Dynamic Scene Recognition using Spatiotemporal based DLTP on Spark

Md Azher Uddin, Mostafijur Rahman Akhond, and Young-Koo Lee
Department of Computer Science and Engineering, Kyung Hee University, Republic of Korea (email: azher006@yahoo.com, mostafij@khu.ac.kr, yklee@khu.ac.kr)
Corresponding author: Young-Koo Lee (e-mail: yklee@khu.ac.kr).
This work was partly supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-2013-1-00717) supervised by the IITP(Institute for Information & communications Technology Promotion), and the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2018-2015-0-00742) supervised by the IITP(Institute for Information & communications Technology Promotion).

ABSTRACT Scene recognition plays a significant role in the fields of pictorial information retrieval and scene understanding. However, the existing methods for scene recognition mostly consider only the images. Moreover, due to the rapid growth of multimedia data, there is a growing demand for distributed processing of large-scale video data. In this work, we present a novel method for dynamic scene recognition from videos that considers spatiotemporal information in a distributed environment. Firstly, to obtain the dynamic information from the videos, we propose the Directional Local Ternary Pattern from Three Orthogonal Planes (DLTP-TOP), which provides valuable information about the nature of dynamic textures. Then, we utilize Apache Spark to conduct distributed computing for large-scale video data. Finally, we employ a convolutional neural network (CNN) to classify the dynamic scenes. The experimental results show that our proposed approach outperforms most of the state-of-the-art methods in terms of accuracy. Moreover, we also experimentally demonstrate the scalability of the proposed method.

INDEX TERMS Dynamic scene recognition, Directional Local Ternary Patterns from Three Orthogonal Planes, Apache Spark, Convolutional Neural Network.

I. INTRODUCTION
Natural scene understanding is a fundamental problem in the fields of computer vision and machine understanding. In general, natural scenes (such as mountains, sunsets, and beaches) include multiple surfaces and objects, organized in a meaningful way. Scene classification is a vital issue for real-world applications, such as, map construction, robot navigation, location and place recommendations for tourists and others. The conventional scene classification approach consists of two key steps: feature extraction and classification. Of these two steps, extracting relevant features from dynamic scenes plays a critical role in dynamic scene recognition. In recent decades, numerous feature extraction techniques have been proposed for scene classification [1-8]. However, most of the existing works perform scene recognition using static images [1-8], while only a few recent works [9-15] have concentrated on videos.

In the past few years, many studies have investigated texture-based feature extraction approaches, due to their outstanding performances and computational efficiency. Local Binary Pattern (LBP) based approaches have emerged as notable texture descriptors due to their effectiveness. The achievements of LBP based methods in different applications have sparked an enormous number of LBP variants that improve robustness, discrimination capability and applicability. However, these existing methods still fail to consider prominent spatial information and/or sensitivity to varying lighting conditions or viewpoint alterations. To resolve these issues, we introduce a novel spatiotemporal-based feature descriptor, for dynamic scene recognition called Directional Local Ternary Patterns from Three Orthogonal Planes (DLTP-TOP), which was inspired by the Local Binary Pattern from Three Orthogonal Planes (LBP-TOP) [23]. The initial LBP [28] was developed for static texture feature analysis in still images, while LBP-TOP was proposed for dynamic texture feature extraction and facial expression classification from videos in [23]. However, the major differences between our proposed DLTP-TOP dynamic feature descriptor and the existing LBP-TOP method are that DLTP-TOP uses three bits instead of two
bits to solve the intensity variations and our feature descriptor also encodes the directional information by employing modified Frei-Chen masks. Furthermore, our method carries out more significant information about the nature of dynamic textures. Recently, many other dynamic feature descriptors [24-26] with different variations have also been proposed, that extract dynamic texture and shape information. However, these existing dynamic descriptors fail to produce consistent patterns in near-uniform areas, where the changes among the center and the neighboring gray level pixels are negligible.

