the s-TCoF.

5. Conclusion and discussion

We have proposed a robust and effective feature extraction approach for dynamic texture and scene classification, termed as Transferred ConvNet Features (TCoF), which was built on the first and the second order statistics of the mid-level features extracted by a ConvNet with transferred knowledge from image domain.

We have investigated two different implementations of the TCoF scheme, i.e., the *spatial* TCoF and the *temporal* TCoF. We have evaluated systematically the proposed approaches on three benchmark data sets and confirmed that: (a) the proposed *spatial* TCoF was effective, and (b) the *temporal* TCoF could provide complementary information when the camera is stabilized.

Unlike images, representing a video sequence needs to consider the following aspects:

1. To depict the spatial information. In most cases, we can recognize the dynamic textures and scenes from a single frame in a video. Thus, extracting the spatial information of each frame in a video might be an effective way to represent the dynamic textures or scenes.
2. To capture the temporal information. In dynamic textures or scenes, there are some specific micro-motion patterns. Capturing these micro-motion patterns might help to better understand the dynamic textures or scenes.
3. To fuse the spatial and temporal information. When the spatial and temporal information are complementary, combining both of them might boost the classification performance.

Different from the rigid or semi-rigid objects (e.g., actions), the dynamics of texture and scene are relatively random and non-directional. Whereas the temporal information might be a more important cue in action recognition [26,47,21], our investigation in this paper suggests that the sequential information in dynamic textures or scenes is *not that critical* for classification.

For future work, there are two methods that may be worth to explore. First, the Recurrent Neural Network [25,18] based method might be a better choice for the dynamic texture and scene classification. The RNN has demonstrated its effectiveness on sequence analysis. The dynamic textures and scenes can be modeled as a temporal process by the RNN. Second, the complementary properties between the proposed method and the

traditional Fisher Vector based method should be investigated. Our approach captures strong structural information. The IFV method which is built on orderless bag of word model captures non-structural information. The combination of the transferred deep features and IFV may further improve the classification accuracy. Notice that in ImageNet 2014 competition, the VGGnet [39], GoogleNet [42], and SPP-net [17] have demonstrated the state-of-the-art performance on image classification. Our approach building on deep image features can be further improved with these advanced deep features.

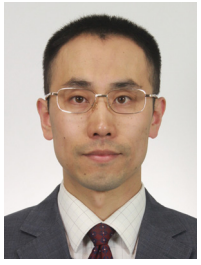
Acknowledgments

X. Qi, G. Zhao, X. Hong, and M. Pietikäinen are supported in part by the Academy of Finland and Infotech Oulu. C.-G. Li is supported partially by NSFC under Grant nos. 61273217 and 61175011, the Scientific Research Foundation for the Returned Overseas Chinese Scholars, Ministry of Education of China. The authors would like to thank Renaud Péteri, Richard P. Wildes and Pavan Turaga for sharing the DynTex dynamic texture, YUPENN dynamic scene, and Maryland dynamic scene data sets.

