Momental directional patterns for dynamic texture recognition

Thanh Tuan Nguyen, Thanh Phuong Nguyen, Frédéric Bouchara, Xuan Son Nguyen

Momental directional patterns for dynamic texture recognition

Thanh Tuan Nguyen
Université de Toulon, Aix Marseille Université, CNRS, LIS, Marseille, France
HCMC University of Technology and Education, Faculty of IT, HCM City, Vietnam

Thanh Phuong Nguyen
Université de Toulon, Aix Marseille Université, CNRS, LIS, Marseille, France
AI Lab, FIT, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Frédéric Bouchara
Université de Toulon, Aix Marseille Université, CNRS, LIS, Marseille, France

Xuan Son Nguyen
Université de Caen Basse-Normandie, CNRS, GREYC, UMR 6072, Caen, France

December 3, 2019

Abstract

Understanding the chaotic motions of dynamic textures (DTs) is a challenging problem of video representation for different tasks in computer vision. This paper presents a new approach for an efficient DT representation by addressing the following novel concepts. First, a model of moment volumes is introduced as an effective pre-processing technique for enriching the robust and discriminative information of dynamic voxels with low computational cost. Second, two important extensions of Local Derivative Pattern operator are proposed to improve its performance in capturing directional features. Third, we present a new framework, called Momental Directional Patterns, taking into account the advantages of filtering and local-feature-based approaches to form effective DT descriptors. Furthermore, motivated

by convolutional neural networks, the proposed framework is boosted by utilizing more global features extracted from max-pooling videos to improve the discrimination power of the descriptors. Our proposal is verified on benchmark datasets, i.e., UCLA, DynTex, and DynTex++, for DT classification issue. The experimental results substantiate the interest of our method.

1 Introduction

Dynamic textures (DTs) are textures repeated in a temporal domain, such as fountain, smoke, foliage, sea-wave, a blowing flag, fire, etc [1]. Efforts of analysis to make them more "understandable" are crucial for important tasks of recognition, segmentation, synthesis, and indexing for retrieval. Those are primary keys in a large range of applications in computer vision, such as visual surveillance of traffic scenes, crowded people [2], human interaction [3, 4, 5, 6], detecting objects and events [7, 8], tracking motion objects [9], etc. The major challenges in DT analysis are due to the wide range of appearances and non-directional motions of DTs. Many works for DT representation have been raised to deal with the problems by exploiting the advantages of spatio-temporal features and other properties of DTs. Roughly, those works can be categorized into six main groups: optical-flow-based, model-based, learning-based, filter-based, geometry-based, and local-feature-based.

First, *optical-flow-based methods*, which efficiently compute and encode videos in natural way, have been taken in remarkable consideration. To shape and trace the path of a motion in a sequence, [10] aggregated spatio-temporal textures formed by magnitudes and directions of the normal flow which are essential to identify motion types. [11] presented a qualitative approach based on the normal vector field and criteria of videos to describe DT features. In another work, these authors combined the normal flow with filtering regularity to capture the revealing properties of DTs [12]. In the meanwhile, [13] utilized the velocity and acceleration properties estimated by a structure tensor to form spatio-temporal multi-resolution histogram. Recently, [14] have proposed Features of Directional Trajectory (FDT) in accordance with Motion Angle Patterns (MAP) for addressing local characteristics and angle information of motion points which are along the paths of dense trajectories of a DT sequence. Due to [15], in the optical flow methods, assumption of brightness constancy and local smoothness is not suitable for

stochastic DTs in reality. Moreover, just motion features of DTs are encoded while their textures and appearances have not been regarded.

Second, *model-based methods* have recently attracted researchers in DT representation. [1] laid the foundation of those with a typical model of Linear Dynamical System (LDS) based on a combination of Hidden Markov Models (HMM) and Gaussian noise. Inspired by the idea of LDS, many works have taken it into account DT estimation for recognition tasks as well as for other problems in computer vision. [16] utilized *kernel*-PCA (Principal Component Analysis) to model LDS's observation component as a non-linear function to apprehend characteristics of dynamic features in complex motions, such as chaotic motions (e.g., turbulent water) and camera motions (e.g., panning, zooming, and rotations). Later, to capture the motions of objects in sequences, they presented a model of DT mixtures (DTMs) based on the LDS's concept. The outputs are then fed into an algorithm of hierarchical expectation-maximization (HEM-DTM) in order to categorize DTMs into k clusters for DT description [17]. Also based on the LDS model, [18] made it be in accordance with a bag-of-words (BoW) method to extract chaotic features in videos, while [19] based on bag-of-systems (BoS) to form the corresponding spatio-temporal patterns. To enhance the lookup speed of BoS's codebook, [20] proposed BoS Tree, in which a bottom-up hierarchy is constructed for indexing the codewords. In terms of efficiency, the model-based methods have achieved the modest results on DT recognition because their major drawback is that their encoding mostly concentrates on spatial-appearance-based characteristics of DTs rather than dynamic-based [1]. Furthermore, efforts taking them into account dynamic features have made the models more complex [19].

Third, *learning-based methods* have been growing into potential approaches as their noteworthy estimations in DT recognition. A well-trained Convolutional Neural Network (CNN) has considerable attention for learning DT characteristics. [21] adopted CNN with the concept of deep structures in still images to learn Transferred ConvNet Features (TCoF) for DT classification. CNN is also utilized in [22] to extract DT features (DT-CNN) from three orthogonal planes of sequences, while [23] took it into account constructing a multi-layer convolutional architecture (PCANet-TOP) in which the PCA procedure is involved with three orthogonal planes of a DT video for learning filters. Lately, a deep dual descriptor [24] is based on characteristics of "key frames" and "key segments" to learn static and dynamic features. In addition, techniques which are based on dictionary learning with kernel sparse

coding to extract local DT features have obtained promising evaluations in DT recognition. In [25], each video is partitioned into patches, known as atoms, in order that local DT features are pointed out using a dictionary learned by the sparse coding method with the input of atoms. However, it is difficult for this work to perform in multi-scale configuration because of the constraint of the atoms in the identical dimension. On the other side, [26] introduced equiangular kernel to learn a dictionary with reasonable size. Although the learning-based approaches have usually outperformed in comparison with others not only in DT representation but also in other tasks of computer vision, they take a long time to encode a huge vector of features using complicated learning algorithms. Our proposed framework below can achieve competitive recognition results with a simple operator and less cost of time for computation.

Fourth, *filter-based methods* have evinced their efficiency in performance of DT recognition. [27] extracted Binarized Statistical Image Features using filtering operations on various spatio-temporal regions and binarizing the filter responses. These filters are learned by employing Independent Component Analysis (ICA) transformation on Three Orthogonal Planes (BSIF-TOP). Then its multi-resolution scheme (MBSIF-TOP) is also introduced to enhance the capacity of DT depiction. [15] presented spatio-temporal Directional Number transitional Graph (DNG) as a dynamic-micro-texture descriptor in which DT appearance and motion features are encoded by capturing directions of natural flow along temporal domains. Experiments illustrate that the filter-based approaches have performed well on DT datasets with simple motions (e.g., UCLA). They, however, either remain several limitations or have not been verified on challenging datasets (e.g., DynTex, DynTex++). In addition, it takes a significant time to learn filters in BSIF-TOP or to divide a sequence into 3D grids for DNG, these constraints can raise the computational complexity.

Fifth, *geometry-based methods* encode DTs based on fractal techniques to tolerate environmental changes of videos. A typical procedure of those is named Dynamic Fractal Spectrum (DFS) [28], (then extended to Multi-Fractal Spectrum (MFS) in [29]), in which DT features are figured out by a combination of capturing stochastic self-similarities and analyzing fractal patterns of DT sequences. However, only spectral information is considered in those, regardless spatial domain. [30] addressed this issue by embedding spatial appearance analysis into MFS in accordance with wavelet coefficients to form Wavelet-based MFS (WMFS) DT representation with more robust-

ness. Lately, [31] presented Spatio-Temporal Lacunarity Spectrum (STLS) descriptor with lacunarity-based features which are captured from lacunarity analysis on local binary patterns in DT slices. Another effort [32] based on Stationary Subspace Analysis (SSA) to extract the stationary components across multiple videos of the same class and then encoded them in a feature vector with lower dimension. It can be seen that the geometry-based methods just perform well on a few particular DT datasets, not on the whole DT benchmarks. In addition, some of them have not been exploited the temporal part, the important information for DT description.

Finally, *local-feature-based methods* have taken into account local features for DT representation. Most of them are derived from the variants of Local Binary Patterns (LBP), adapted to deal with DT videos, to take advantage of their simplicity and efficiency in computation for carrying out various issues of DT representation. [4] proposed two LBP-based variants to encode videos: Volume-LBP (VLBP) operator acquires spatio-temporal characteristics by considering three consecutive frames of DT sequences, while the other addresses LBP operator on Three Orthogonal Planes (LBP-TOP). Inspired by those works, several efforts extended them to treat the LBP's conventional shortcomings in order to enhance the performance of DT representation: rotation-invariant [33], complementary components and problems of sensitivity to noise [34, 35, 36, 37]. In other ways, [38] focused on feature vector reduction through a technique of learning data-driven LBP (DDLBP) structures. To eliminate misleading patterns in encoding LBP features, [39] then presented Principal Histogram Analysis (PHA) in which PCA is exploited to improve the reliability of LBP histograms. This method, however, only concentrates on the appearances of DT and has high-dimensional feature space (i.e. $256 \times 256 \times$ number of patches).

Beside their promising results, the local-feature-based methods remain several essential issues, such as large dimension [4, 38, 34], sensitivity to noise, near uniform regions [35, 36]. To address these problems, we present in this paper a new framework for DT representation based on three stages: At first, a new model of r-order moment volumes is introduced by considering the local neighbors of each dynamic voxel sampled by an element of structuring volume. Second, an extended operator of Local Derivative Patterns (LDPs) is proposed by integrating two major extensions to improve the performance of the typical LDP operator. Third, we present a framework for integration of these components in order to produce effective DT descriptors, named Momental Directional Patterns (MDP). To verify our works, we have

experimented on benchmark DT datasets (UCLA, DynTex, and DynTex++) for the recognition task. Evaluations show that our framework outperforms significantly compared to the existing approaches. Consequently, the major contributions of this paper can be listed as follows.

- A new model of moment volumes extracting from different orders is proposed in order to make the descriptor more robust against noise and illumination sensitivity, near-uniform regions. This operation enriches the robust and discriminative information of dynamic voxels with low time cost of computation and is also regarded as a pre-processing step of our framework.

- A novel concept of completed second-order LDP operator, which allows to exploit more efficiently different local higher derivative variations (sign, magnitudes, etc.) to enhance discrimination power by addressing three complementary components, is introduced.

- Adaptive directional thresholds for the components are also proposed to mitigate the near uniform image problem. They are then taken into account the above completed model to construct an extended version of LDP operator making it even more robust and discriminative.

- A new framework for DT description, called Momental Directional Patterns (MDP), is formulated by integrating the above complementary components in several ways. It allows to obtain both shape and motion cues of directional DT features through filtered videos figured out by the model of r-order moment volumes.

- Inspired from CNN architecture, the max-pooling operator is addressed in the encoding of MDP descriptor to improve its performance by capturing more global features.

The rest of this paper is structured as follows. Section 2 recalls some related works of LBP-based variants in still images and DT description as well. The proposed model of r-order moment volumes is presented in Section 3. The next section is to introduce some crucial extensions of LDP operator for capturing local features of textured images. Then DT descriptors based on the above components are detailed in Portion 5. Section 6 addresses max-pooling properties to boost the performance. Section 7 presents settings for experiments and evaluations of the proposed approach compared to the

state-of-the-art results on various benchmark DT datasets. The last section states conclusions and several suggestions for the future works.

2 Related work

As mentioned in the previous section, image texture description based on the LBP operator has significantly obtained outperformance of recognition issue thanks to its simple calculation and efficient discrimination power. In this section, we briefly expose the performance of LBP and its variants on encoding features in still images and dynamic textures.

2.1 A brief review of LBP

In order to structure spatial features of a local image texture as binary codes, [40] proposed Local Binary Patterns (LBP), in which a center pixel $\mathbf{q}_c$ is involved with its neighbors. Let $\mathcal{I}$ be a 2D gray-scale image, LBP code of $\mathbf{q}_c$ is formed as follows.

$$\mathrm{LBP}_{P,R}(\mathbf{q}_c) = \sum_{i=0}^{P-1} s\big(\mathcal{I}(\mathbf{p}_i) - \mathcal{I}(\mathbf{q}_c)\big)2^i \tag{1}$$

where $\mathcal{I}(.)$ means gray-scale level of a pixel, $\{\mathbf{p}_i\}$ denotes P interpolated neighbors on a circle of the center $\mathbf{q}_c$ with radius R, and function s(.) is defined as

$$s(x) = \begin{cases} 1, x \geq 0 \\ 0, \text{otherwise.} \end{cases} \tag{2}$$

Accordingly, this basic encoding leads to the fact that up to 2^P discrete values are utilized to depict an image texture. To address this problem, three popular mapping techniques are usually employed in practice as follows to decrease the dimension of descriptor. First, based on the fact that most of the LBP patterns in natural images are uniform, [40] proposed $u2$ mapping to capture uniform patterns (LBP^{u2}), those which have number of bitwise transitions (1-0 or 0-1) in their binary string at most two. The uniform descriptor then has $P(P-1)+3$ distinct values including $P(P-1)+2$ of uniform patterns and one of that all non-uniform patterns are grouped into. Second, rotation invariant (ri) mapping is also introduced in [40] to make

the LBP patterns more resistant to image rotation as follows.

$$\mathrm{LBP}^{ri}_{P,R}(\mathbf{q}_c) = \min_{0 \le i < P} \{ROR(\mathrm{LBP}_{P,R}(\mathbf{q}_c), i)\} \quad (3)$$

where $ROR(\mathrm{LBP}, i)$ computes the distribution of the right circular bitwise shift of i bits on the LBP binary chain. In case of that, uniform patterns LBP^{u2} are taken into account ri mapping, the third mapping is formed, named $riu2$, to structure rotation invariant uniform descriptor (LBP^{riu2}) with a much lower dimension of $P + 2$ compared to 2^P of the basic LBP. Thereafter, other important mappings are also presented to enhance the performance of representation, such as Local Binary Count (LBC) [41] - an alternative of uniform patterns, extra uniform patterns [42] for taking advantage of useful non-uniform patterns, $TAP^{\mathcal{A}}$ mapping [43] for acquiring topological information.

2.2 LBP-based variants in still images

The typical LBP operator has been prominently taken into account diverse applications in computer vision owing to its imposing performance with simplicity in computation and implementation as well. However, it remains several internal restrictions, such as small supporting regions, lack of local and global textural information, and noise sensitivities. A lot of efforts have then addressed these shortcomings in order to improve the LBP's execution.

[44] proposed a completed LBP (CLBP) model by adopting local variations of magnitudes which include useful information of local textural patterns. More specifically, CLBP consists of three complementary components: CLBP_S that is identical to the basic LBP, CLBP_M for acquiring local variations of magnitudes, and CLBP_C for measuring the difference between gray-scale level of each center pixel and the global one of a texture image. Integrating these complementary components in different ways can significantly ameliorate the performance. One of the most favorite combinations is that probability distributions of those are joined to form a descriptor with more robust discrimination. Furthermore, [45] took variance into account LBP-based encoding as a regional contrast estimation to exploit valuable information which is not considered in the typical LBP model. [46] then advanced this idea to investigate various order moments as LBP-based filters to capture Statistical Binary Patterns (SBP).

Other proposals have attempted to handle the inherent restrictions of LBP in several ways, such as multi-scale analysis [47], pattern encoding and

selecting [48, 49, 50], feature training [51], preprocessing [52, 46], thresholding [53, 54], mapping issues [42, 43], etc.