Dynamic feature extraction is one of the most time-consuming tasks in the field of scene understanding. The rapid growth in multimedia data (i.e., images and videos) from numerous sources, has raised the demand for distributed computing to deliver such services efficiently. Recently, several distributed computing methods have been proposed. In [16], the authors performed near-duplicate video detection on Spark by developing a parallel system with multi-feature descriptors that combined the local features SIFT and Local Maximal Occurrence (LOMO) with the global feature Color Name (CN). Deep Convolutional SparkNet was introduced in [17] to perform image retrieval and classification. Here, the authors extended existing convolutional neural networks through batch normalization and a multi-cropping scheme. Other relevant distributed video processing systems were presented in [18-22]. To improve the efficiency of dynamic scene recognition for large-scale video datasets, we utilized the Apache Spark framework. Furthermore, existing dynamic feature descriptors (e.g., LBP-TOP [23]) compute the spatial feature and spatial-temporal co-occurrence features in the XY, XT and YT planes sequentially, which requires more computational time. Therefore, to address the above mentioned issues, in this work, we introduce a novel distributed approach based on a two-level map-reduce framework on Apache Spark to compute the DLTP-TOP features in the XY, XT and YT planes more efficiently. Finally, the extracted spatiotemporal feature is fed to a convolutional neural network (CNN) to classify the dynamic scenes from the videos. In this research, we employed three benchmark datasets: the Maryland dataset [9], the Yupenn dataset [10] and the subset of the YouTube-8M dataset [27] to perform dynamic scene recognition from videos. This paper primarily provides the following contributions:

- We introduce a novel dynamic feature descriptor, Directional Local Ternary Patterns from Three Orthogonal Planes (DLTP-TOP) that produces stable features that are robust to intensity variations. Moreover, the descriptor contains more detailed discriminative features, since it encodes high-order derivative information.
- We perform an extensive experimental analysis to demonstrate the scalability and effectiveness of the proposed method.

In this paper, we further improve the proposed method by considering the generation of the dynamic threshold value used to produce the ternary pattern, and we also propose a holistic and region-based approach for DLTP-TOP. The remainder of this paper is organized as follows. Section II provides a review of the related literature. A brief review of the variants of local patterns is presented in Section III. In Section IV, we introduce the proposed method for dynamic scene recognition using the spatiotemporal feature. The experimental results are presented in Section V. Finally, we draw the conclusions of the paper in Section VI.

II. RELATED WORK

Several studies have been conducted to investigate dynamic scene understanding and classification. In this section, we introduce prior studies for dynamic scene recognition that are related to our study.

Derpanis et al. [9] investigated the role of spatiotemporal-based oriented energy features for scene recognition and introduced a new dataset that included dynamic scenes. The slow feature analysis principle was presented in [10] for dynamic scene classification. Their proposed approach represents principal and stable motion components. Moreover, they also employed a global coding and pooling architecture, and finally, applied a linear support vector machine (SVM) classifier along with a leave-one-out procedure for classification. Feichtenhofer et al. [11] introduced the bag-of-visual-words-based dynamic scene classification that first extracted local feature information in a temporal sliding window. Finally, they proposed dynamic energy pooling approach to aggregate the encoded features. A complementary space-time orientation descriptor was proposed in [12] for dynamic scene classification. The authors extracted the spatial and temporal information from scenes and applied two-dimensional Gaussian third derivative filters in the spatial domain. Finally, they employed a forest-based classifier for scene recognition. Convolutional neural network-based two-level feature extraction was proposed in [13]. Initially, the authors used the first and second order statistics computed over the features generated by ConvNet. Finally, they investigated the performance of the spatial and the temporal transferred ConvNet feature. Later, Feichtenhofer et al. [14] presented dynamically pooled complementary features constructed to recognize dynamic scenes, which constituted an
extension of previous studies [11] [12]. Zheng et al. [15] presented an active discriminative dictionary learning approach based on multicriteria for scene classification; however, their work focuses on static images. Ullah et al. [38] presented 3DPyraNet-F, which learns a spatiotemporal feature. Hong et al. [39] introduced a deep dual descriptor to recognize dynamic scenes, which considers both key-frame- and key-segment-based approaches. Vasudevan et al. [41] recognized dynamic scenes by utilizing the SIFT descriptor to obtain the spatial information and introduced a 5-dimensional motion-flow vector (5DMFV) for the temporal information. In [42], the statistical aggregation performance of pre-trained convolutional neural network (CNN) models was analysed to perform dynamic scene classification.

