LBP-flow and hybrid encoding for real-time water and fire classification

Konstantinos Avgerinakis Panagiotis Giannakeris Alexia Briassouli
Anastasios Karakostas Stefanos Vrochidis
Ioannis Kompatsiaris
Centre for Research and Technology Hellas (CERTH) - Information Technologies Institute (ITI)
koafgeri@iti.gr
giannakeris@iti.gr abria@iti.gr stefanos@iti.gr akarakos@iti.gr ikom@iti.gr

Abstract

The analysis of dynamic scenes in video is a very useful task especially for the detection and monitoring of natural hazards such as floods and fires. In this work, we focus on the challenging problem of real-world dynamic scene understanding, where videos contain dynamic textures that have been recorded in the "wild". These videos feature large illumination variations, complex motion, occlusions, camera motion, as well as significant intra-class differences, as the motion patterns of dynamic textures of the same category may be subject to large variations in real world recordings. We address these issues by introducing a novel dynamic texture descriptor, the "Local Binary Pattern-flow" (LBP-flow), which is shown to be able to accurately classify dynamic scenes whose complex motion patterns are difficult to separate using existing local descriptors, or which cannot be modelled by statistical techniques. LBP-flow builds upon existing Local Binary Pattern (LBP) descriptors by providing a low-cost representation of both appearance and optical flow textures, to increase its representation capabilities. The descriptor statistics are encoded with the Fisher vector, an informative mid-level descriptor, while a neural network follows to reduce the dimensionality and increase the discriminability of the encoded descriptor. The proposed algorithm leads to a highly accurate spatio-temporal descriptor which achieves a very low computational cost, enabling the deployment of our descriptor in real world surveillance and security applications. Experiments on challenging benchmark datasets demonstrate that it achieves recognition accuracy results that surpass State-of-the-Art dynamic texture descriptors.

1. Introduction

Dynamic scene analysis is a very useful task in numerous real world applications, such as security and disaster management, where surveillance videos can be used for the early detection, classification and monitoring of natural hazards like floods and fires. The analysis of such videos is considered of utmost importance during natural disasters, since it can improve situational awareness by providing early detection of floods and fires. The motions in dynamic scenes are complex, highly non-rigid, with many auto-correlations and occlusions that render their analysis costly in terms of computation and memory requirements. Surveillance videos of interest in natural disasters contain dynamic textures (fire, smoke, flood) for which various methods have been deployed either to model them using statistical and dynamic system modeling, or to represent them with a local descriptor, to discriminate between them and classify them. Dynamic textures also appear in videos containing small motions, analyzed for emotion/facial recognition using approaches based on Local Binary Patterns (LBP), which have led to accurate classification in several applications, such as clothing (fabric in motion), facial expressions recognition and non-rigid pedestrian motion estimation.

In this work, we introduce the LBP-flow, which extends and re-designs the LBP to the spatio-temporal domain, by applying it both to illumination values and optical flow estimates in highly complex dynamic textures. Additionally, inspired from recent work in hybrid classification architectures [4], we propose a novel encoding scheme that combines Fisher encoding with Neural Network in order to recognize crisis events or classify dynamic textures in videos samples. As shown in the experimental section (Sec. 4), not only our LBP-Flow descriptor is capable of achieving near real-time performance at a low computational cost, and with improved accuracy in comparison to existing dynamic texture recognition methods, but also lead to State-of-the-Art accuracy rates when encoded with the suggested hybrid architecture. It is thus a promising solution for analyzing dynamic scenes in security and surveillance applications, as it can provide early warning and timely detection of events such as smoke, fire, flooding, among others.