2.3 LBP-based variants for dynamic textures

Inspired by the leverage of LBP-based variants in still images, several attempts have taken them into account dynamic texture processing. [4] proposed VLBP in which $3P$ neighbors are located on three circles of center voxels with the same coordinate from three consecutive frames of DTs. The center voxel at the middle frame is then binarized by exploiting the typical LBP operator for these neighbors and two other centers of the first and last frames. As a result of that, VLBP binary codes are formed with length of $3P + 2$, leading to the descriptor with large dimension of 2^{3P+2} bins. To treat this problem, [4] introduced LBP-TOP in which LBP operator is applied for a center voxel in consideration of Three Orthogonal Planes (TOP) of a sequence to capture spatial structures on XY plane and motion cues on XT, YT planes. The final descriptor of DT sequence is then shaped by concatenating the histograms calculated on these corresponding planes, i.e., $[\mathrm{LBP}_{XY}, \mathrm{LBP}_{XT}, \mathrm{LBP}_{YT}]$ with dimension of 3×2^P bins.

Thereafter, most of works for DT representation are stated as variants of above approaches in order to enhance the discriminative power of DT descriptor, such as merging CLBP into VLBP to form CVLBP framework [34], taking advantage properties of both LBP-TOP and VLBP to procedure Helix Local Binary Patterns (HLBP) [35], adopting adaptative thresholds to encode Local Structure Patterns on Three Orthogonal Planes (LSP-TOP) [36].

3 Moment models

Taking into account the advantages of filter-bank approaches, we propose in this section a new concept of moment volumes as a filtering technique in which different order moments of a sequence are calculated from a pre-defined element of supporting volume. In our framework, this operation is regarded as a preprocessing with a low cost of computation to enrich robustness and discrimination for local DT features.

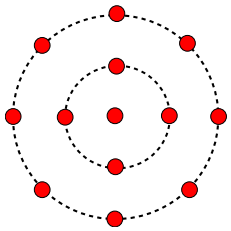

Figure 1: A sample of structuring element with {(P,R)}={(4,1),(8,2)} [46].

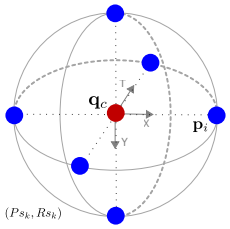

Figure 2: (Best viewed in color) A pattern of volume support $\Omega = \{(6, 1)\}$ which has $P_k = 6$ blue neighbors sampled on a sphere with the center of red point and radius $R_k = 1$.

3.1 Moment images

In [46], the authors presented a model of moment images, also known as a pre-processing step of image texture classification, in which still images are filtered by exploiting a LBP-based filter with a pre-defined supporting region. Encoding based on the filtered images points out local relationships with more stable textural structures against changes of environment. Two types of local statistical moments are produced as follows. First, the r-order moment image calculates the statistic distribution around a pixel $\mathbf{q}_c$ as

$$m^r_{(\mathcal{I},B)}(\mathbf{q}_c) = \frac{1}{|B|} \sum_{\mathbf{p}_i \in B} \Big(\mathcal{I}(\mathbf{q}_c + \mathbf{p}_i)\Big)^r \tag{4}$$

in which $\mathcal{I}$ means a 2D gray-scale image texture, $\mathbf{q}_c$ is a center pixel (i.e., $\mathbf{q}_c \in \mathcal{I}$), B is a supporting regional element consisting of points sampled by one or more concentric circles of the center $\mathbf{q}_c$ with different radii R, i.e., $\{(P, R)\}$ (see Fig. 1), $|B|$ is the cardinality of B.
Second, the r-order centered moment image ($r > 1$) defines the statistic distribution around a pixel $\mathbf{q}_c$ as follows.

$$\mu^r_{(\mathcal{I},B)}(\mathbf{q}_c) = \frac{1}{|B|} \sum_{\mathbf{p}_i \in B} \Big(\mathcal{I}(\mathbf{q}_c + \mathbf{p}_i) - m^1_{(\mathcal{I},B)}(\mathbf{q}_c)\Big)^r \tag{5}$$

where $m^1_{(\mathcal{I},B)}(\mathbf{q}_c)$ denotes the mean value (1-order moment) formed around pixel $\mathbf{q}_c$.

[46] have also shown that working on a series of moment images of different orders brings more textural information because the regional gray distribution is better captured using different statistical moments.

3.2 Moment volumes

The model of moment images has just considered spatial relations of a center pixel with its neighbors for image texture classification. To be in accordance with video representation, we hereafter propose a new local statistical model, called moment volumes as an extension of moment images, based on statistical moments calculated from a pre-defined spherical support. Similar to [46], our idea is motivated from filter-bank approaches to exploit more rich and robust information of shape and motion cues of DT videos by addressing different statistic distributions.

Let $\mathcal{V}$ denote a 3D gray-scale level of a video and $\mathbf{q}_c$ an arbitrary voxel of $\mathcal{V}$. Let $\Omega = \{S_1, S_2, ..., S_n\}$ be a local supporting volume as union of discrete spheres, centered at the same spatial coordinate, for calculating the statistic distributions at each voxel of $\mathcal{V}$. Each single spheric structuring support $S_k = (P_k, R_k)$ expresses that P_k neighbors are located on a sphere with radius R_k. In order to compute local statistic distribution at voxel $\mathbf{q}_c$, it is simply to settle the center of Ω at $\mathbf{q}_c$ and then to determine its neighbors defined by Ω. To simplify the presentation, we adopt hereafter an assumption that coordinate of $\mathbf{q}_c$ is $(0, 0, 0)$, it is possible to situate P_k neighbors on the sphere S_k in two following configurations:

- First, six points are placed on the endings of its orthogonal diameters, i.e., $\{(0, 0, R_k),\ (0, 0, -R_k),\ (-R_k, 0, 0),\ (R_k, 0, 0),\ (0, -R_k, 0),\ (0, R_k, 0)\}$.
- Second, in addition to the above set, this also consists of eight radial points. Each of which is located on the center of each one-eighth of the sphere S_k, i.e., its coordinate can be referred as one of different instances of $\big(\pm R_k/\sqrt{3}, \pm R_k/\sqrt{3}, \pm R_k/\sqrt{3}\big)$. As the result, there are 14 neighbors in this configuration which can be considered for each supporting volume.

A sample of $S_k = (6, 1)$ for the center $\mathbf{q}_c$ can be formed by $P_k = 6$ local neighbors on a sphere of $R_k = 1$ as graphically illustrated in Fig. 2. On the other hand, a local supporting volume may be unions of different discrete

spheres. For example, $\Omega = \{(6,1),(14,2)\}$ consists of two spheric structuring supports.

Given a pre-defined supporting volume Ω, we propose to consider two following statistic distributions. Firstly, the r-order moment volume of association between video $\mathcal{V}$ and the local supporting volume Ω is defined as follows.

$$m^r_{\mathcal{V},\Omega}(\mathbf{q}_c) = \frac{1}{|\Omega|}\sum_{i=1}^{|\Omega|}\Big(g(\mathbf{q}_c + \mathbf{p}_i)\Big)^r \tag{6}$$

in which $\mathbf{q}_c \in \mathcal{V}$ is the current voxel with its surrounding neighbors $\mathbf{p}_i \in \Omega$, the volume element Ω can be structured by one or more spheres with the same center voxel and different radii, $|\Omega|$ is the total of considered neighbors, function $g(.)$ returns the gray level value of a voxel. Secondly, the r-order centered moment volume ($r > 1$) can also be defined as

$$\mu^r_{\mathcal{V},\Omega}(\mathbf{q}_c) = \frac{1}{|\Omega|}\sum_{i=1}^{|\Omega|}\Big(g(\mathbf{q}_c + \mathbf{p}_i) - m^1_{\mathcal{V},\Omega}(\mathbf{q}_c)\Big)^r \tag{7}$$

where $m^1_{\mathcal{V},\Omega}(\mathbf{q}_c)$ is the 1-order moment volume at the voxel $\mathbf{q}_c$. In practice, our model particularly considers two following types of moment volumes: the mean m^1 and the variance μ^2 that are complementary and exploit well shape and motion cues of DT videos.

3.3 Advantages of moment volume model

By addressing different statistic distributions calculated from a pre-defined structuring volume, the proposed model of moment volumes has several following beneficial properties.

- *Insensitivity to noise:* Considering local statistic distributions (mean and variance) calculated from neighbors allows moment volumes to be more robust against noise than the raw video because the proposed model works like a low-pass filter which is able to eliminate dynamic voxels with intensely high frequency corresponding to noise.

- *Invariance to rotation:* Our model is independent on angle changes of frames in DT sequences because the pre-defined supporting region for calculating volume of moments is a union of discrete spheres, which is isotropic and so on discards all orientation information. Therefore, the moment volumes are invariant against rotation.

- *Information richness:* The concept of moment volume, which exploits textural information about local structures, allows to capture more global information. In addition, taking into account the advantages of filter-bank methods, our model permits to obtain more numerous types of local structures by using various moment distributions with different elements of the structuring volume. In practice, two order moments "mean" and "variance" are complementary, so these convey richer information than the original video.

- *Low computational cost:* Concerning the computational complexity, as filtered sequences are calculated on a pre-defined structuring volume, their calculation is simple and efficient along with the same computing cost like the typical LBP operator. Our algorithm in raw MATLAB code runs impressively fast on a Linux laptop of CPU Intel Core i7 1.9 Ghz with 4G RAM. It just takes less than 0.11s to handle a video with dimension of $48 \times 48 \times 75$ for a 3D spherical supporting volume of $\Omega = \{(6, 1)\}$ (see Fig. 2).

4 Some crucial extensions of Local Derivative Patterns

The typical LDP operator has been initially proposed for face recognition [55] by exploiting local derivative direction variations and then successfully applied to other applications, such as action recognition [3]. We adopt in this work for the first time this operator to capture shape and motion cues for DT representation. Moreover, we also propose hereafter two following important extensions of LDP operator to improve its discriminative power: adaptative directional thresholds and a completed model of LDP.

4.1 Local Derivative Patterns

[55] introduced Local Derivative Patterns (LDPs), a directional extension of LBP, by taking into account local high-order derivative variations based on considering a pixel and its neighbors in different directions to capture more robust features.

The first-order LDP at a pixel for a set of considered directions $\mathcal{D}$ is

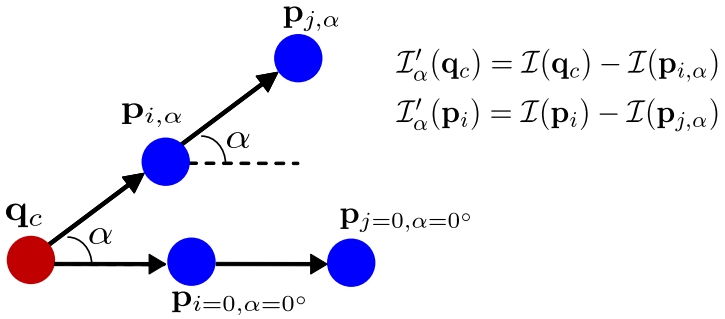

Figure 3: (Best viewed in color) Model of the first-order LDP patterns of $\mathbf{q}_c$ ($\mathcal{I}'_\alpha(\mathbf{q}_c)$) and $\mathbf{p}_i$ ($\mathcal{I}'_\alpha(\mathbf{p}_i)$) pixels in directions $\alpha \in \mathcal{D}$ in which $\mathbf{q}_c$ (in red) is the considered point, $\mathbf{p}_i$ is the i^{th} neighbor of $\mathbf{q}_c$, and $\mathbf{p}_j$ is the j^{th} neighbor of $\mathbf{p}_i$.

defined as follows.

$$\mathcal{I}'_\alpha(\mathbf{q}_c) = \mathcal{I}(\mathbf{q}_c) - \mathcal{I}(\mathbf{p}_{i,\alpha}) \tag{8}$$

where $\mathbf{p}_{i,\alpha}$ is the i^{th} neighbor of a center point $\mathbf{q}_c$ in a concerning direction α, $\mathcal{I}(.)$ is gray-scale image level of a pixel. Fig. 3 graphically illustrates the regular computation of the first-order LDP patterns corresponding to directions $\alpha \in \mathcal{D}$.

In general, the n^{th}-order LDP is defined as follows, for the center pixel $\mathbf{q}_c$ and its P neighbors circled with radius R.

$$\text{LDP}^n_{P,R,\alpha}(\mathbf{q}_c) = \left\{ f\left(I^{n-1}_\alpha(\mathbf{q}_c), I^{n-1}_\alpha(\mathbf{p}_i)\right) \right\}_{1 \le i \le P} \tag{9}$$

where $I^{n-1}_\alpha(.)$ means the $(n-1)^{th}$-order derivative in direction α at a pixel, $\mathbf{p}_i$ is the i^{th} neighbor of the center point $\mathbf{q}_c$, and function $f(.)$ is defined as

$$f(x, y) = \begin{cases} 1, & \text{if } x * y \le 0 \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

The detail of other LDP's formulations as well as samples of its calculation is discussed in [55]. In practice, four directions are often considered, i.e., $\alpha \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$, to capture directional mutual relations of pixels [3, 55]. In case of inspecting the first-order derivative variations in all of directions, LDP is simply identical to the basic LBP.

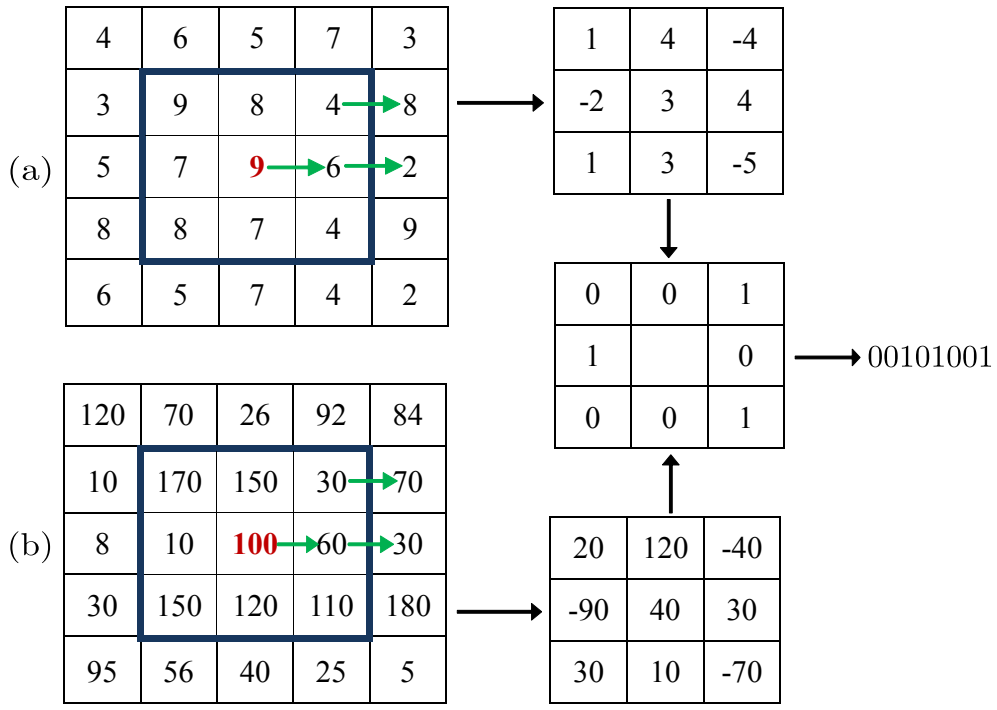

Figure 4: (Best viewed in color) An example of two different local structures (marked in red color) are encoded by the same LDP pattern in concerned direction $\alpha = 0°$.

4.2 Adaptative directional thresholds

Similar to a well-known restriction of the typical LBP, LDP is not occasionally able to judge different structure patterns because its encoding is still thresholded by the center with around neighbors. It can be observed in Fig. 4 that two different local structures (a) and (b) are figured out by the same pattern. In order to handle this issue, we propose to define three following adaptative thresholds[1] for LDP operator. The key idea for that is the consideration of the first-order LDP. These thresholds will be then exploited in Section 4.3 to construct the completed model of LDP.

First, Global Directional Difference (GDD) of an image texture is calculated as the mean of absolute directional differences on the entire of concerned

[1]Contrary to two last thresholds, the first one is empirically proposed without depending on α because this leads to more robust and stable results.

directions.

$$\mathrm{GDD}(\mathcal{I}) = \frac{1}{|\mathcal{D}| * \mathcal{N}} \Big(\sum_{\mathbf{q}_j \in \mathcal{I}} |\mathcal{I}'_\alpha(\mathbf{q}_j)| \Big) \Big|_{\alpha \in \mathcal{D}} \tag{11}$$

where $\mathcal{N} = (\mathcal{W} - 2) * (\mathcal{H} - 2)$, $\mathcal{W}$ and $\mathcal{H}$ are width and height dimensions of 2D image $\mathcal{I}$ respectively, $|\mathcal{D}|$ is the total of considered directions, $\mathcal{I}'_\alpha(.)$ is the first-order local derivative pattern of a pixel in regarding direction α.