III. BRIEF REVIEW OF LOCAL PATTERN VARIANTS

In this section, we review several Local Binary Pattern (LBP) variants, including LBP itself and the Local Directional Pattern (LDP).

A. LOCAL BINARY PATTERN (LBP)

The local binary pattern (LBP) was introduced for texture classification in [28]. The basic LBP code for a pixel is computed by,

\[ LBP_{p,r}(x, y) = \sum_{n=0}^{p-1} S (i_n - i_c) \times 2^n \quad (1) \]

where, \( i_n \) denotes the neighboring pixels intensity values, \( i_c \) denotes the center pixel intensity value, while \( p \) and \( r \) represent the number of neighbor pixels and the radius of the circle respectively. The process of computing LBP for a 3 × 3 pattern is presented in Fig. 1.

![FIGURE 1. The process of computing LBP for 3 × 3 pattern.](image)

B. LOCAL DIRECTIONAL PATTERN (LDP)

Jabid et al. [29] introduced a novel feature descriptor that represents image texture by calculating the edge responses while employing Kirsch masks in eight different directions. These Kirsch masks \( KM_{ij} = 0...,7 \) are illustrated in Fig. 2, while the process of computing LDP pattern is demonstrated in Fig. 3. The LDP code of a pixel is computed by,

\[ LDP = \sum_{n=0}^{7} S(|g_k| - |g_k|) \times 2^n \quad (3) \]

\[ g_k = I(x, y) \times KM_{i}(x, y) \quad (4) \]

where, \( I(x, y) \) denotes the gray value of an image and \( g_k \) is the k-th most significant response. Finally, the top-k values of eight directional edge responses are set to 1 and the other responses are set to 0.

![FIGURE 2. Kirsch masks in eight directions.](image)

![FIGURE 3. The process of computing LDP for 3 × 3 pattern.](image)

IV. DYNAMIC SCENE RECOGNITION USING SPATIOTEMPORAL BASED FEATURE ON SPARK

In this section, we explain the proposed method which recognizes dynamic scenes on Spark. Here, we extract low-level information, e.g., directional information in the neighbor pixels to recognize scenes from videos. Fig. 4 demonstrates the proposed approach for dynamic scene recognition from the...
videos on Spark. Initially, all the videos are stored in the Hadoop Distributed File System (HDFS). Subsequently, the video data are loaded into the Spark cluster. Then, partitions of this video dataset are cached in each worker node which are represented by the Spark Resilient Distributed Dataset (RDD). RDD operations are performed in parallel to obtain a distributed result. In each partition, we perform preprocessing and feature extraction using our proposed descriptor, Directional Local Ternary Pattern from three orthogonal planes. Lastly, the spatiotemporal features generated by the feature extraction step are fed to the Convolutional Neural Network (CNN) classifier, which classifies the dynamic video scenes.

The existing dynamic feature descriptors [23-26] compute spatial features and spatio-temporal co-occurrence features in a sequential manner which requires more computation than the volume local binary pattern (VLBP) operator [23]. However, VLBP generates a large feature vector when the number of neighboring pixels employed is increased. To address the above issues, we introduce a novel framework that uses a two-level map-reduce operation. Our proposed DLTP-TOP feature descriptor, extracts directional features in the neighboring area, using a two-level map-reduce operation. Firstly, all the videos are loaded into the Spark RDD. These videos are processed in a distributed manner by the first mapper function. Here we pre-process the videos which involves frame extraction, frame conversion from RGB to grayscale and frame resizing. Then, the spatial features and spatiotemporal co-occurrence features in three orthogonal planes are computed by the second mapper function. Then, the reducer combines the features produced by the second mapper function for each video. Finally, the last reducer function combines the feature vectors generated from all videos and stores them on HDFS. Fig. 5 depicts the framework for dynamic feature extraction based on DLTP-TOP on Spark employing a two-level map-reduce operation.