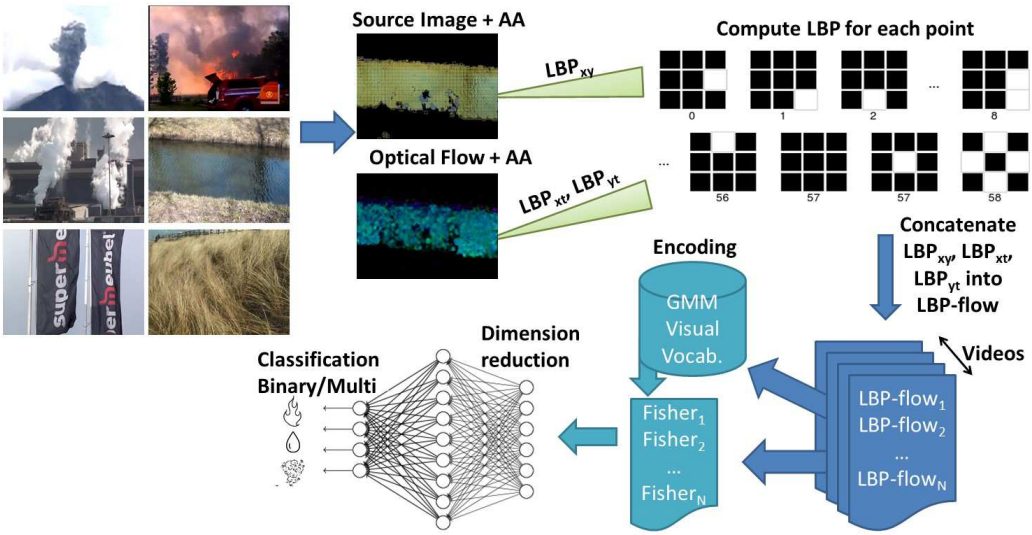

Figure 1. Block diagram of the overall framework: Activity Areas(AAs) are applied on each video frame and optical flow matrix in order to sample and aggregate LBP descriptors in the three spatio-temporal domains (x,y,t). LBP-Flow is constructed by concatenating them and fed to a GMM to extract the visual vocabulary. Fisher vectors are extracted for each video sample in the training and testing set based on this vocabulary and the results are given in a Neural Network to reduce the dimensionality of the descriptor and make the descriptors even more discriminative.

2. Related Work

Dynamic texture recognition methods can be separated into roughly two categories. The first one involves global modelling of video sequences, based on Linear Dynamical Systems (LDS) [3], [7], [12]. More recently, LDS have been extended to a stabilized higher order LDS (shLDS) in [6], whose multidimensional time series are transformed into Grassmannian points to form a novel descriptor, the Histograms of Grassmannian Points (HoGP). The resulting algorithm [6] achieved very high accuracy rates both in the recognition of dynamic textures, as well as of more rigid motions that characterize human actions. However, the conversion of the multidimensional motion time-series to the Grassmannian manifold is computationally costly, and thus not appropriate for security and real-time surveillance purposes.

While LDS models seem quite promising for representing dynamic textures, their application to classifying the wider set of motion patterns found in dynamic scenes has been shown to perform poorly [16]. The complex, stochastic character of dynamic textures makes their precise modeling very challenging, so a new category of dynamic texture representations is created, based on local, spatio-temporal features that describe moving texture dynamics by estimating local variations and statistics of intensity and optical flow values. Early techniques involved the accumulation of local spatio-temporal features to describe texture dynamics using appearance features like GIST [13], motion histograms such as HOF [10], and spatio-temporal oriented energy features [5], which describe energy changes in the intensity of sampled images. However, the coarse quantization of GIST and the orientation invariance of HOF do not allow them to detect dynamic textures with accuracy. The computational cost of spatio-temporal oriented energy features [5] makes them impractical for surveillance and security purposes. Spatial energies are also used in a Bag of Words (BoW) framework in [8], where they are combined with color information and a space-time pyramidal representation, leading to very accurate dynamic scene recognition on challenging benchmark datasets. However, this method is computationally costly due to the estimation of spatio-temporal energies, their encoding and learning by linear SVM, making it inappropriate for real time, or near-real time, applications.

Accurate texture classification has been achieved in images using Local Binary Patterns (LBP's), whose promising results have led to a number of extensions of it as a dynamic texture descriptor. Volume Local Binary Patterns (VLBP) [19] and LBP-TOP [18] are among the earlier methods, however they can easily reach a dimensionality of 2^{14} to 2^{26}, which is impractical in real-world applications involving large amounts of data that are to be processed in near real time. Mettes *et al*. recently introduced a hybrid spatio-temporal extension of LBP in [11], which stacks the descriptor in time to obtain temporal information. It complements this descriptor with a Fourier temporal statistic, but requires large training datasets. The method achieved very high accuracy rates when discriminating between wa-

ter and non-water scenes, however it missed many video frames containing a crisis (i.e. fire, smoke, flood), as it would need related samples in the training data in order to detect them. Thus, the method of [11] cannot be realized in real world security/surveillance scenarios, where crisis events may be unknown, with no related training samples.

In this work, we overcome the limitations of the State-of-the-Art (SoA) by extending LBP over time to build our LBP-Flow descriptor. LBP-Flow follows the same concept as VLBP and LBP-TOP by its application to the temporal domain. However, it goes beyond both of these methods as:

1. it is applied to the optical flow values, whereas VLBP, LBP-TOP are applied to intensity differences over short temporal windows, which cannot sufficiently capture the motion taking place, or may be sensitive to local noise, such as camera noise/motion and compression artefacts

2. it addresses the issue of the high-dimensionality of the descriptor by effectively reducing its size without losing discriminative ability.

3. it leverages both shallow and deep attributes by combining Fisher and Neural Networks in a robust, SoA scheme that can accurately recognize and discriminate fire, smoke and water patterns from other moving pattern, and also perform multi-class recognition at no further computational burden.