Second, to capture the information of Directional Magnitudes (DM_α) for each direction α, we compute the mean of absolute multiplication of directional differences on the whole image as follows.

$$\mathrm{DM}_\alpha(\mathcal{I}) = \frac{1}{\mathcal{N} * P} \Big(\sum_{i=0}^{P-1} |\mathcal{I}'_\alpha(\mathbf{q}_j) * \mathcal{I}'_\alpha(\mathbf{p}_i)| \Big) \Big|_{\mathbf{q}_j \in \mathcal{I}} \tag{12}$$

in which $\mathbf{p}_i$ is the i^{th} neighbor of current pixel $\mathbf{q}_j$ of image $\mathcal{I}$, P is the number of considered neighbors.

Third, the Directional Center (DC) threshold is defined as the average of directional centered differences on the whole image.

$$\mathrm{DC}_\alpha(\mathcal{I}) = \frac{1}{\mathcal{N}} \sum_{\mathbf{q}_j \in \mathcal{I}} \big|\mathcal{I}'_\alpha(\mathbf{q}_j)\big| \tag{13}$$

4.3 Completed model of LDP

[44] showed that considering local variations of magnitudes together with the typical LBP makes the descriptor more robust and discriminant because they are complementary. Inspired by this idea, we introduce in this portion, a completed model of the second order LDP using adaptative thresholds, which are presented in Section 4.2. Similar to [44], it also consists of three following complementary components.

First, we propose LDP-D operator as the first component in order to capture the second-order local derivative patterns adjusted by an adaptive thresholding GDD (see Equation (11)) as

$$\text{LDP-D}_{P,R,\alpha}(\mathbf{q}_c) = \sum_{i=0}^{P-1} \psi\big(\mathcal{I}'_\alpha(\mathbf{q}_c), \mathcal{I}'_\alpha(\mathbf{p}_i), \mathrm{GDD}(\mathcal{I})\big) 2^i \tag{14}$$

where $\mathbf{p}_i$ is the i^{th} neighbor of the center pixel $\mathbf{q}_c$ in accordance with direction α, P is number of considered neighbors circled by radius R, and function $\psi(.)$

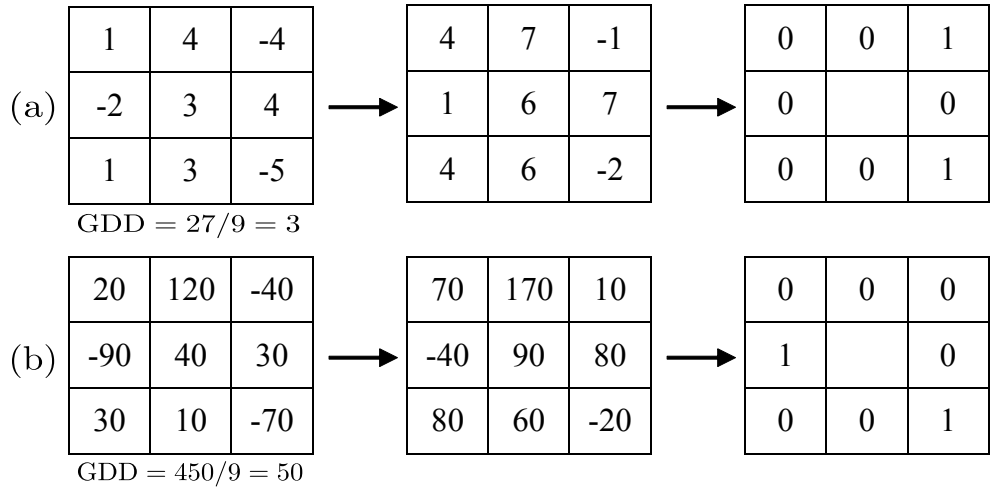

Figure 5: Two different local structures (a) and (b) in Fig. 4 are encoded by different LDP-D patterns in direction $\alpha = 0°$.

is estimated as

$$\psi(x, y, z) = \begin{cases} 1, & \text{if } (x + z) * (y + z) \leq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

In assumption of just considering one direction of $\alpha = 0°$ (i.e., $|\mathcal{D}| = 1$), in contrast to the basic LDP, the proposed operator LDP-D is able to differentiate the local structures (a) and (b) as detailed in Fig. 5.
Second, LDP-M component exploits the information of magnitudes in a direction α by using adaptative threshold DM_α (see Equation (12)) and is formed as

$$\text{LDP-M}_{P,R,\alpha}(\mathbf{q}_c) = \sum_{i=0}^{P-1} h\big(\mathcal{I}'_\alpha(\mathbf{q}_c), \mathcal{I}'_\alpha(\mathbf{p}_i), \mathrm{DM}_\alpha(\mathcal{I})\big) 2^i \tag{16}$$

where $h(.)$ is defined as

$$h(x, y, z) = \begin{cases} 1, & \text{if } |x * y| \geq z \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

Third, LDP-C regards to the directional contrast of a center against the mean of directional differences on the whole image.

$$\text{LDP-C}_\alpha(\mathbf{q}_c) = s\big(\mathcal{I}'_\alpha(\mathbf{q}_c) - \mathrm{DC}_\alpha(\mathcal{I})\big) \tag{18}$$

in which $s(.)$ is defined by Equation (2).

Three above complements (abbreviated to LDP_D, LDP_M, and LDP_C) should be combined in different ways to produce extended LDP operator, named xLDP, for investigation to find out an enhanced operator LDP for encoding DT features. An instance of those is $\text{xLDP} = \text{LDP}_{D\_M/C}$, in which the signals of "$\_$" and "$/$" in the style $_{D\_M/C}$ mean that histograms obtained by the corresponding components are concatenated and jointed respectively. It should be noted that our operator can be also generated in high-order derivative patterns (xLDP^n) by exploiting the n^{th}-order directional LDPs ($n > 2$) [55] for calculation of the proposed components above.

Our xLDP operator is different from the typical LDP [55] in several properties to enhance the performance:

- The xLDP operator considers local structures in diversity of directional relations based on three complemented components, in contrast to LDP with only in consideration of local derivative patterns.

- Our operator is more insensitive to noise when exploiting adaptative directional thresholds (see an instance of encoding patterns in Fig. 4 and Fig. 5).

- To encode a local structure in each direction, LDPs are separately computed by using the corresponding components. In the meanwhile, the basic LDP encodes a pixel in a long binary chain for all concerned directions, e.g., a string of 32 bits for four 8-bit LDPs.

- Thanks to structuring patterns in separative strings of binary codes, two popular mappings of $riu2$ and $u2$ for the processing of description can be utilized to advance the performance of descriptor with practical dimension. In contrast, LDPs are calculated on sub-regions of an image texture with various parameters of histogram bins.

4.4 Assessing our proposed extensions of LDP

In order to evaluate the proposed complementary components for LDP operator, we also implement the basic LDP [55] for DT description based on the filtered videos captured by the proposed model of r-order moment volumes. For a center pixel $\mathbf{q}_c$ and its P considered neighbors sampled by a circle with

radius R, the second-order typical local derivative pattern (LDP) of $\mathbf{q}_c$ in direction α, named $\mathrm{LDP}_{P,R,\alpha}$, is defined as

$$\mathrm{LDP}_{P,R,\alpha}(\mathbf{q}_c) = \sum_{i=0}^{P-1} f\big(\mathcal{I}'_\alpha(\mathbf{q}_c), \mathcal{I}'_\alpha(\mathbf{p}_i)\big) 2^i \tag{19}$$

where the function $f(.)$ is defined by Equation (10). Actually, this operator is the same LDP-D without exploiting the adaptative threshold of GDD.

5 Momental Directional Patterns for DT representation

In this section, we propose a new operator, named Momental Directional Patterns (MDP), to efficiently capture directional DT patterns from filtered videos obtained by the r-order moment volume model. Our idea is to take into account the advantages of filter bank approaches and a complementary LBP-based variant allowing to obtain more textural information in DT videos. We then consider our extended xLDP operator, presented in Section 4, on a series of moment volumes which are introduced in Section 3 to result in Momental Directional Patterns for DT representation. Let us recall that the extended operator xLDP is introduced to work in still images. For that reason, in order to take it into account describing shape and motion cues of a DT video, we adopt the idea of [4] to address xLDP on three orthogonal planes of moment volumes.

Let $\mathcal{V}$ denote a video and $\mathcal{D}$ be a set of considered directions. The r-order moment volumes with supporting region Ω are utilized to point out filtered sequences, i.e., mean (m^r) and variance (μ^r) videos. DT characteristics in each of these are then encoded by exploiting the proposed operator xLDP with directions $\alpha \in \mathcal{D}$ on three orthogonal planes XY, XT, YT of these moment volumes to compute the corresponding probability distributions, as graphically demonstrated in Fig. 6. The obtained histograms are concatenated and normalized to form the final descriptor of video $\mathcal{V}$ as follows.

$$\begin{aligned}\mathrm{MDP}_{\Omega,\mathcal{D}}(\mathcal{V}) = \big[&\mathrm{xLDP}_{P,R,\mathcal{D}}({m^r}_{XY}), \mathrm{xLDP}_{P,R,\mathcal{D}}({m^r}_{YT}),\\ &\mathrm{xLDP}_{P,R,\mathcal{D}}({m^r}_{XT}), \mathrm{xLDP}_{P,R,\mathcal{D}}(\mu^r_{XY}),\\ &\mathrm{xLDP}_{P,R,\mathcal{D}}(\mu^r_{XT}), \mathrm{xLDP}_{P,R,\mathcal{D}}(\mu^r_{YT})\big]\end{aligned} \tag{20}$$

From now on, we use the combination way of the extended xLDP operator to denote the corresponding descriptor MDP. For example, $\text{MDP}_{D\text{-}M/C}$ means that it is based on the extended operator xLDP = LDP_D-$\text{LDP}_M/\text{LDP}_C$, which is the concatenation between LDP_D and the joint of two components LDP_M and LDP_C.

In order to evaluate the contribution of the proposed extensions of LDP operator, a basic descriptor MDP-B, which is based on the second-order LDPs, is also considered by using the similar construction.

$$\begin{aligned}\text{MDP-B}_{\Omega,\mathcal{D}}(\mathcal{V}) = \big[&\text{LDP}_{P,R,\mathcal{D}}({m^r}_{XY}), \text{LDP}_{P,R,\mathcal{D}}({m^r}_{YT}),\\ &\text{LDP}_{P,R,\mathcal{D}}({m^r}_{XT}), \text{LDP}_{P,R,\mathcal{D}}(\mu^r_{XY}),\\ &\text{LDP}_{P,R,\mathcal{D}}(\mu^r_{XT}), \text{LDP}_{P,R,\mathcal{D}}(\mu^r_{YT})\big]\end{aligned} \tag{21}$$

On the other hand, to verify the eminent contribution of our model of moment volumes, we also structure LDP-TOP patterns to depict the original DT sequence $\mathcal{V}$ with non-supporting volume elements. These patterns are encoded by the typical second-order LDP operator on three orthogonal planes.

$$\begin{aligned}\text{LDP-TOP}_{\mathcal{D}}(\mathcal{V}) = \big[&\text{LDP}_{P,R,\mathcal{D}}(\mathcal{V}_{XY}), \text{LDP}_{P,R,\mathcal{D}}(\mathcal{V}_{XT}),\\ &\text{LDP}_{P,R,\mathcal{D}}(\mathcal{V}_{YT})\big]\end{aligned} \tag{22}$$

Two possible mappings can be taken into account encoding DT features in order to reduce the dimension of representation: $riu2$ and $u2$ giving $L_{riu2} = (P+2)$ and $L_{u2} = \big(P(P-1)+3\big)$ distinct values for each pixel pattern respectively, in which P is the considered neighbors. Particularly, the size of MDP descriptor depends on the combination ways of complemented components to form xLDP. For instance, descriptor $\text{MDP}_{D\text{-}M/C}$, computed by a style of xLDP = LDP_D-$\text{LDP}_M/\text{LDP}_C$ with $3 \times |\mathcal{D}| \times L_{riu2/u2}$ bins, has dimension of $9 \times |\mathcal{D}| \times L_{riu2/u2}\big(|m^r| + |\mu^r|\big)$ for $riu2$ and $u2$ mappings. Therein, $|\mathcal{D}|$ denotes the number of concerned directions. $|m^r|$ and $|\mu^r|$ explain the quantity of "mean" and "variance" videos filtered by the r-order moment volume model. Towards the MDP-B and LDP-TOP descriptors, their dimensions are respectively fixed as $3 \times |\mathcal{D}| \times L_{riu2/u2}\big(|m^r| + |\mu^r|\big)$ and $3 \times |\mathcal{D}| \times L_{riu2/u2}$ bins corresponding to the mappings.

Furthermore, we also take advantage of the multi-scale performance [47] to enhance the discriminative power of DT descriptors. According to that, the proposed operators are utilized to calculate concerning probability distributions with different samples of neighbors $\{(P, R)\}$. The output histograms

are then concatenated and normalized to produce multi-scale descriptors MMDP, MMDP-B, and MLDP-TOP.

6 Enhancing the performance with max-pooling features

Inspirited by CNNs [56, 57], we exploit the stage of max-pooling to obtain more intensity of global characteristics and "deep" patterns for DT representation (hereunder referred as max-pooling features). Accordingly, for a filtering window with size of $\omega \times \omega$, the max-pooling process is taken into account to analyze a video $\mathcal{V}$ by striding the filter at 1 for calculating $\mathcal{V}_1$ of "deep" features and at ω for capturing $\mathcal{V}_2$ of global characteristics. Then MDPs of the obtained sequences are computed and concatenated with those of $\mathcal{V}$ to form an enhanced MDP (EMDP) descriptor as

$$\mathrm{EMDP}_{\Omega,\mathcal{D}}(\mathcal{V}) = \left[\mathrm{MDP}_{\Omega,\mathcal{D}}(\mathcal{V}), \mathrm{MDP}_{\Omega,\mathcal{D}}(\mathcal{V}_1), \mathrm{MDP}_{\Omega,\mathcal{D}}(\mathcal{V}_2)\right] \tag{23}$$

Figure 7 graphically demonstrates an example of this computation. Similarity to MDP operator, EMDP is also considered in multi-scale regions to capture the further local features for structuring a more robust descriptor MEMDP.

7 Experiments

We verify our method on different benchmark DT datasets: UCLA [1], DynTex [58], and DynTex++ [59]. For DT recognition task, the final histogram, calculated by our proposed descriptor on a DT video, is used as an input feature vector for classification utilizing a linear SVM (Support Vector Machine) which is trained according to specific experimental protocols. Then the obtained results are compared to those of the state-of-the-art approaches. In our experiments, we conduct the default parameters of LIBLINEAR[2] tool in which learning algorithms of linear SVMs have been implemented into [60].

[2]https://www.csie.ntu.edu.tw/ cjlin/liblinear

7.1 Experimental settings

7.1.1 Settings for moment volumes

Since encoding dynamic textures on the high-order moment volumes results out DT descriptor with a large dimension, it should be considered in this work two first orders of moment volume model to calculate mean (m^1) and variance (μ^2) sequences, i.e., $|m^1| = |\mu^2| = 1$. Structuring volume elements adopted to this filtering process are a set of supporting 3D spheres as $\mathcal{S} = \{\Omega_1, \Omega_2, ..., \Omega_m\}$. Particularly, we have experimented on various elements of $\{\{(6,1)\}, \{(14,1)\}, \{(6,1),(6,2)\}, \{(6,1),(14,2), \{(14,1),(14,2)\}\}$. In the coming sections, we only present experiments using supporting volume of $\Omega = \{(6,1)\}$ owing to its better performance on the different DT datasets. An instance of filtering process exploiting this structuring element in two first-order moment volumes (i.e., m^1 and μ^2) is graphically illustrated in Fig. 8.