**FIGURE 5.** A two-level Map-reducer for dynamic feature extraction on Spark.

**FIGURE 6.** Modified Frei-Chen masks in eight directions.

**A. DIRECTIONAL LOCAL TERNARY PATTERNS FROM THREE ORTHOGONAL PLANES (DLTP-TOP)**

The main basic descriptor for texture information extraction is the Local Binary Pattern (LBP) [28]. The key idea of LBP is that it assesses the picture of an image or video frame by thresholding a circular neighborhood region. However, the drawback of the LBP operator is that- LBP codes are susceptible to noise, even a slight alteration in the neighboring pixel intensities can completely change the resultant binary code. Therefore LBP fails to produce consistent patterns in near-uniform areas, where the changes between the center and the neighbor gray level pixels are minuscule. Moreover, LBP can extract only static texture information. Therefore, to obtain the dynamic texture information from the videos, VLBP [23] and LBP-TOP [23] were proposed. However, still these methods have issues with varied illumination and noise. To resolve these issues, we introduce the spatiotemporal-based Directional Local Ternary Patterns from Three Orthogonal Planes (DLTP-TOP), which takes advantage of the more robust directional information by computing the edge responses in eight different directions rather than simply comparing intensity values. Furthermore, a three-level coding scheme is employed to differentiate between highly-textured and smooth areas, which also helps guarantee the production...
of robust texture patterns under varying illumination conditions. DLTP-TOP deals with both spatial and temporal information. An LDP variant called the Local Directional Ternary Pattern (LDTP) was introduced in [30-31], however, these approaches are different from ours in terms of applying the compass mask and producing the bit vector to generate the feature. Furthermore, we generate an adaptive threshold value to compare the edge responses. Finally, we employ our proposed descriptor in three orthogonal planes (represented as XY-DLTP, XT-DLTP and YT-DLTP) to extract both spatial and spatiotemporal information, whereas LDTP [30-31] extracts only the spatial information.

Firstly, DLTP-TOP applies the proposed modified Frei-Chen masks, $MFCM_i(x, y)$ in eight different directions to obtain the edge responses, which are represented by $D_{ER}$. The modified Frei-Chen masks are an extended version of Frei-Chen masks [32] and are presented in Fig. 6. The main advantage of using the modified Frei-Chen masks over the original Frei-Chen masks [32] is that they are computationally efficient because they do not require any square root computation. Next, we obtain the average value of the edge responses $D_T$, which is used as a threshold value for comparison with the individual edge responses. Furthermore, we obtain an adaptive threshold value $T$ by taking the square-root of the center pixel value, $T = \sqrt{CenterPixel\_value}$, which is then applied to produce the ternary pattern. Fig. 7 illustrates the process used in DLTP-TOP, which is computed by applying equations 5-8.

$$DLTP = \sum_{n=0}^{p-1} S(D_{ER} - D_T) \times 2^n$$

$$D_{ER} = I(x, y) \times MFCM_i(x, y)$$

$$D_T = \frac{\sum_{n=0}^{p-1} D_{ER}}{p}$$

$$S(x) = \begin{cases} 1, & x > T \\ -1, & x < -T \\ 0, & \text{otherwise} \end{cases}$$

The DLTP values are calculated on three different planes and then combined into a single histogram, which is used as input to the convolutional neural network (CNN) classifier. The feature size of the DLTP-TOP is $3\times2\times256$, where two histograms are produced by upper and lower patterns.