References

[1] B. Afsari, R. Chaudhry, A. Ravichandran, R. Vidal, Group action induced distances for averaging and clustering linear dynamical systems with applications to the analysis of dynamic scenes, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Providence, RI, USA, 2012, pp. 2208–2215.
[2] H. Azizpour, A.S. Razavian, J. Sullivan, A. Maki, S. Carlsson, From generic to specific deep representations for visual recognition, 2014, arXiv preprint arXiv:1406.5774.
[3] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 27.
[4] K. Chatfield, V. Lempitsky, A. Vedaldi, A. Zisserman, The devil is in the details: an evaluation of recent feature encoding methods, in: British Machine Vision Conference, 2011.
[5] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: delving deep into convolutional nets, in: British Machine Vision Conference, 2014.
[6] R. Chaudhry, G. Hager, R. Vidal, Dynamic template tracking and recognition, Int. J. Comput. Vis. 105 (1) (2013) 19–48.
[7] J. Chen, G. Zhao, M. Salo, E. Rahtu, M. Pietikäinen, Automatic dynamic texture segmentation using local descriptors and optical flow, IEEE Trans. Image Process. 22 (1) (2013) 326–339.
[8] R. Collobert, J. Weston, A unified architecture for natural language processing: deep neural networks with multitask learning, in: Proceedings of the 25th International Conference on Machine Learning, ACM, Helsinki, Finland, 2008, pp. 160–167.
[9] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12 (2011) 2493–2537.
[10] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, et al., Recent advances in deep learning for speech research at microsoft, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Florence, Italy, 2013. pp. 8604–8608.
[11] 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: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE,Providence, RI, USA, 2012, pp. 1306–1313.
[12] K.G. Derpanis, R.P. Wildes, Dynamic texture recognition based on distributions of spacetime oriented structure, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, San Francisco, CA, USA, 2010, pp. 191–198.
[13] 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.
[14] G. Doretto, A. Chiuso, Y.N. Wu, S. Soatto, Dynamic textures, Int. J. Comput. Vis. 51 (2) (2003) 91–109.
[15] C. Feichtenhofer, A. Pinz, R.P. Wildes, Spacetime forests with complementary features for dynamic scene recognition, in: British Machine Vision Conference, 2013.
[16] C. Feichtenhofer, A. Pinz, R.P. Wildes, Bags of spacetime energies for dynamic scene recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Columbus, OH, USA, 2014.
[17] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, in: Computer Vision-ECCV 2014, 2014, Springer, Zurich, Switzerland, pp. 346–361.
[18] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[19] 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.
[20] Y. Jia, Caffe: an open source convolutional architecture for fast feature embedding, 2013, ⟨http://caffe.berkeleyvision.org⟩.
[21] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[22] K. Konda, R. Memisevic, V. Michalski, Learning to encode motion using spatio-temporal synchrony, in: International Conference on Learning Representations, 2013.
[23] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[24] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86(11) (1998) 2278–2324.
[25] D. Mandic, J. Chambers, Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures, and Stability, John Wiley & Sons, Inc., New York, NY, USA, 2001.
[26] M. Marszalek, I. Laptev, C. Schmid, Actions in context, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Miami, FL, USA, 2009, pp. 2929–2936.
[27] A. Mumtaz, E. Coviello, G. Lanckriet, A. Chan, A scalable and accurate descriptor for dynamic textures using bag of system trees, IEEE Trans. Pattern Anal. Mach. Intell. 37 (4) (2015) 697–712.
[28] 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.
[29] 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.
[30] F. Perronnin, J. Sánchez, T. Mensink, Improving the Fisher kernel for large-scale image classification, in: European Conference on Computer Vision, Springer, Heraklion, Crete, Greece, 2010, pp. 143–156.
[31] R. Péteri, S. Fazekas, M.J. Huiskes, Dyntex: A comprehensive database of dynamic textures, Pattern Recognit. Lett. 31 (12) (2010) 1627–1632.
[32] M. Pietikäinen, A. Hadid, G. Zhao, T. Ahonen, Computer Vision Using Local Binary Patterns, vol. 40, Springer, London, UK, 2011.
[33] X. Qi, R. Xiao, J. Guo, L. Zhang, Pairwise rotation invariant co-occurrence local binary pattern, in: European Conference on Computer Vision, Springer, Florence, Italy, 2012, pp. 158–171.
[34] E. Rahtu, J. Heikkilä, V. Ojansivu, T. Ahonen, Local phase quantization for blur-insensitive image analysis, Image Vis. Comput. 30 (8) (2012) 501–512.
[35] 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.
[36] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: integrated recognition, localization and detection using convolutional networks, 2013, arXiv preprint arXiv:1312.6229.
[37] A. Sharif Razavian, H. Azizpour, J. Sullivan, S. Carlsson, Cnn features off-the-shelf: an astounding baseline for recognition, 2014, arXiv preprint arXiv:1403.6382.
[38] N. Shroff, P. Turaga, R. Chellappa, Moving vistas: exploiting motion for describing scenes, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, San Francisco, CA, USA, 2010, pp. 1911–1918.
[39] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[40] Y. Sun, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, 2014, arXiv preprint arXiv:1406.4773.
[41] Y. Sun, X. Wang, X. Tang, Deep learning face representation from predicting 10,000 classes, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, 2014, arXiv preprint arXiv:1409.4842.
[43] C. Theriault, N. Thome, M. Cord, Dynamic scene classification: learning motion descriptors with slow features analysis, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Portland, OR, USA, 2013, pp. 2603–2610.
[44] L. van der Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (2579–2605) (2008) 85.
[45] A. Vedaldi, K. Lenc, Matconvnet-convolutional neural networks for matlab, 2014, arXiv preprint arXiv:1412.4564.
[46] Y. Xu, Y. Quan, H. Ling, H. Ji, Dynamic texture classification using dynamic fractal analysis, in: IEEE International Conference on International Conference on Computer Vision (ICCV), IEEE, Barcelona, Spain, 2011, pp. 1219–1226.
[47] 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.