7.1.2 Parameter settings for DT descriptors

Based on the two first-order filtered sequences to structure DT descriptors in justifiable dimension, we compute MDP, MDP-B, and LDP-TOP descriptors in 4 directions of $\mathcal{D} = \{0°, 45°, 90°, 135°\}$. For MDP descriptor, formed by the extended xLDP operator, three kinds of integrating complementary components can be experimented as $\{\text{MDP}_{D\_M}, \text{MDP}_{D\_M/C}, \text{MDP}_{D\_M\_C}\}$ (hereunder called MDP descriptors for all) corresponding to dimensions of $\{48L_{riu2/u2}, 72L_{riu2/u2}, 48(L_{riu2/u2}$ $2)\}$ with $riu2$ and $u2$ mappings respectively. In respect of MDP-B and LDP-TOP descriptors, their lengths in this case are $24L_{riu2/u2}$ and $12L_{riu2/u2}$. Several particular dimensions of these descriptors of $riu2$ mapping can be seen in Table 1, in which it is possible for our operators to compute multi-scale descriptors for capturing more robust structural relations while retaining their sizes in reasonable dimensions compared to other LBP-based methods. Similarity to the settings for encoding MDP, descriptor EMDP is extra enhanced with the enhanced features computed from max-pooling videos which are formed with the *vl_nnpool*() function[3] using the default parameters except Square filter = 2×2, Stride = 1 for "deep" features, and Stride = 2 for global characteristics.

[3]http://www.vlfeat.org/matconvnet/mfiles/vl_nnpool

7.2 Datasets and experimental protocols

In this section, at first, features of DT datasets and corresponding protocols are presented in detail. Then their key properties are shortened in Table 2 for a convenient glance.

7.2.1 UCLA dataset

UCLA dataset [1] consists of 50 categories with four sequences for each, i.e., 200 DT sequences in total, those which are recorded in different conditions to picture *fountain*, *fire*, *boiling water*, *waterfall*, *plant*, and *flower*. Each original sequence is captured in 75 frames with dimension of 160×110 for each frame. A slight version of UCLA often utilized for DT classification task is structured by splitting the initial videos into sub-sequences with a 48×48 pixel window located at the major region of dynamical features (see Fig. 9 for several DT samples). Hereinafter, three benchmark schemes are popularly used for evaluations of DT recognition.

- *50-class breakdown:* Two experimental settings are usually focused on this scheme:

  Leave-one-out (50-LOO): Following the protocol in [1, 64, 27], just one sample in the scheme is taken out for testing and the rest for training. This trial is performed in repetition for all samples and the final estimation is resulted by the mean of all obtained rates.

  Four cross-fold validation (50-4fold): Similarity to [28, 27, 35], one-fourth of each class is addressed for testing and the remain for learning. The experiment is repeated four times with distinct test samples for each runtime. The final recognition rate is reported by the average of all repetitions.

- *9-class breakdown:* This scheme is reorganized from the 50-class model by categorizing its DT sequences into 9 classes named as *boiling water(8), fire(8), flowers(12), fountains(20), plants(108), sea(12), smoke(4), water(12), and waterfall(16)*, where the numbers in parentheses denote total of sequences of each class (see Fig. 9 for several samples corresponding to their groups). The experimental setting is adopted as that in [19, 59, 28], in which one half of DT sequences in each category are randomly selected for training and the remain for testing. The average of 20 runtimes is reported as the output rate.

- *8-class breakdown:* As the dominant cardinality of the *plants(108)* group in 9-class, it is eliminated to form 8-class scheme with more challenges for DT evaluation. Following [19, 28], the configuration for experiment is set like that 50% of DT sequences randomly taken out from each class is utilized for training and the rest for testing. Similar to 9-class, the trial on this scheme is also run 20 times and the mean of those forms the final rate.

7.2.2 DynTex dataset

DynTex dataset [58] originally consists of more than 650 videos captured under various changes of environmental elements and taped in AVI format. In our experiments, we utilize *"pr1"* DynTex version[4] of 679 DT sequences arranged in 10 seconds with justifiable dimension of 352×288 and 250 colour frames. Several videos along with their classification labels are illustrated in Fig. 10. Following the experimental settings in [4, 27, 35], a sub-dataset for DT recognition is founded by selecting 35 videos from *"pr1"*, called as DynTex35. Each video is treated as a category comprising 8 non-overlapping sub-DTs that are separated from this sequence using random clipping points along axes of X, Y, and T, but not at half in these. An instance of splitting points is sampled as $x = 170, y = 130, t = 100$ in [4]. Furthermore, partitioning along T axis of each sequence results out two sub-DTs. Consequently, 10 sub-DTs with different spatio-temporal measurements are collected for each category.

In addition, three challenging benchmark datasets[5] are also composed in [58] for DT classification as follows: *Alpha* comprises 60 videos categorized into three groups with 20 DTs per each as *"grass"*, *"sea"*, and *"trees"*. *Beta* consists of 162 sequences grouped into 10 classes: *"sea"*, *"vegetation"*, *"trees"*, *"flags"*, *"calm water"*, *"fountains"*, *"smoke"*, *"escalator"*, *"traffic"*, and *"rotation"* with different cardinality of sequences for each. *Gamma* includes 264 DTs divided into 10 categories as *"flowers"*, *"sea"*, *"naked trees"*, *"foliage"*, *"escalator"*, *"calm water"*, *"flags"*, *"grass"*, *"traffic"*, and *"fountains"* with various numbers of sequences for each. Similar to the protocol set up in [27, 65, 23], leave-one-out cross validation is utilized to verify the performance of our framework on those for the DT recognition problem.

[4] http://dyntex.univ-lr.fr/download.html

[5] http://dyntex.univ-lr.fr/classification_datasets/classification_datasets.html

7.2.3 DynTex++ dataset

The sequences in DynTex dataset are restructured to form a richer benchmark for DT recognition, named DynTex++ [59]. Accordingly, 345 DynTex's raw videos are split into sub-sequences with the fixed size of $50 \times 50 \times 50$ so that they just include the main dynamic texture without any background or other dynamic structures. The clipped DTs are then filtered by some techniques to expose 3600 sequences, those which are grouped into 36 categories with 100 DTs for each. We follow the same experimental setting as that in [59, 27, 62] for evaluation. One half of samples from each class is randomly selected for training, and the remain for testing. The experiment is repeated 20 times to report the average performance as the final result.

7.3 Experimental results

Performances on different benchmark DT datasets (UCLA, DynTex, and DynTex++) of our framework, in which the proposed operators along with $riu2$ and $u2$ mappings are ultilized to encode filtered videos in single-scale and multi-scale analyses for DT description, are detailed in corresponding Tables 4, 5, and 7 respectively. Based on the experimental results, we could make some crucial statements as follows.

First, as mentioned in Sections 3.2 and 3.3, exploiting moment volumes makes DT representation more insensitive to noise and illumination. Our experiments have verified that the DT descriptors MDP and MDP-B, computed on the filtered videos, have outstanding performance in comparison to the LDP-TOP's, encoded on the raw DT sequence with non-supporting volumes (see Tables 4, 5, 7 for MDP, MDP-B, and Table 8 for LDP-TOP descriptor). In this regard, two first-order filtered ("mean" and "variance") videos have notably contributed to the performance of the proposed descriptors (see Table 3). Second, it is in accordance with our evaluation in Section 4.4 that the combination of complemented components comprises additional discriminant information. As expected, all MDP descriptors outperform significantly compared to MDP-B with a single complementary element (see Tables 4, 5, and 7). Third, MDP descriptors exploiting the factor of directional center contrast (the component LDP_C of an extended operator xLDP) are often more informative than others. Therein, jointing with this factor makes those more robust to noise (see Tables 4 and 7). Fourth, MDP descriptors with $riu2$ mapping not only have tiny dimension but also deal

with more efficiently than $u2$. Fifth, it is consistent with our analysis in Section 5 that multi-scale encoding allows to capture more local directional structures in larger regions. More specifically, multi-scale descriptors of $riu2$ mapping ($\{(P,R)\}^{riu2}$) are more efficient than single-scale. Therein, 2-scale (e.g., $\{(8,1),(16,2)\}$) achieves good results but the performance of 3-scale, i.e., $\{(8,1),(16,2),(24,3)\}$ seems more "stable" on most of the benchmark DT datasets thanks to considering spatial features on the broad locality. Consequently, it should be recommended for implementation in practice, and also be the setting chosen for comparing with the state-of-the-art performances.

Furthermore, the MDP-B, which is based on the basic LDP (see Sections 4.4 and 7.3.4), has not performed as efficiently as MDP descriptors structured by the extended LDP operator. MDP also outperforms in comparison with LDP-TOP using the same configuration. These facts prove the effectiveness of our proposed components: the important extensions of LDP operator and the model of moment volumes. However, it should be noted that MDP-B also obtains promising results compared to existing LBP-based methods thanks to the contribution of the r-order moment volume model.

In aspect of comparison with the existing approaches, our proposed method with a simple encoding technique conducts outstandingly in DT recognition issue compared to LBP-based variants for DT representation. In addition, its ability is the same as that of deep-learning-based frameworks in several circumstances (see Table 6). Hereafter, comprehensive evaluations of our proposal on different DT datasets are expressed clearly, in which if MDP descriptors are not specified their implemented configurations in detail, the default setting is mentioned for them, i.e., $\{(8,1),(16,2),(24,3)\}^{riu2}$.

7.3.1 Recognition on UCLA dataset

It can be verified in Tables 4 and 6 that the proposed method obtains the best recognition rates of 100% for both 50-LOO and 50-4fold schemes compared to the state-of-the-art results. For 9-class and 8-class scenarios, our proposal also acquires competitive performances. Hereafter, estimations on each of UCLA's sub-datasets are detailed specifically.

50-class: It can be realized in Table 6 that $\text{MMDP}_{D\_M\_C}$ and MMDP_{D_M} achieve good results with 100% and 99.5% on 50-LOO and 50-4fold scenarios respectively. In aspect of the chosen comparing setting (see Section 7.3), $\text{MMDP}_{D\_M/C}$ with only 3888 bins outperforms with rate of 100% on both scenarios. It is the best performance in comparison to all existing methods

including deep-learning-based approaches PCANet-TOP [23], D3 [24], and DT-CNN [22]. The filter-based method, MBSIF-TOP [27], achieves rate of 99.5% using a 7-scale descriptor of larger dimension (5376 bins). Utilizing multi-fractal analysis to measure spatio-temporal features, DFS [66] obtains the same ours (100%) on 50-4fold scheme but it has not dealt with well on other challenging DT datasets (e.g., DynTex). Similarly, PI/PD-LBP variants [39] structure DT descriptors with grand dimensions using complicated learning procedures, and they have not been tested on DynTex.

9-class: In this scheme, MMDP_{D_M} with rate of 98.90% is the best performance compared to other MDP descriptors. In the meanwhile, accuracies of $\text{MMDP}_{D\_M\_C}$ and $\text{MMDP}_{D\_M/C}$ are 98.35% and 98.70%, slightly lower rates of 99.20%, 99.35%, and 99.60% which are reported by CVLBC [63], FD-MAP [14], and DNGP [15] respectively. However, CVLBC and FD-MAP is not better than ours on other scenarios (except 8-class) of UCLA dataset, while DNGP has a complex representation. It should be noted that our method outperforms lightly compared to DT-CNN's [22], 98.05% for AlexNet and 98.35% for GoogleNet deep learning framework. Specific recognition rate on each category in Fig. 11 illustrates that $\text{MMDP}_{D\_M/C}$ has mainly confused sequences of *"Fire"* with *"Plants"*, *"Water"* with *"Waterfall"*, and *"Smoke"* with *"Water"*. The reason for that may be the similar characteristics of those.

8-class: Obtaining rate of 98.7% with $\text{MMDP}_{D\_M/C}$ in the more challenging scheme (see Table 4), it is interesting to note that the ability of our method is nearly the same as DT-CNN's [22] utilizing deep-learning-based frameworks: AlexNet (98.48%) and GoogleNet (99.02%). It can be also observed in Table 6 that our method has the best performance among LBP-based methods, excluding CVLBC [63]. As mentioned above, it does not handle well on other schemes and has not been verified on the more challenging subsets of DynTex (i.e., Alpha, Beta, Gamma). Other non-LBP-based approaches, like Orthogonal Tensor DL (99.50%) [25], STLS (99.5%) [31], DNGP (99.4%) [15], DFS (99.2%) [66], 3D-OTF (99.5%) [29], FDT (99.35%) [14], FD-MAP (99.57%) [14], deal with more effectively than ours but their drawbacks are either sophisticated computation (e.g., Orthogonal Tensor DL, DNGP) or inefficient operation on other DT datasets (e.g., Orthogonal Tensor DL, DFS, 3D-OTF, STLS, FDT, FD-MAP). The confusion matrix confusion of each class in Fig. 12 indicates that $\text{MMDP}_{D\_M/C}$ has principally confused the properties of *"Smoke"* sequences with *"Water"* due to their alike features.

7.3.2 Recognition on DynTex dataset

Tables 5 and 6 indicate that our method obtains the best results compared to existing LBP-based methods and other non-deep-leaning techniques on this scheme. Specific evaluations on each of DynTex's variants are expressed in detail as follows.

DynTex35: It can be observed in Table 5 that the highest rate of recognition on this scenario is 100% reported by $\text{MDP}^{u2}_{D\_M}(24,3)$ and $\text{MDP}^{u2}_{D\_M\_C}(24,3)$. In the meanwhile, MMDP_{D_M}, $\text{MMDP}_{D\_M\_C}$, and $\text{MMDP}_{D\_M/C}$ result out lightly lower rate of 99.43%. This is because of the similarity of features in two classes c and d, as shown in Fig. 14, that they are not able to differentiate. The detail of classification rate of $\text{MMDP}_{D\_M/C}$ is exposed in Fig. 13. CVLBC [63] obtains accuracy of 99.71% on this scheme (see Table 6), sightly higher than our MMDP descriptors' but it has not verified on other challenging variants of DynTex (i.e., Alpha, Beta, Gamma).

Alpha: In this scheme, MMDP_{D_M} and $\text{MMDP}_{D\_M\_C}$ with rate of 98.33% (see Table 5) outperform compared to that of $\text{MMDP}_{D\_M/C}$ with 96.67% due to the confusion of two DT sequences (see Fig. 15). Those results are also the best in comparison with all existing methods excluding deep-leaning-based approaches st-TCoF [21], DT-CNN [22], and D3 [24].

Beta: It can be realized in Tables 5 and 6 that our MDP descriptors have the best performance compared to all non-deep-leaning-based methods. More specifically, $\text{MMDP}_{D\_M\_C}$ of $(16,2)(24,3)^{riu2}$ gains the highest rate of 98.15%, slightly better than MMDP_{D_M} and $\text{MMDP}_{D\_M/C}$ with (96.91%) and (97.53%) respectively. Those performances are much better than PCANet-TOP's [23] and about 1% to 3% lower than st-TCoF's [21], DT-CNN's [22], and D3 [24], in which exploiting complicated learning algorithms along with tremendous dimension of DT representation while those are crucial to ensure feasible implementations in practice. The confusion matrix of $\text{MMDP}_{D\_M/C}$ in Fig. 16 indicates that it has mostly confused *"Rotation"* sequences with *"Vegetation"* and *"Trees"*.

Gamma: In this scenario, rate of 94.68% is the best recognition pointed out by $\text{MDP}^{u2}_{D\_M/C}(16,2)$, while multi-scale MMDP also obtains good results from 92% to 93%. Towards the setting chosen for comparison, $\text{MMDP}_{D\_M/C}$ achieves rate of 92.05%, better than all existing methods excepting LBP-TOP's implemented in [21] and that of deep-leaning-based approaches. In order to address which categories have enforced the misunderstanding of $\text{MMDP}_{D\_M/C}$ for the improvement, the confusion matrix is figured out as in

Fig. 17. According to that, mutual confusion between sequences of *"Fountains"* and *"Calm water"* should be concentrated on for perspectives.