### B. CONVOLUTIONAL NEURAL NETWORK FOR CLASSIFICATION

The spatiotemporal features extracted by DLTP-TOP from the scene videos are employed in the Convolutional Neural Network (CNN) to classify the dynamic scenes. CNNs are mainly used to perform deep feature learning from images and have been demonstrated to perform better than other deep learning-based approaches [5, 13, 33, 34]. Similar to [33], we also employed a 1-D CNN in our work, however, the structure of our 1-D CNN is different from that of [33]. Usually, 1-D CNNs are computationally efficient and easier to train with less number of epochs than typical 2-D CNN’s [40]. Thus, a 1-D CNN can achieve the fast classification of dynamic scenes using extracted features. Fig. 8 demonstrates the structure of our 1-D CNN for dynamic scene classification.

The DLTP-TOP features from a video of 60 frames are arranged as a $60\times1536$ vector and used as input to the CNN. Our CNN includes four convolutional layers, four max-pooling layers, and one fully connected layer. In the first and second convolutional layer, 16 convolution kernels with sizes of $1\times7$ and $1\times6$ are convolved with the processed feature, while in the third and fourth convolution layers, 32 convolution kernels with sizes of $1\times5$ and $1\times4$ are convolved with the processed feature. In contrast, the first, third and fourth pooling layers perform $1\times2$ sliding window based max-pooling.
pooling, while the second pooling layer employs $1 \times 4$ sliding window based max-pooling. The weight and bias values are randomly initialized when executing the CNN on the spatiotemporal features, and the weight and bias values of all layers are updated by employing the backpropagation algorithm.

![Architecture of the 1-D CNN applied in our work.](image)

**V. EXPERIMENTAL RESULTS AND ANALYSIS**

**A. EXPERIMENTAL SETUP**

To investigate the scalability and performance of DLTP-TOP on Spark, we employed a Hadoop Spark cluster and three benchmark datasets. In the experiments, we adopted a cluster with 5 nodes: one is the master node and the other four are worker nodes. Each node had the same configuration—4 cores running at 3.0 GHz and 16 GB memory. The Hadoop version was 2.7.1 and the Spark version was 1.6.2. We evaluated the proposed approach on the Maryland dataset [9], YUPENN dataset [10] and a subset of the YouTube 8 M dataset [27]. In the experiment, for all three datasets, we employed 5-fold cross-validation using 70 percent of the video data for training and the remaining 30 percent for testing.

**B. MARYLAND DATASET**

The Maryland dataset was first proposed in [9]. The dataset includes 13 dynamic scene categories and a total of 130 videos. The videos have large variations in illumination, different viewpoints and, most importantly, camera-induced motion. The categories are avalanche, boiling water, chaotic traffic, forest fire, fountain, iceberg collapse, landslide, smooth traffic, tornado, volcano eruption, waterfall, waves and whirlpool. Fig. 9 shows some sample frames from the Maryland dataset.

![An example of Maryland dataset.](image)

**C. YUPENN DATASET**

The Yupenn stabilized dynamic dataset was first introduced in [10]. In these videos, the camera remains in a static position. The dataset consists of 14 scene groups, and each group contains 30 videos. The groups are beach, forest fire, lightning storm, elevator, fountain, ocean, railway, sky-clouds, snowing, highway, street, waterfall, rushing river, and...
windmill farm. The size of the videos is $250 \times 370$ (pixels) $\times$ 145 (frames). Fig. 10 shows some example frames from the Yupenn dataset.

D. YOUTUBE 8M DATASET
The YouTube-8M is a benchmark dataset for video understanding and was presented in [27]. In our work, we used a subset of the YouTube 8 M dataset. From the travel section of the YouTube 8 M dataset, we selected 7 categories: beach, lake, amusement park, sunset, zoo, airport and street. We employed a total of 700 videos—100 videos from each category. The videos have different viewpoints and include camera motion. Fig. 11 shows some sample frames from the YouTube-8 M dataset.