Xianbiao Qi received his B.E. degree in 2008 and Ph.D. degree in 2015 from Beijing University of Posts and Telecommunications (BUPT). He visited the Web Search and Mining Group in Microsoft Research Asia (MSRA) from January 2011 to May 2012. Currently, he is a researcher in the Center of Machine Vision group in Oulu university of Finland. His research interests lie on computer vision, pattern recognition and medical image analysis. Specifically, he focuses on local feature design, texture and material classification, object recognition and medical image classification.



Chun-Guang Li received his B.E. degree in telecommunication engineering from Jilin University in 2002 and Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications (BUPT) in 2007. Currently he is a lecturer with the School of Information and Communication Engineering, BUPT, and as a member of Pattern Recognition and Intelligent System (PRIS) lab. He visited the Visual Computing group, Microsoft Research Asia, from July 2011 to April 2012. From December 2012 to November 2013, he visited the Vision, Dynamics and Learning Lab, Johns Hopkins University. His research interests are statistical machine learning, compressive sensing, and pattern recognition. He is a member of the IEEE, ACM, and CCF.

Guoying Zhao received the Ph.D. degree in computer science from the Chinese Academy of Sciences, Beijing, China, in 2005. From July 2005 to August 2010, she was a Senior Researcher with the Center for Machine Vision Research, University of Oulu, Oulu, Finland, where she has been an Adjunct Professor since September 2010 and Associate Professor since January 2014. She has authored over 110 papers in journals and conferences, and has served as a reviewer for many journals and conferences. Her research interests include gait analysis, dynamic texture recognition, facial-expression recognition, human motion analysis, and person identification. Dr. Zhao was a co-chair of the European Conference on Computer Vision 2008 Workshop on Machine Learning for Vision-based Motion Analysis (MLVMA) and the MLVMA workshop at the ICCV 2009 and the IEEE Conference on Computer Vision and Pattern Recognition 2011, ECCV 2014 workshop on Spontaneous Facial Behavior Analysis, and ACCV 2014 workshop on RoLoD: Robust Local Descriptors for Computer Vision.

Xiaopeng Hong received the B.Eng., M.Eng., and Ph.D. degrees in computer application from the Harbin Institute of Technology, Harbin, China, in 2004, 2007, and 2010, respectively. He has been a Scientist Researcher with the Center for Machine Vision Research, University of Oulu, since 2011. He has authored and co-authored more than 20 peer-reviewed articles in journals and conferences, and has served as a reviewer for several journals and conferences. His current research interests include pose and gaze estimation, texture classification, object detection and tracking, and visual speech recognition.

Matti Pietikäinen received his Doctor of Science in Technology degree from the University of Oulu, Finland. He is currently a Professor, Scientific Director of the Infotech Oulu, and Director of Center for Machine Vision Research at the University of Oulu. From 1980 to 1981 and from 1984 to 1985, he visited the Computer Vision Laboratory at the University of Maryland. He has made pioneering contributions, e.g., to Local Binary Pattern (LBP) methodology, texture-based image and video analysis, and facial image analysis. He has authored over 300 refereed papers in international journals, books, and conferences. His research is frequently cited, and its results are used in various applications around the world. Dr. Pietikäinen was Associate Editor of IEEE Transactions on Pattern Analysis and Machine Intelligence and Pattern Recognition journals, and currently serves as Associate Editor of Image and Vision Computing and IEEE Transactions on Forensics and Security journals. He was the President of the Pattern Recognition Society of Finland from 1989 to 1992, and was named its Honorary Member in 2014. From 1989 to 2007 he served as Member of the Governing Board of International Association for Pattern Recognition (IAPR), and became one of the founding fellows of the IAPR in 1994. He is an IEEE Fellow for contributions to texture and facial image analysis for machine vision. In 2014, his research on LBP-based face description was awarded the Koenderink Prize for Fundamental Contributions in Computer Vision.