7.3.3 Recognition on Dyntex++ dataset

It can be observed from Tables 6 and 7 that MDP descriptors have performed well in comparison to the existing approaches. Specifically, the best recognition rate on this scheme is 96.51% (see Table 7) reported by $\text{MDP}^{u2}_{D\_M\_C}(8,1)$. The descriptors of MMDP_{D_M}, $\text{MMDP}_{D\_M\_C}$, and $\text{MMDP}_{D\_M/C}$ obtain 95.58%, 95.7%, and 95.86% respectively, those which are the highest rates compared to the existing methods using SVM algorithm for classification. In aspect of the comparing setting, the performance of $\text{MMDP}_{D\_M/C}$ is nearly the same MBSIF-TOP's (97.12%) [27] with 8-scale descriptor formed by 8 learned filters, and about 3% lower than DT-CNN's (98.18%) [22] using deep learning techniques of AlexNet for learning DT features. The LBP-based method, MEWLSP [62], acquires the highest recognition rate of 98.48% on this scheme, even better than DT-CNN's (98.18%) [22]. However, it does not outperform on UCLA dataset compared to ours as well as has not been justified on other challenging DynTex variants (i.e., Alpha, Beta, Gamma). Another sophisticated method utilizing deep learning framework of GoogleNet [22] has prominent classification rate but it takes a long time to handle DT features with a huge complicated computation while these costs are crucial in real-time applications of computer vision. Accuracies of $\text{MMDP}_{D\_M/C}$ on each categories are detailed in Fig. 18. Accordingly, our descriptor outperforms on most of categories, only five of them are really challenges for the future work (see Fig. 19).

7.3.4 Assessing the proposed components: Recognition with MDP-B and LDP-TOP

We address in this section some experiments for verifying our proposed components. Two following descriptors (see also Section 5 for more details) are considered: i) LDP-TOP that applies directly the second-order LDP operator on three orthogonal planes of raw videos; ii) MDP-B has the same architecture as that of MDP descriptors but on the contrary, it is based only on LDP operator. It is evident that the comparisons between LDP-TOP and MDP-B, between MDP-B and MDP, allow to highlight respectively the contribution of moment volumes, and that of the extended operator xLDP.

It could be seen from Tables 4, 5, 6, 7, and 8 that MDP descriptors are more efficient and "stable" than MDP-B and LDP-TOP ones. Table 6 shows that our proposals permit to prominently improve MDP's performance compared to the straightforward LDP-TOP version on most of DT datasets (e.g., up to 8.64% on Beta dataset). It also outperforms in comparison with MDP-B on various datasets (e.g., up to 9.26% on Beta).

Moreover, the execution of LDP-TOP is impaired in comparison to MDP-B's on most of the DT datasets (see also Table 6) due to non-supporting volume taken into account. This fact proves that considering moment volumes inspite of raw videos allows to capture more robust and discriminative features to enhance the performance of DT descriptors.

In the meanwhile, with the same configuration, MDP-B fails behind MDP descriptors on most of DT datasets because the typical second-order LDP is used instead of our extended operator xLDP (see Section 4.4). This shows the important contribution of two proposed extensions for LDP operator to make DT descriptors more robust and discriminative. However, it should be noted that MDP-B's performance produces competitive results that are still comparable with the existing methods in several circumstances thanks to the collaboration of the filtered videos figured out by the proposed model of r-order moment volumes.

Because of those, the below evaluations mainly focus on the performance of MDP-B compared to the existing approaches.

UCLA: The performance of MMDP-B with multi-scale setting of $\{(8,1),(16,2),(24,3)\}^{riu2}$ acquires recognition rates of 99.5%, 98.5%, 98.05%, and 97.61% for 50-LOO, 50-4fold, 9-class, and 8-class scenarios respectively, those which are comparable to the LBP-based methods (see Table 6). In 50-LOO and 50-4fold schemes, the results of LDP-TOP$^{u2}(16,2)$ are also promising with rates of 99% and 99.5% (see Table 8).

DynTex: In this scheme, MMDP-B and MLDP-TOP with comparing configuration just break down on Beta with classification rate of 88.27% and 88.89% respectively while they and their other settings perform well on other variants of DynTex dataset (see Table 5). More specifically, the best recognition rates on DynTex35 is 99.43% resulted by MDP-B$^{u2}(24,3)$, LDP-TOP$^{u2}(24,3)$, and 99.14% reported by MDP-B$^{riu2}(24,3)$ with only 624 dimensions. Towards the comparing setting, MMDP-B and MLDP-TOP achieve rate of 98.86% on DynTex35, the best classification among the LBP-based variants except MEWLSP's [62] (99.71%) (see Table 6). Although not better than the ability of MDP on Beta, MDP-B obtains comparable rates

against those of all existing techniques excepting deep learning methods, i.e., st-TCoF [21], D3 [24], DT-CNN [22]. Furthermore, it is interesting to note that the operation of MMDP-B is slightly better than MMDP's in verifying on Gamma scheme with 93.18% in contrast to 92.05% of MMDP.

DynTex++: Utilizing complicated learning algorithms, DT-CNN [22] outperforms dominantly on this scenario (98.58%). The MMDP-B descriptor of $\{(8,1),(16,2),(24,3)\}^{riu2}$ with only size of 1,350 bins gains the promising results with rate of 95.82%, lightly better than that of MMDP_{D_M} and $\text{MMDP}_{D\_M\_C}$. Thanks to exploiting spatio-temporal information of the moment volumes, $\text{MDP-B}^{u2}(8,1)$ results out the highest rate of 96.51%, just about 2% lower than DT-CNN's [22].

7.3.5 Assessing impact of max-pooling features: Recognition with EMDP descriptor

We conduct in this section several experiments for investigating the impact of max-pooling features on encoding MDP patterns. As validated in Section 7.3 that the configurations of $riu2$ mapping and $_{D\_M/C}$ integration reported the best performance, we just address these settings to compute EMDP descriptor.

It could be verified from Tables 4, 5, 7, and 9 that EMDP descriptor is more discriminative than MDP thanks to the contribution of max-pooling features. Specifically, the performance of its single-scale variants has significantly improved in the recognition issue of 50-class schemes in the UCLA dataset. For instance, with $\big\{(P,R)\big\} = \big\{(8,1)\big\}$ of $riu2$ mapping, EMDP obtains 1.5% better than that of MDP (see Tables 4 and 9). In the setting chosen for comparison with the state of the art $\big($i.e., $\big\{(8,1),(16,2),(24,3)\big\}^{riu2}\big)$, EMDP outperforms about 0.3% compared to MDP (99.43%) on DynTex35. Particularly, it gains 93.94% rate of recognition on the complicated dataset, Gamma, about 2% higher than MDP's. In terms of classification on DynTex++, the operation of EMDP looks more "stable" and achieves a little better rate with 96.03% in comparison to those of MDP with 95.86% (see Tables 7 and 9).

In general, it is validated that the impact of the max-pooling features is positive in enhancing the performance of the proposed descriptors. Table 10 indicates the important contribution of these enhanced features in shallow analysis. It can be realized from this table that it is possible to take advantage of these in the deeper max-pooling layers as well as to combine this

computation with other advance components of CNNs in the further context.

7.4 Global discussion

Based on the above experimental results on benchmark DT datasets, it can be derived several general findings as follows.

- The proposed moment volume model can be judged as a filter bank approach for pre-processing techniques since its principle is a local filter in which its operator is inherited from the basic LBP concept with low computing costs (see Sections 3.2 and 3.3) to exploit robust and discriminant features of DT videos. Outputs of this process, i.e., "mean" and "variance" videos, are regarded as complementary parts to boost the discriminative power of DT representation (see Table 3).

- Considering larger supporting volumes to construct moment volumes can be lead to outputs of blurred videos. This induces that encoding on these videos of our proposed operators reduces their performance due to the increase of noise patterns structured from the near uniform voxels. It can be seen from Tables 5, 7, and 11 that the performances of DT descriptors are affected significantly by blurred videos dealt with by the model of two first-order moment volumes with large supporting regions $\Omega = \{(14, 1), (14, 2)\}$. Moreover, bigger elements of supporting volumes also increase the time cost of filtering voxel features without enhancing the operation of recognition as expected. In practice, the setting of regional volume $\Omega = \{(6, 1)\}$ should be empirically recommended for the proposed model of r-order moment volumes.

- Two proposed extensions for LDP operator resulting in the extended operator xLDP make our descriptor MDP even more robust and discriminative than the straightforward version MDP-B, which is based on LDP, in spite of the fact that this simple descriptor is also very competitive compared to the state-of-the-art results.

- MDP descriptors, based on the configuration of $\{(8, 1), (16, 2), (24, 3)\}^{riu2}$, have more substantial performance compared to others thanks to more relationships of local directional structures involved in.

- Directional complement of center contrast level LDP_C has a trivial impact on improving the performance of DT descriptor in our framework

(see Tables 4, 5, and 7). Concatenating it to form the corresponding descriptor would just grow up 2 bins for each concerned direction, i.e., $L_{riu2/u2} + 2$, while that would be double size in case of jointing, i.e., $2L_{riu2/u2}$. Therefore, it is possible to make a trade-off between accuracy of recognition and the computing consumption in particular applications.

8 Conclusions and perspectives

We have presented effective descriptors for DT representation. Our main contribution is four-fold. We have introduced model of moment volumes as a simple yet efficient pre-processing techniques to take into account robust and discriminative features of DT videos. We then proposed two major extensions for LDP operator making it more distinctive than the typical version for capturing local derivative variations. Finally, we address different efficient descriptors based on above propositions for DT recognition. The experimental results on various benchmark DT datasets have demonstrated that our approach significantly outperforms compared to the existing methods.

Due to turbulent motions of DTs, full directions should be addressed for the future works to entirely investigate the relations of local informative directions for an image texture. Furthermore, in consideration of treating the large dimension problem, encoding DT features with n-order MDP^n ($n \geq 3$) operator can be done on high-order moment volumes or Gaussian-based outcomes [67, 68]. This may obtain more robust spatio-temporal relationships to boost the discriminative power of DT description.

Acknowledgment

We would like to express our sincere appreciation for the insightful and valuable comments of the editors and reviewers which allow us to clarify the presentation of this work.

References

[1] Saisan, P., Doretto, G., Wu, Y.N., Soatto, S.: Dynamic texture recognition. In: CVPR. (2001) 58–63

[2] Chan, A.B., Vasconcelos, N.: Modeling, clustering, and segmenting video with mixtures of dynamic textures. IEEE Trans. PAMI **30** (2008) 909–926

[3] Nguyen, X.S., Nguyen, T.P., Charpillet, F., Vu, N.S.: Local derivative pattern for action recognition in depth images. Multimedia Tools Appl **77** (2018) 8531–8549

[4] Zhao, G., Pietikäinen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. PAMI **29** (2007) 915–928

[5] Zhang, W., Smith, M.L., Smith, L.N., Farooq, A.R.: Gender and gaze gesture recognition for human-computer interaction. CVIU **149** (2016) 32–50

[6] Maqueda, A.I., del-Blanco, C.R., Jaureguizar, F., García, N.N.: Human-computer interaction based on visual hand-gesture recognition using volumetric spatiograms of local binary patterns. CVIU **141** (2015) 126–137

[7] Barmpoutis, P., Dimitropoulos, K., Grammalidis, N.: Smoke detection using spatio-temporal analysis, motion modeling and dynamic texture recognition. In: EUSIPCO. (2014) 1078–1082

[8] Mettes, P., Tan, R.T., Veltkamp, R.C.: Water detection through spatio-temporal invariant descriptors. CVIU **154** (2017) 182–191

[9] Nguyen, T.P., Manzanera, A., Garrigues, M., Vu, N.: Spatial motion patterns: Action models from semi-dense trajectories. IJPRAI **28** (2014)

[10] Peh, C., Cheong, L.F.: Synergizing spatial and temporal texture. IEEE Trans. IP **11** (2002) 1179–1191

[11] Péteri, R., Chetverikov, D.: Qualitative characterization of dynamic textures for video retrieval. In Wojciechowski, K.W., Smolka, B., Palus, H., Kozera, R., Skarbek, W., Noakes, L., eds.: ICCVG. Volume 32 of Computational Imaging and Vision. (2004) 33–38

[12] Péteri, R., Chetverikov, D.: Dynamic texture recognition using normal flow and texture regularity. In Marques, J.S., de la Blanca, N.P., Pina, P., eds.: IbPRIA. Volume 3523 of LNCS. (2005) 223–230

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

[14] Nguyen, T.T., Nguyen, T.P., Bouchara, F., Nguyen, X.S.: Directional beams of dense trajectories for dynamic texture recognition. In Blanc-Talon, J., Helbert, D., Philips, W., Popescu, D., Scheunders, P., eds.: ACIVS. (2018) 74–86

[15] Rivera, A.R., Chae, O.: Spatiotemporal directional number transitional graph for dynamic texture recognition. IEEE Trans. PAMI **37** (2015) 2146–2152

[16] B. Chan, A.B., Vasconcelos, N.: Classifying video with kernel dynamic textures. In: CVPR. (2007) 1–6

[17] Mumtaz, A., Coviello, E., Lanckriet, G.R.G., Chan, A.B.: Clustering dynamic textures with the hierarchical EM algorithm for modeling video. IEEE Trans. PAMI **35** (2013) 1606–1621

[18] Wang, Y., Hu, S.: Chaotic features for dynamic textures recognition. Soft Computing **20** (2016) 1977–1989

[19] Ravichandran, A., Chaudhry, R., Vidal, R.: View-invariant dynamic texture recognition using a bag of dynamical systems. In: CVPR. (2009) 1651–1657

[20] Mumtaz, A., Coviello, E., Lanckriet, G.R.G., Chan, A.B.: A scalable and accurate descriptor for dynamic textures using bag of system trees. IEEE Trans. PAMI **37** (2015) 697–712

[21] Qi, X., Li, C.G., Zhao, G., Hong, X., Pietikainen, M.: Dynamic texture and scene classification by transferring deep image features. Neurocomputing **171** (2016) 1230 – 1241

[22] Andrearczyk, V., Whelan, P.F.: Convolutional neural network on three orthogonal planes for dynamic texture classification. Pattern Recognition **76** (2018) 36 – 49

[23] Arashloo, S.R., Amirani, M.C., Noroozi, A.: Dynamic texture representation using a deep multi-scale convolutional network. JVCIR **43** (2017) 89 – 97

[24] Hong, S., Ryu, J., Im, W., Yang, H.S.: D3: recognizing dynamic scenes with deep dual descriptor based on key frames and key segments. Neurocomputing **273** (2018) 611–621

[25] Quan, Y., Huang, Y., Ji, H.: Dynamic texture recognition via orthogonal tensor dictionary learning. In: ICCV. (2015) 73–81

[26] Quan, Y., Bao, C., Ji, H.: Equiangular kernel dictionary learning with applications to dynamic texture analysis. In: CVPR. (2016) 308–316

[27] Arashloo, S.R., Kittler, J.: Dynamic texture recognition using multiscale binarized statistical image features. IEEE Trans. Multimedia **16** (2014) 2099–2109

[28] Xu, Y., Quan, Y., Ling, H., Ji, H.: Dynamic texture classification using dynamic fractal analysis. In: ICCV. (2011) 1219–1226

[29] Xu, Y., Huang, S.B., Ji, H., Fermüller, C.: Scale-space texture description on sift-like textons. CVIU **116** (2012) 999–1013

[30] Ji, H., Yang, X., Ling, H., Xu, Y.: Wavelet domain multifractal analysis for static and dynamic texture classification. IEEE Trans. IP **22** (2013) 286–299

[31] Quan, Y., Sun, Y., Xu, Y.: Spatiotemporal lacunarity spectrum for dynamic texture classification. CVIU **165** (2017) 85–96

[32] Baktashmotlagh, M., Harandi, M.T., , A., C. Lovell, B.C., Salzmann, M.: Discriminative non-linear stationary subspace analysis for video classification. IEEE Trans. PAMI **36** (2014) 2353–2366

[33] Zhao, G., Ahonen, T., Matas, J., Pietikäinen, M.: Rotation-invariant image and video description with local binary pattern features. IEEE Trans. IP **21** (2012) 1465–1477

[34] Tiwari, D., Tyagi, V.: Dynamic texture recognition based on completed volume local binary pattern. MSSP **27** (2016) 563–575

[35] Tiwari, D., Tyagi, V.: A novel scheme based on local binary pattern for dynamic texture recognition. CVIU **150** (2016) 58–65