E. RESULTS
We conducted numerous experiments to investigate the scalability of the proposed method on the Apache Spark framework. Fig. 12 depicts the experimental results for the proposed dynamic descriptor (DLTP-TOP) when employing different numbers of nodes. This experiment was designed to demonstrate the scalability of the proposed method. From the experiment, we can see that when using two nodes, our feature extraction technique speeds up by almost 1.5x, while introducing four nodes does not result in a large speed improvement (3.3x). However, using three nodes achieves an averaged speedup of 2.6x.
Figs. 13, 14, and 15 demonstrate the performance of the proposed method for the three datasets. The average recognition rates when using the proposed method on the Maryland dataset, Yupenn dataset and YouTube-8 M dataset were 84.6%, 97.0% and 94.14%, respectively. For the Maryland dataset, the avalanche, chaotic traffic, tornado, waterfall and waves categories resulted in better performances than did the other categories, while the iceberg collapse and landslide categories showed very poor results. These poor results might be due to the substantial camera motions in these scene categories. In contrast, on the Yupenn dataset, the categories beach, elevator and rushing river achieved the best results. However, the fountain and windmill farm categories showed low accuracy. Finally, for the YouTube-8 M dataset, the sunset category achieved the best accuracy, while the airport category showed poor performance due to the complexity of those videos.

We also performed experiments using a region-based approach and a holistic-based approach. In the region-based method, each frame of the video scene is divided into six different regions. Then, DLTP-TOP is employed on each region, and the resultant feature histograms of all the regions are concatenated to produce a single histogram, which is used as the input to the CNN. In contrast, for the holistic approach, DLTP-TOP was employed on each video. The region-based approach for DLTP-TOP is demonstrated in Fig. 16, while a comparison of DLTP-TOP’s performances using the holistic-based approach and the region-based approach is shown in Fig. 17. From the experiment, we can see that the holistic-based approach outperforms the region-based approach by a firm margin because a dynamic scene represents global information about a video obtained by the holistic-based approach, whereas the region-based method determines only the influence of the micro-information obtained from the different regions. Furthermore, in the region-based approach, the increased number of regions cause the dimensionality of the feature vector to increase, which leads to many ambiguities.

**FIGURE 16.** Region based approach for the DLTP-TOP.

**FIGURE 17.** Comparison of the proposed method in holistic approach and region based approach.

**FIGURE 18.** Comparison of the proposed method with existing dynamic feature extraction methods on different datasets.
Fig. 18 depicts a comparison of the proposed method with existing dynamic feature extraction methods on numerous datasets. In this experiment, we employed 1-D CNN as a classifier for all the dynamic feature descriptors. From this figure, we can see that our proposed method outperforms the existing dynamic feature extraction methods by some distance, since the proposed dynamic feature descriptor obtains more robust information from the scenes. On the Maryland dataset, the average accuracy of the proposed approach is superior to VLBP [23] and LBP-TOP [23] by 25.4% and 23.2% respectively. On the other hand, for the Yupenn dataset, our approach outperforms the 2-D Histogram Fourier LBP-TOP (2DHFLBP-TOP) [24] by 5.3%. Lastly, on the YouTube-8M dataset, DLTP-TOP significantly outperforms the Histogram of Oriented Gradients from Three Orthogonal Planes (HOG-TOP) [26] by 5.04%.