[36] Nguyen, T.T., Nguyen, T.P., Bouchara, F.: Completed local structure patterns on three orthogonal planes for dynamic texture recognition. In: IPTA. (2017) 1–6

[37] Nguyen, T.T., Nguyen, T.P., Bouchara, F.: Completed statistical adaptive patterns on three orthogonal planes for recognition of dynamic textures and scenes. J. Electronic Imaging **27** (2018) 053044

[38] Ren, J., Jiang, X., Yuan, J., Wang, G.: Optimizing LBP structure for visual recognition using binary quadratic programming. IEEE Signal Processing Letters **21** (2014) 1346–1350

[39] Ren, J., Jiang, X., Yuan, J.: Dynamic texture recognition using enhanced LBP features. In: ICASSP. (2013) 2400–2404

[40] Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. PAMI **24** (2002) 971–987

[41] Zhao, Y., Huang, D.S., Jia, W.: Completed Local Binary Count for Rotation Invariant Texture Classification. IEEE Trans. IP **21** (2012) 4492–4497

[42] Fathi, A., Naghsh-Nilchi, A.R.: Noise Tolerant Local Binary Pattern Operator for Efficient Texture Analysis. Pattern Recognition Letters **33** (2012) 1093–1100

[43] Nguyen, T.P., Manzanera, A., Kropatsch, W.G., N'Guyen, X.S.: Topological attribute patterns for texture recognition. Pattern Recognition Letters **80** (2016) 91–97

[44] Guo, Z., Zhang, L., Zhang, D.: A completed modeling of local binary pattern operator for texture classification. IEEE Trans. IP **19** (2010) 1657–1663

[45] Guo, Z., Zhang, L., Zhang, D.: Rotation Invariant Texture Classification using LBP Variance (LBPV) with Global Matching. Pattern Recognition **43(3)** (2010) 706–719

[46] Nguyen, T.P., Vu, N., Manzanera, A.: Statistical binary patterns for rotational invariant texture classification. Neurocomputing **173** (2016) 1565–1577

[47] Mäenpää, T., Pietikäinen, M.: Multi-scale binary patterns for texture analysis. In: SCIA. (2003) 885–892

[48] Liao, S., Law, M.W.K., Chung, A.C.S.: Dominant local binary patterns for texture classification. IEEE Trans. IP **18** (2009) 1107–1118

[49] Tan, X., Triggs, B.: Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions. IEEE Trans. IP **19** (2010) 1635–1650

[50] Guo, Y., Zhao, G., Pietikäinen, M.: Discriminative features for texture description. Pattern Recognition **45** (2012) 3834–3843

[51] Nanni, L., Brahnam, S., Lumini, A.: A Simple Method for Improving Local Binary Patterns by considering Non-uniform Patterns. Pattern Recognition **45** (2012) 3844–3852

[52] Vu, N.S., Nguyen, T.P., Garcia, C.: Improving texture categorization with biologically-inspired filtering. Image and Vision Computing **32** (2014) 424–436

[53] Shrivastava, N., Tyagi, V.: An effective scheme for image texture classification based on binary local structure pattern. The Visual Computer **30** (2014) 1223–1232

[54] Liu, L., Lao, S., Fieguth, P.W., Guo, Y., Wang, X., Pietikäinen, M.: Median robust extended local binary pattern for texture classification. IEEE Trans. IP **25** (2016) 1368–1381

[55] Zhang, B., Gao, Y., Zhao, S., Liu, J.: Local derivative pattern versus local binary pattern: Face recognition with high-order local pattern descriptor. IEEE Trans. IP **19** (2010) 533–544

[56] LeCun, Y., Bottou, L., Bengio, Y., Haffner., P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE (1998) 2278–2324

[57] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In Bartlett, P.L., Pereira, F.C.N., Burges, C.J.C., Bottou, L., Weinberger, K.Q., eds.: NIPS. (2012) 1106–1114

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

[59] Ghanem, B., Ahuja, N.: Maximum margin distance learning for dynamic texture recognition. In Daniilidis, K., Maragos, P., Paragios, N., eds.: ECCV. Volume 6312 of LNCS. (2010) 223–236

[60] Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: LIBLINEAR: A library for large linear classification. JMLR **9** (2008) 1871–1874

[61] Tiwari, D., Tyagi, V.: Improved weber's law based local binary pattern for dynamic texture recognition. Multimedia Tools Appl. **76** (2017) 6623–6640

[62] Tiwari, D., Tyagi, V.: Dynamic texture recognition using multiresolution edge-weighted local structure pattern. Computers & Electrical Engineering **62** (2017) 485–498

[63] Zhao, X., Lin, Y., Heikkilä, J.: Dynamic texture recognition using volume local binary count patterns with an application to 2d face spoofing detection. IEEE Trans. Multimedia **20** (2018) 552–566

[64] Derpanis, K.G., Wildes, R.P.: Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. IEEE Trans. PAMI **34** (2012) 1193–1205

[65] Dubois, S., Péteri, R., Ménard, M.: Characterization and recognition of dynamic textures based on the 2d+t curvelet transform. Signal, Image and Video Processing **9** (2015) 819–830

[66] Xu, Y., Quan, Y., Zhang, Z., Ling, H., Ji, H.: Classifying dynamic textures via spatiotemporal fractal analysis. Pattern Recognition **48** (2015) 3239–3248

[67] Nguyen, T.T., Nguyen, T.P., Bouchara, F.: Smooth-invariant gaussian features for dynamic texture recognition. In: ICIP. (2019) 4400–4404

[68] Nguyen, T.T., Nguyen, T.P., Bouchara, F., Vu, N.: Volumes of blurred-invariant gaussians for dynamic texture classification. In Vento, M., Percannella, G., eds.: CAIP. (2019) 155–167

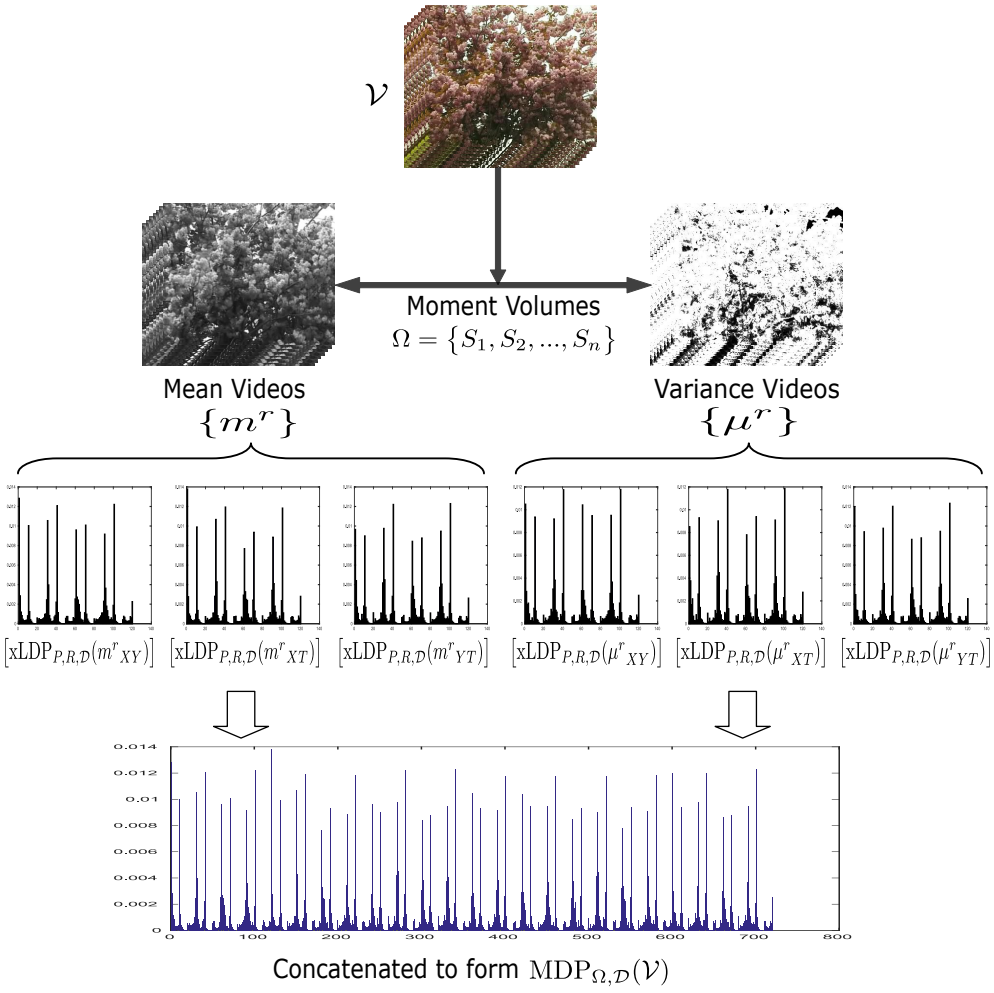

Figure 6: Illustration of structuring proposed DT descriptor.

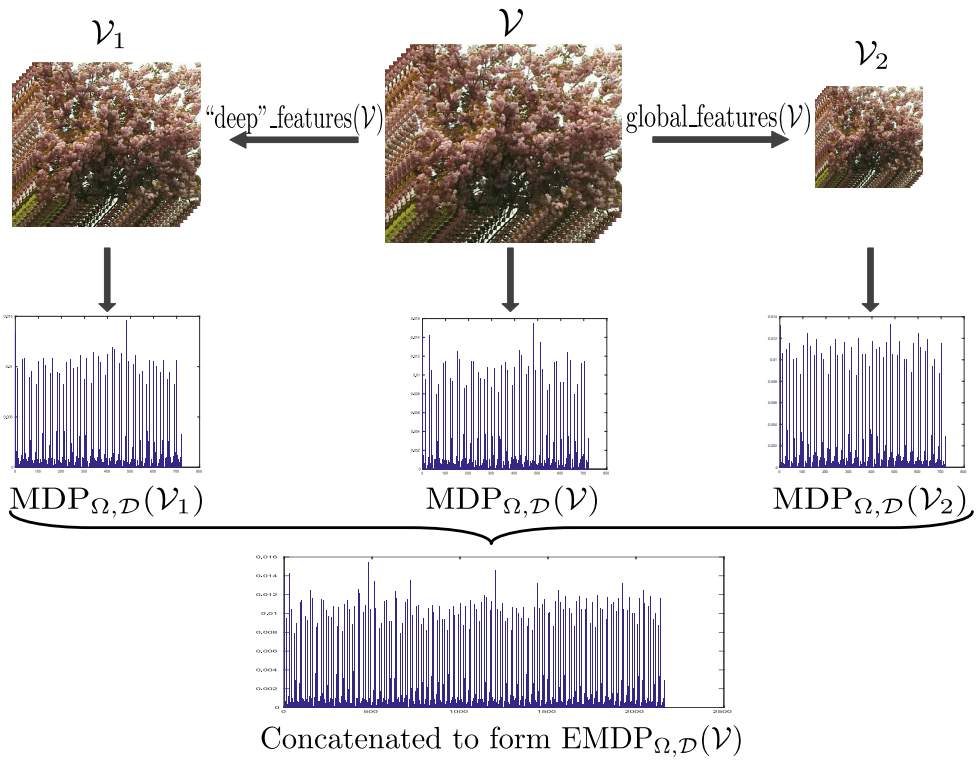

Figure 7: Illustration of constructing EMDP descriptor.

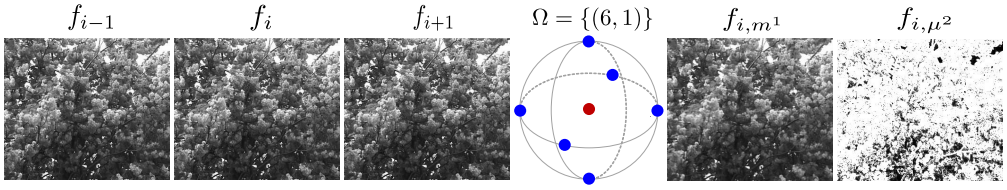

Figure 8: (Best viewed in color) An example of filtering process using two first-order moment volumes (i.e., m^1 and μ^2) with a supporting element of 3D sphere $\Omega = \{(6,1)\}$. Based on frames f_{i-1} and f_{i+1} of a video, frame f_i is filtered to form two corresponding frames f_{i,m^1} and f_{i,μ^2}.

Table 1: Several comparative dimensions of LBP-based descriptors for DT recognition.

Method	Dimensions	with $P = 8$	with $P = 16$
LBP-TOP^{u2} [4]	$3\big(P(P-1)+3\big)$	177	729
VLBP [4]	2^{3P+2}	-	-
CVLBP [34]	$3 \times 2^{3P+2}$	-	-
HLBP [35]	6×2^{P}	1536	-
CLSP-TOP^{riu2} [36]	$6(P+2)^2$	600	1944
WLBPC [61]	6×2^{P}	1536	-
MEWLSP [62]	6×2^{P}	1536	-
CVLBC [63]	$2(3P+3)^2$	1458	5202
$\mathbf{MDP}^{riu2}_{D\_M}$	$48(P+2)$	480	864
$\mathbf{MDP}^{riu2}_{D\_M\_C}$	$48(P+4)$	576	960
$\mathbf{MDP}^{riu2}_{D\_M/C}$	$72(P+2)$	720	1296
MDP-B^{riu2}	$24(P+2)$	240	432
LDP-TOP^{riu2}	$12(P+2)$	120	216

Note: P is the considered neighbors. $riu2$ and $u2$ are two popular mappings for LBP-based variants. MDP-B and MDP descriptors are structured in 4 directions on two first-order filtered videos (also the settings for comparison their performance with the state-of-the-art in DT recognition). In the meanwhile, LDP-TOP is computed on the original videos. "-" denotes that the corresponding descriptor is not implemented in practice due to its huge dimension.

Figure 9: Sample sequences of UCLA dataset

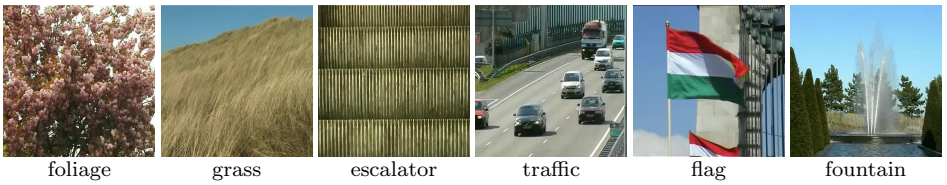

Figure 10: Sample subsets of DynTex dataset

Table 2: A brief of key features of DT datasets and protocols of recognition.

Dataset	Sub-dataset	#Videos	Resolution	#Classes	Protocol of recognition
UCLA	50-class	200	$48 \times 48 \times 75$	50	LOO and 4fold
	9-class	200	$48 \times 48 \times 75$	9	50%/50%
	8-class	92	$48 \times 48 \times 75$	8	50%/50%
DynTex	DynTex35	350	different dimensions	10	LOO
	Alpha	60	$352 \times 288 \times 250$	3	LOO
	Beta	162	$352 \times 288 \times 250$	10	LOO
	Gamma	264	$352 \times 288 \times 250$	10	LOO
	DynTex++	3600	$50 \times 50 \times 50$	36	50%/50%

Note: 50%/50% means a protocol of taking randomly 50% items for training and the rest (50%) for testing. LOO and 4fold are leave-one-out and four cross-fold validation respectively.

Table 3: Recognition (%) on "mean" (m^1) and "variance" (μ^2) videos.

Dataset	50-LOO (UCLA)			Beta (DynTex)			DynTex++		
Descriptor	m^1	μ^2	$\{m^1, \mu^2\}$	m^1	μ^2	$\{m^1, \mu^2\}$	m^1	μ^2	$\{m^1, \mu^2\}$
$\mathbf{MDP}_{D\_M}$	99.50	99.50	100	96.30	95.06	97.53	94.87	94.89	95.58
$\mathbf{MDP}_{D\_M\_C}$	99.50	99.50	100	96.30	95.68	97.53	94.88	94.68	95.70
$\mathbf{MDP}_{D\_M/C}$	99.50	100	100	95.68	95.68	96.91	94.41	94.89	95.86
MDP-B	98.00	98.00	99.50	88.89	89.51	88.27	93.98	94.02	95.82

Note: D_M, D_M_C, and D_M/C are different integrations of complemented components of the extended operator xLDP to form the corresponding MDP descriptors. 50-LOO means results on 50-class breakdown using leave-one-out validation.

Table 4: Classification rates (%) on UCLA using MDP, MDP-B descriptors and their multi-scale settings with mappings of $riu2/u2$.

Scheme	50-LOO				50-4fold				9-class				8-class			
$\{(P,R)\}^{riu2/u2}$	D_M	D_M_C	D_M/C	MDP-B	D_M	D_M_C	D_M/C	MDP-B	D_M	D_M_C	D_M/C	MDP-B	D_M	D_M_C	D_M/C	MDP-B
$\{(8,1)\}^{riu2}$	98.00	98.00	98.50	96.00	97.50	97.50	98.50	96.00	97.60	98.60	98.40	94.50	95.33	96.85	96.41	94.89
$\{(16,2)\}^{riu2}$	99.50	99.00	99.50	98.50	99.00	99.00	**100**	98.50	97.70	97.85	97.90	96.10	96.63	95.33	96.74	96.20
$\{(24,3)\}^{riu2}$	99.50	99.50	97.00	98.50	**100**	**100**	97.50	98.00	96.85	98.25	97.45	95.50	96.96	97.17	97.39	95.54
$\{(8,1),(16,2)\}^{riu2}$	99.50	99.50	**100**	98.00	99.00	99.00	**100**	98.00	98.45	**99.00**	98.20	96.45	97.71	97.71	97.07	95.22
$\{(8,1),(24,3)\}^{riu2}$	**100**	99.50	**100**	98.00	99.50	99.50	**100**	97.50	98.20	98.65	98.40	96.55	97.83	97.50	98.15	97.28
$\{(16,2),(24,3)\}^{riu2}$	**100**	**100**	**100**	99.00	**100**	**100**	**100**	98.50	98.10	98.05	98.55	96.40	97.61	97.50	98.40	96.41
$\{(8,1),(16,2),(24,3)\}^{riu2}$	**100**	**100**	**100**	**99.50**	99.50	99.50	**100**	98.50	**98.90**	98.35	**98.70**	**98.05**	98.15	**98.59**	**98.70**	97.61
$\{(8,1)\}^{u2}$	99.00	99.00	99.00	98.00	99.00	99.00	99.00	97.50	98.60	98.25	97.35	97.65	**98.80**	98.37	97.93	95.00
$\{(16,2)\}^{u2}$	99.50	99.50	99.50	99.00	99.50	99.50	99.50	98.00	96.95	98.00	97.30	95.65	96.96	97.50	96.52	**98.80**
$\{(24,3)\}^{u2}$	99.50	99.50	-	**99.50**	99.50	99.50	-	**99.00**	96.40	96.60	-	94.65	97.07	96.10	-	95.54

Note: D_M, D_M_C, and D_M/C are different integrations of complemented components of the extended operator xLDP to form the corresponding MDP descriptors. 50-LOO and 50-4fold denote results on 50-class breakdown using leave-one-out and four cross-fold validation respectively. "-" means that the corresponding MDP is not implemented due to the problem of large dimension.

Table 5: Rates (%) on DynTex using MDP, MDP-B descriptors and their multi-scale settings with mappings of $riu2/u2$.

Scheme	DynTex35				Alpha				Beta				Gamma			
$\{(P,R)\}^{riu2/u2}$	D_M	D_M_C	D_M/C	MDP-B	D_M	D_M_C	D_M/C	MDP-B	D_M	D_M_C	D_M/C	MDP-B	D_M	D_M_C	D_M/C	MDP-B
$\{(8,1)\}^{riu2}$	96.86	96.57	97.43	97.43	95.00	93.33	95.00	**96.67**	94.44	95.06	95.68	90.12	92.42	92.05	92.80	91.29
$\{(16,2)\}^{riu2}$	98.00	97.71	98.86	98.57	**98.83**	**98.83**	**98.83**	**96.67**	95.68	95.68	96.30	90.74	**93.18**	92.05	91.67	**93.94**
$\{(24,3)\}^{riu2}$	99.43	99.43	**99.43**	99.14	**98.83**	**98.83**	96.67	**96.67**	96.91	96.91	**96.91**	88.89	**93.18**	92.80	93.18	90.15
$\{(8,1),(16,2)\}^{riu2}$	97.71	98.00	98.86	98.86	**98.33**	**98.33**	**98.33**	**96.67**	95.06	95.06	96.30	90.74	92.80	92.42	92.05	92.05
$\{(8,1),(24,3)\}^{riu2}$	99.43	99.43	**99.43**	98.86	**98.33**	**98.33**	96.67	**96.67**	96.91	97.53	**96.91**	89.51	**93.18**	92.80	91.67	90.91
$\{(16,2),(24,3)\}^{riu2}$	99.43	99.43	**99.43**	98.57	**98.33**	**98.33**	96.67	**96.67**	96.91	**98.15**	96.30	88.89	92.42	92.42	92.80	**93.94**
$\{(8,1),(16,2),(24,3)\}^{riu2}$	99.43	99.43	**99.43**	98.86	**98.33**	**98.33**	96.67	**96.67**	**97.53**	97.53	**96.91**	88.27	92.42	92.42	92.05	93.18
$\{(8,1)\}^{u2}$	97.14	97.14	98.00	98.57	95.00	95.00	95.00	**96.67**	92.59	93.83	93.21	90.12	92.80	92.42	92.80	89.77
$\{(16,2)\}^{u2}$	98.86	99.14	**99.43**	99.14	96.67	96.67	96.67	**96.67**	93.83	94.44	95.06	91.36	**93.18**	92.80	**94.68**	91.67
$\{(24,3)\}^{u2}$	**100**	**100**	-	**99.43**	96.67	96.67	-	95.00	93.83	93.83	-	**92.59**	**93.18**	**93.18**	-	90.91

Note: D_M, D_M_C, and D_M/C are different integrations of complemented components to form the corresponding MDP descriptors. "-" denotes that the corresponding MDP is not implemented due to the problem of large dimension.

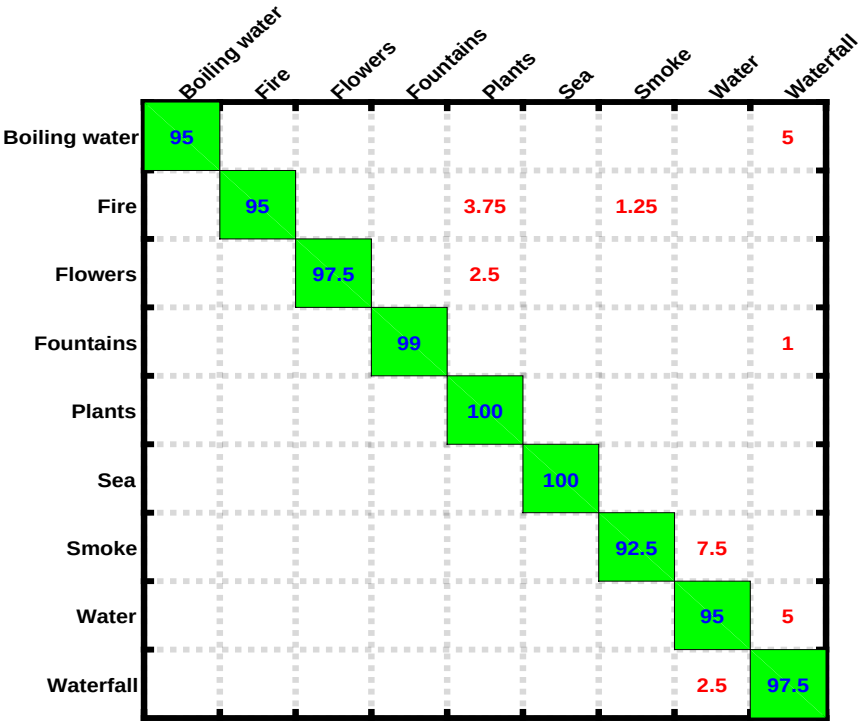

Figure 11: Confusion matrix (%) of MMDP$_{D_M/C}$ on 9-class.

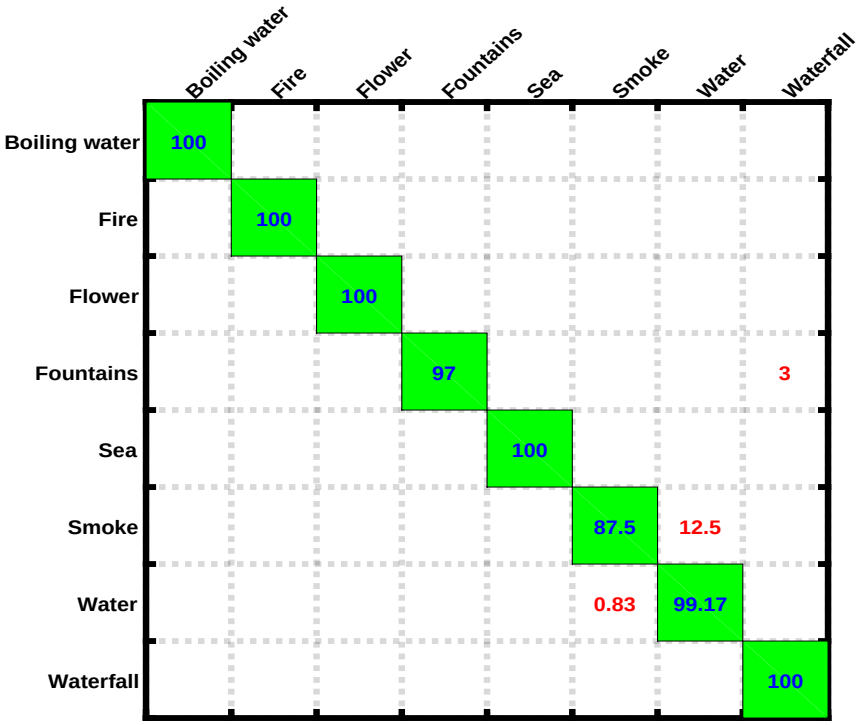

Figure 12: Confusion matrix (%) of MMDP$_{D_M/C}$ on 8-class.

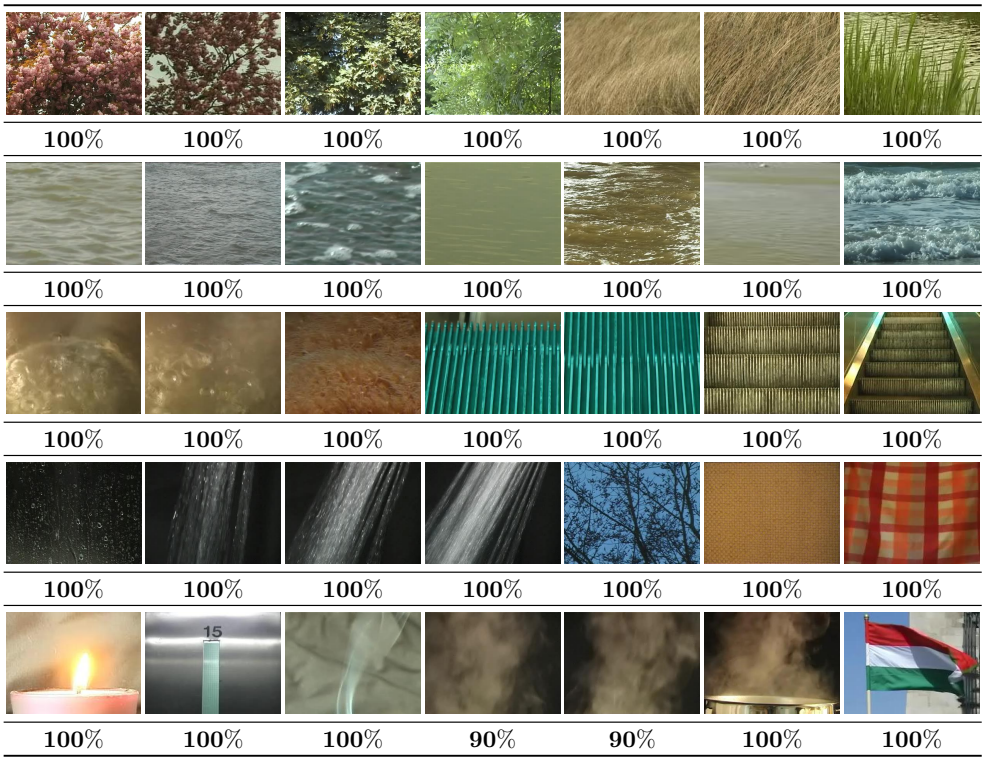

Figure 13: Specific recognition of $\text{MMDP}_{D\_M/C}$ on each class of DynTex35.

c d

Figure 14: Two mutual confused categories in recognition on DynTex35.

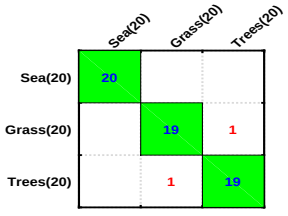

Figure 15: Confusion matrix of $\text{MMDP}_{D\_M/C}$ on Alpha.

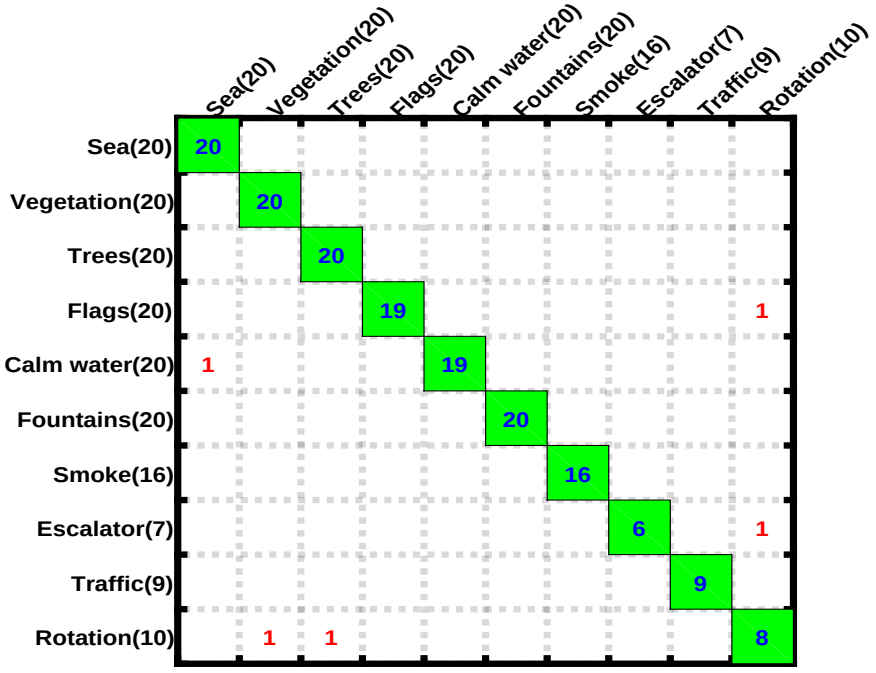

Figure 16: Confusion matrix of MMDP$_{D_M/C}$ on Beta.