<table>
<thead>
<tr>
<th>Method</th>
<th>Maryland dataset</th>
<th>Yupenn dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOE [9]</td>
<td>43</td>
<td>81</td>
</tr>
<tr>
<td>SFA [10]</td>
<td>60</td>
<td>85</td>
</tr>
<tr>
<td>CSO [12]</td>
<td>68</td>
<td>86</td>
</tr>
<tr>
<td>DPCF [14]</td>
<td>80</td>
<td>99</td>
</tr>
<tr>
<td>Single-Frame CNN [35]</td>
<td>78</td>
<td>87</td>
</tr>
<tr>
<td>t-TCoF [13]</td>
<td>66.1</td>
<td>97.8</td>
</tr>
<tr>
<td>st-TCoF [13]</td>
<td>88.4</td>
<td>97.8</td>
</tr>
<tr>
<td>3DPyraNet-F [38]</td>
<td><strong>94.8</strong></td>
<td>93.6</td>
</tr>
<tr>
<td>D3 [39]</td>
<td>90.7</td>
<td>98.3</td>
</tr>
<tr>
<td>D3 [39]</td>
<td>92.3</td>
<td><strong>99</strong></td>
</tr>
<tr>
<td>SIFT+5DMFV [41]</td>
<td>-</td>
<td>85.61</td>
</tr>
<tr>
<td>Hybrid-CNN [42]</td>
<td>93.85</td>
<td>98.1</td>
</tr>
<tr>
<td>DLTP-TOP</td>
<td>84.6</td>
<td>97</td>
</tr>
</tbody>
</table>

Table I shows a comparison of the proposed method with other state-of-the-art methods. Our proposed approach greatly outperforms the Spatiotemporal Oriented Energy (SOE) [9], Slow Feature Analysis (SFA) [10], Complementary Space-time Orientation (CSO) [12], Bag of Space-time Energy (BoSE) [11], Dynamically Pooled Complementary Features (DPCF) [14], Single-frame CNN [35] and temporal Transferred ConvNet Feature (t-TCoF) [13] on the Maryland dataset, however, the spatial and temporal Transferred ConvNet Feature (st-TCoF) [13], Hybrid-CNN [42], 3DPyraNet-F [38] and Deep dual descriptor (D3) [39] achieved better performances than our approach. These results occurred because these approaches have more discriminative power; moreover, the proposed approach does not consider camera motion. For the Yupenn dataset, the proposed approach shows superior performance than SOE [9], SFA [10], CSO [12], BoSE [11], SIFT+5DMFV [41], Single-frame CNN [35] and 3DPyraNet-F [38]. Furthermore, DLTP-TOP’s performance is competitive with that of t-TCoF [13], Hybrid-CNN [42] and the Dynamic Deep dual descriptor (D3) [39] on the Yupenn dataset. However, on the Yupenn dataset, DLTP-TOP achieves lower accuracy than DPCF [14] because DPCF collects its primitive feature descriptors in spatiotemporal grids to obtain the neighbourhood structure. Moreover, it also considers colour properties.

We also performed experiments using different classifiers. Fig. 19 illustrates the results of using DLTP-TOP with different classifiers. From this figure, we can see that the CNN performs better than the other classifiers due to its discriminating power, while the SVM [37] classifier performed better than the Random Forest [36] classifier.

In this study, we selected the threshold value dynamically. After extensive experimental analysis, we set the value of threshold $T$ to the square root of the center pixel intensity. Fig. 20 clarifies the reason for why we elected to set threshold $T$ to the square root of the center pixel value. The average threshold represents the average of the neighboring pixels, while the median threshold represents the median of the absolute value.
VI. CONCLUSION
In this paper, we introduced a new approach for dynamic scene recognition from videos in a distributed environment. We introduced the dynamic descriptor DLTP-TOP to extract dynamic spatiotemporal features from videos. Finally, we used a CNN to classify dynamic video scenes. We conducted experiments to investigate applying DLTP-TOP in two different scenarios: region-based and holistic-based. Moreover, we also compared our results with existing dynamic feature extraction methods and with other state-of-the-art methods to demonstrate the effectiveness of our novel approach. Finally, we demonstrated the scalability of our proposed dynamic feature descriptor.

ACKNOWLEDGMENT
This work was partly supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-1-00717) supervised by the IITP(Institute for Information & communications Technology Promotion), and the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2018-15-0-00742) supervised by the IITP(Institute for Information & communications Technology Promotion).

REFERENCES