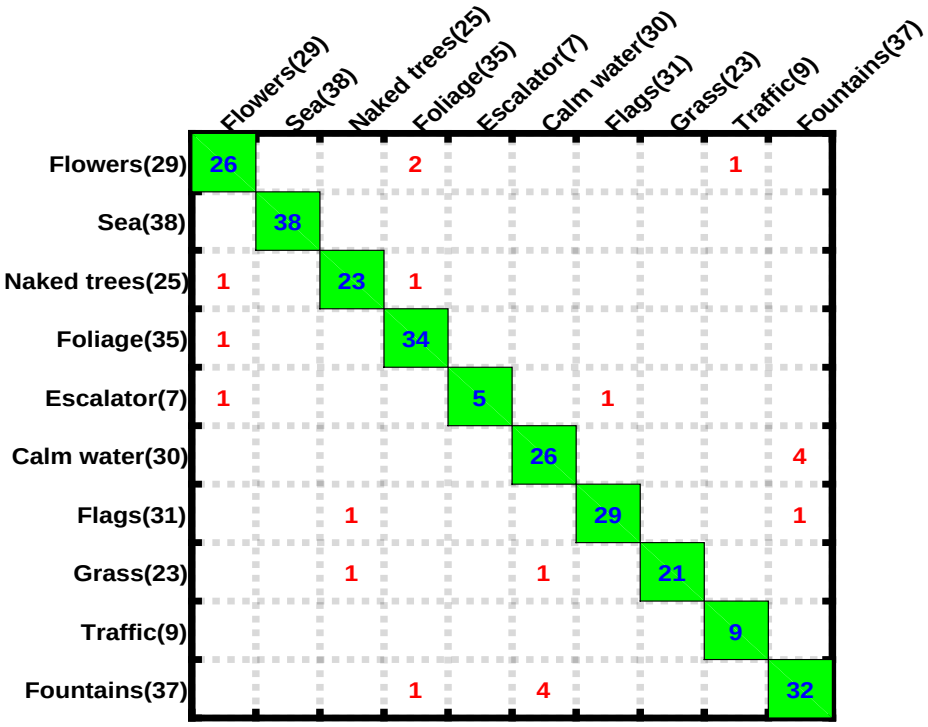

Figure 17: Confusion matrix of MMDP$_{D_M/C}$ on Gamma.

Table 6: Comparison of recognition rates (%) on benchmark DT datasets

Category	Dataset	UCLA				DynTex				
	Encoding method	50-LOO	50-4fold	9-class	8-class	Dyn35	Alpha	Beta	Gamma	Dyn++
Optical-flow-based	FDT [14]	98.50	99.00	97.70	99.35	98.86	98.33	93.21	91.67	95.31
	FD-MAP [14]	99.50	99.00	99.35	**99.57**	98.86	98.33	92.59	91.67	95.69
Model-based	AR-LDS [1]	89.90^{N}	-	-	-	-	-	-	-	-
	KDT-MD [16]	-	97.50	-	-	-	-	-	-	-
	NLDR [19]	-	-	-	80.00	-	-	-	-	-
	Chaotic vector [18]	-	-	85.10^{N}	85.00^{N}	-	-	-	-	-
Geometry-based	3D-OTF [29]	-	87.10	97.23	99.50	96.70	83.61	73.22	72.53	89.17
	WMFS [30]	-	-	97.11	96.96	-	-	-	-	-
	NLSSA [32]	-	-	-	-	-	-	-	-	92.40
	KSSA [32]	-	-	-	-	-	-	-	-	92.20
	DKSSA [32]	-	-	-	-	-	-	-	-	91.10
	DFS [66]	-	**100**	97.50	99.20	97.16	85.24	76.93	74.82	91.70
	2D+T [65]	-	-	-	-	-	85.00	67.00	63.00	-
	STLS [31]	-	99.50	97.40	99.50	98.20	89.40	80.80	79.80	94.50
Filter-based	MBSIF-TOP [27]	99.50^{N}	-	-	-	98.61^{N}	90.00^{N}	90.70^{N}	91.30^{N}	97.12^{N}
	DNGP [15]	-	-	**99.60**	99.40	-	-	-	-	93.80
Local-feature-based	VLBP [4]	-	89.50^{N}	96.30^{N}	91.96^{N}	81.14^{N}	-	-	-	94.98^{N}
	LBP-TOP [4]	-	94.50^{N}	96.00^{N}	93.67^{N}	92.45^{N}	98.33	88.89	84.85^{N}	94.05^{N}
	DDLBP with MJMI [38]	-	-	-	-	-	-	-	-	95.80
	CVLBP [34]	-	93.00^{N}	96.90^{N}	95.65^{N}	85.14^{N}	-	-	-	-
	HLBP [35]	95.00^{N}	95.00^{N}	98.35^{N}	97.50^{N}	98.57^{N}	-	-	-	96.28^{N}
	CLSP-TOP [36]	99.00^{N}	99.00^{N}	98.60^{N}	97.72^{N}	98.29^{N}	95.00^{N}	91.98^{N}	91.29^{N}	95.50^{N}
	MEWLSP [62]	96.50^{N}	96.50^{N}	98.55^{N}	98.04^{N}	**99.71^{N}**	-	-	-	98.48^{N}
	WLBPC [61]	-	96.50^{N}	97.17^{N}	97.61^{N}	-	-	-	-	95.01^{N}
	CVLBC [63]	98.50^{N}	99.00^{N}	99.20^{N}	99.02^{N}	98.86^{N}	-	-	-	91.31^{N}
	MMDP$_{\mathbf{D_M}}$ of $\{(8,1),(16,2),(24,3)\}^{riu2}$	**100**	99.50	98.90	98.15	99.43	98.33	97.53	92.42	95.58
	MMDP$_{\mathbf{D_M_C}}$ of $\{(8,1),(16,2),(24,3)\}^{riu2}$	**100**	99.50	98.35	98.59	99.43	98.33	97.53	92.42	95.70
	MMDP$_{\mathbf{D_M/C}}$ of $\{(8,1),(16,2),(24,3)\}^{riu2}$	**100**	**100**	98.70	98.70	99.43	98.33	96.91	92.05	95.86
	MEMDP$_{\mathbf{D_M/C}}$ of $\{(8,1),(16,2),(24,3)\}^{riu2}$	**100**	**100**	98.90	98.70	**99.71**	96.67	96.91	93.94	96.03
	MMDP-B of $\{(8,1),(16,2),(24,3)\}^{riu2}$	99.50	98.50	98.05	97.61	98.86	96.67	88.27	93.18	95.82
	MLDP-TOP of $\{(8,1),(16,2),(24,3)\}^{riu2}$	97.00	97.00	96.50	96.09	98.86	96.67	88.89	92.80	94.02
Learning-based	DL-PEGASOS [59]	-	97.50	95.60	-	-	-	-	-	63.70
	PI-LBP+super hist [39]	-	**100^{N}**	98.20^{N}	-	-	-	-	-	-
	PD-LBP+super hist [39]	-	**100^{N}**	98.10^{N}	-	-	-	-	-	-
	PCA-cLBP/PI-LBP/PD-LBP [39]	-	-	-	-	-	-	-	-	92.40
	Orthogonal Tensor DL [25]	-	99.80	98.20	99.50	-	87.80	76.70	74.80	94.70
	Equiangular Kernel DL [26]	-	-	-	-	-	88.80	77.40	75.60	93.40
	st-TCoF [21]	-	-	-	-	-	**100***	**100***	98.11*	-
	PCANet-TOP [23]	99.50*	-	-	-	-	96.67*	90.74*	89.39*	-
	D3 [24]	-	-	-	-	-	**100***	**100***	98.11*	-
	DT-CNN-AlexNet [22]	-	99.50*	98.05*	98.48*	-	**100***	99.38*	**99.62***	98.18*
	DT-CNN-GoogleNet [22]	-	99.50*	98.35*	99.02*	-	**100***	**100***	**99.62***	**98.58***

Note: "-" means "not available". Superscript "*" indicates results using deep learning algorithms. "N" indicates rates with 1-NN classifier. 50-LOO and 50-4fold denote results on 50-class breakdown using leave-one-out and four cross-fold validation respectively. Dyn35 and Dyn++ are abbreviated for DynTex35 and DynTex++ datasets respectively. Evaluations of VLBP and LBP-TOP operators are referred to the evaluations of implementations in [35, 21].

Table 7: Recognition (%) on DynTex++ using MDP, MDP-B descriptors and their multi-scale settings with mappings of $riu2/u2$.

Dataset	DynTex++			
$\big\{(P,R)\big\}^{riu2/u2}$	D_M	D_M_C	D_M/C	MDP-B
$\big\{(8,1)\big\}^{riu2}$	93.93	94.28	94.52	92.71
$\big\{(16,2)\big\}^{riu2}$	95.27	94.70	95.18	94.25
$\big\{(24,3)\big\}^{riu2}$	93.92	94.09	93.71	92.16
$\big\{(8,1),(16,2)\big\}^{riu2}$	95.47	95.59	95.56	95.38
$\big\{(8,1),(24,3)\big\}^{riu2}$	94.92	95.10	94.88	94.92
$\big\{(16,2),(24,3)\big\}^{riu2}$	95.37	94.85	95.11	95.07
$\big\{(8,1),(16,2),(24,3)\big\}^{riu2}$	95.58	95.70	95.86	95.82
$\big\{(8,1)\big\}^{u2}$	95.97	**96.51**	**96.18**	**96.51**
$\big\{(16,2)\big\}^{u2}$	**96.37**	96.28	95.92	96.39
$\big\{(24,3)\big\}^{u2}$	95.72	95.68	-	94.79

Note: D_M, D_M_C, and D_M/C are different integrations of complemented components of the extended operator xLDP to form the corresponding MDP descriptors. "-" denotes that the corresponding MDP is not implemented due to the problem of large dimension.

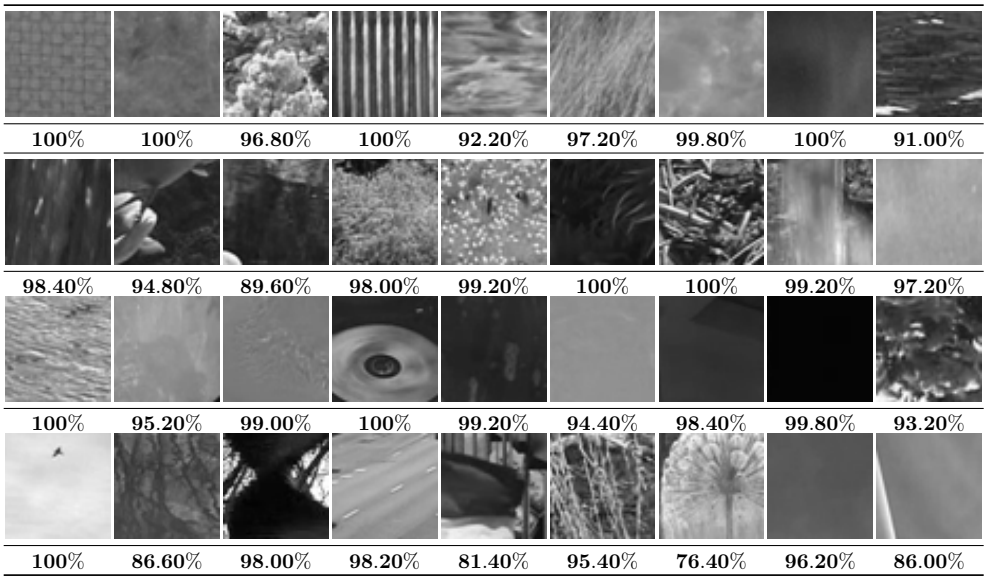

Figure 18: Specific recognition of $\text{MMDP}_{D\_M/C}$ on each class of DynTex++.

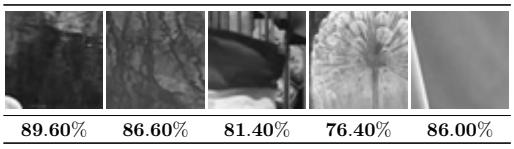

Figure 19: Challenging categories of DynTex++ for $\text{MMDP}_{D\_M/C}$.

Table 8: Classification rates (%) of LDP-TOP descriptor and its multi-scale settings with mappings of $riu2/u2$ on DT datasets without applying the proposed moment volume model.

Dataset	UCLA				DynTex				
$\{(P,R)\}^{riu2/u2}$	50-LOO	50-4fold	9-class	8-class	Dyn35	Alpha	Beta	Gamma	Dyn++
$\{(8,1)\}^{riu2}$	93.00	96.00	96.30	96.09	96.00	**98.33**	87.04	87.12	89.82
$\{(16,2)\}^{riu2}$	96.50	98.00	96.55	**96.74**	97.14	96.67	**90.74**	89.39	91.02
$\{(24,3)\}^{riu2}$	86.00	92.50	93.40	93.48	97.43	96.67	86.42	88.26	87.01
$\{(8,1),(16,2)\}^{riu2}$	97.50	97.00	96.75	95.98	97.71	96.67	89.51	**92.05**	93.61
$\{(8,1),(24,3)\}^{riu2}$	95.50	96.00	96.85	92.72	97.71	96.67	88.27	90.53	92.84
$\{(16,2),(24,3)\}^{riu2}$	95.00	96.50	96.25	95.33	98.57	96.67	87.65	**92.05**	92.52
$\{(8,1),(16,2),(24,3)\}^{riu2}$	97.00	97.00	96.50	96.09	98.86	96.67	88.89	92.80	94.02
$\{(8,1)\}^{u2}$	97.00	97.50	96.40	95.54	97.71	95.00	**90.74**	91.29	95.31
$\{(16,2)\}^{u2}$	**99.00**	**99.50**	**96.90**	96.41	98.86	96.67	88.27	90.91	**95.86**
$\{(24,3)\}^{u2}$	92.00	95.50	92.65	95.00	**99.43**	93.33	90.12	90.53	93.26

Note: 50-LOO and 50-4fold mean rates on 50-class breakdown using leave-one-out and four cross-fold validation respectively. Dyn35 and Dyn++ are shortened for DynTex35 and DynTex++ datasets.

Table 9: Recognition rates (%) of $\text{EMDP}_{D\_M/C}$ descriptor and its multi-scale settings with mapping of $riu2$ on DT datasets.

Dataset	UCLA				DynTex				
$\{(P,R)\}$	50-LOO	50-4fold	9-class	8-class	Dyn35	Alpha	Beta	Gamma	Dyn++
$\{(8,1)\}$	99.50	98.50	98.40	97.07	97.71	95.00	95.68	92.80	95.17
$\{(16,2)\}$	**100**	100	97.15	97.07	99.14	**98.33**	96.91	93.18	95.27
$\{(24,3)\}$	99.50	99.50	98.25	98.04	**99.71**	95.00	96.91	93.56	94.67
$\{(8,1),(16,2)\}$	**100**	**100**	97.90	97.61	99.43	96.67	96.91	93.18	95.90
$\{(8,1),(24,3)\}$	**100**	**100**	98.55	98.26	99.43	96.67	96.91	93.56	95.66
$\{(16,2),(24,3)\}$	**100**	99.50	97.05	97.17	**99.71**	96.67	**97.53**	93.18	95.68
$\{(8,1),(16,2),(24,3)\}$	**100**	**100**	**98.90**	**98.70**	**99.71**	96.67	96.91	**93.94**	**96.03**

Note: 50-LOO and 50-4fold denote rates on 50-class breakdown using leave-one-out and four cross-fold validation respectively. Dyn35 and Dyn++ are shortened for DynTex35 and DynTex++ datasets.

Table 10: Contribution of max-pooling features for the performance (%) of descriptors using settings of $_{D\_M/C}$, and $\{(P,R)\} = \{(8,1),(16,2),(24,3)\}$ with $riu2$ mapping.

Descriptors	DynTex35	Gamma	DynTex++
MMDP	99.43	92.05	95.86
MMDP + "deep" features	99.71	91.30	95.85
MMDP + global features	99.14	**93.94**	95.34
MMDP + "deep" and global features (e.g., MEMDP)	**99.71**	**93.94**	**96.03**

Table 11: Recognition rates (%) of MDP descriptors encoded on filtered videos with supporting elements of $\Omega = \{(14,1),(14,2)\}$.

Dataset	Beta (DynTex)			DynTex++		
$\{(P,R)\}^{riu2/u2}$	D_M	D_M_C	D_M/C	D_M	D_M_C	D_M/C
$\{(8,1)\}^{riu2}$	93.21	93.21	93.83	92.74	93.44	93.76
$\{(16,2)\}^{riu2}$	92.59	92.59	**95.06**	93.88	94.24	93.92
$\{(24,3)\}^{riu2}$	**95.06**	94.44	93.21	94.04	93.96	93.07
$\{(8,1),(16,2)\}^{riu2}$	93.21	93.21	94.44	94.61	94.60	**94.82**
$\{(8,1),(24,3)\}^{riu2}$	93.83	93.83	94.44	94.27	94.54	94.58
$\{(16,2),(24,3)\}^{riu2}$	94.44	**95.06**	94.44	94.49	94.36	94.62
$\{(8,1),(16,2),(24,3)\}^{riu2}$	94.44	93.83	94.44	**95.27**	**94.85**	94.70

Note: D_M, D_M_C, and D_M/C are different integrations of complemented components to form the corresponding MDP descriptors.