3. Methodology

The LBP-Flow descriptor introduced in this work, is a spatio-temporal descriptor tailored to the nature of dynamic scenes, as it encodes not only appearance, but also motion-induced variations as texture. This results in the improved discrimination between dynamic textures with different appearance, such as water and vegetation, but also with different motion statistics, such as rapidly versus slowly flowing water. Inspired by the success of LBP in face recognition [1] and pedestrian detection [17] where it is combined with HOG appearance descriptor, we extend it over time to encode the optical flow values estimated in videos of dynamic textures. Furthermore, inspired from the recent successful combination of intermediate(Fisher) and high level encoding schemes(NN) in [4], we propose a hybrid classification architecture that leverages both the discriminative and computational power of the proposed descriptor.

As a first step, the background is detected using Activity Areas (AA) [2] separating background from foreground motion, and interest points are sampled on a dense grid in each video frame. Quantized-LBP descriptors are then estimated for each sampled grid point over a temporal window of W_{LBP} video frames in order to form the LBP-Flow, which will characterize the spatio-temporal dynamics of the region. Fisher encoding [14] is used to compute the first and second order differences from a spatio-temporal dynamics visual vocabulary, and fed to a Neural Network which reduces the dimensionality and improves the discriminability of the descriptor. Both binary and multi-class recognition is achieved by using either Sigmoid or Softmax as a final NN layer respectively. Figure 1 shows the framework of the overall system.

3.1. Activity Areas

Activity Areas (AA) are binary masks [2] that separate background from foreground motion, based on the assumption that flow estimates originate either from actual motion, or noise, e.g. from the video capture or compression process. These two hypotheses can be formulated as:

$$H_0 : u_k^0 = z_k(\overline{r})$$
$$H_1 : u_k^1 = u_k(\overline{r}) + z_k(\overline{r}),$$

where $\overline{r} = (x, y)$ is the pixel under consideration, $u_k(\overline{r})$ its actual motion value and $z_k(\overline{r})$ is induced by noise. As shown in [2], $z_k(\overline{r})$ can be modelled by a Gaussian pdf when optical flow is taken into account for the computation of motion vectors. Under the Gaussianity assumption for the noisy flow values, and accumulating these motion vectors over a temporal window W_{AA}, we can easily eliminate them by estimating their Kurtosis, which is equal to zero for Gaussian data:

$$G_2[u_k(\overline{r})] = \frac{3}{W_{AA}(W_{AA}-1)} \sum_{k=1}^{W_{AA}} u_k(\overline{r})^4 -$$
$$- \frac{W_{AA}+2}{W_{AA}(W_{AA}-1)} (\sum_{k=1}^{W_{AA}} u_k(\overline{r})^2)^2.$$

Thus the AA has zero values at pixels where the kurtosis values tend to zero, as in that case the motion is noise-induced:

$$AA(\overline{r}) = \begin{cases} 0 & \text{if } G_2(\overline{r}) < th_{AA} \\ 1 & \text{else} \end{cases}$$

where th_{AA} is statistically determined equal to $2 \cdot 10e^{-2}$ based on the experiments that we performed on Dyntex dynamic texture dataset.

3.2. LBP-Flow computation

After extracting the AA in each video frame, we apply a dense grid to sample a set of interest points from the regions of interest(i.e. where $AA(\overline{r}) = 1$), and use them to define LBP-Flow. LBP-Flow builds upon the original LBP, which is defined at each pixel $\overline{r}$ by the difference between its intensity value $f(\overline{r})$ and that of neighboring pixels $(\overline{r}_p)$ within a radius R. In this work, LBP-Flow is extended to

include the values of the optical flow around pixel $\overline{r}$ so as to include motion information. Thus:

$$LBP_{P,R}(\overline{r}) = \sum_{p=0}^{P-1} s(f(\overline{r}) - f(\overline{r}_p))2^p,$$

where $f(\overline{r})$ corresponds to its intensity or optical flow value, P represents the number of neighbour points around each sampled interest point $\overline{r}$, and R is the radius of its neighbourhood, with the neighbouring points around pixel $\overline{r}$ at coordinates $\overline{r}_p = (r_x + R\cos(2\pi p/P), r_y - R\sin(2\pi p/P))$. The threshold function $s(z)$ of LBP is given by:

$$s(z) = \begin{cases} 0 & z < 0 \\ 1 & z \geq 0. \end{cases}$$

The novelty of LBP-Flow is that it represents both intensity and optical flow variations over space and time. In our LBP representation, texture is spatially represented by calculating local binary patterns in two directions, on the $x-y$ axes, as in the original LBP. A novel representation of motion as a temporal texture is introduced by calculating LBP over the optical flow values in the x and y directions, $x-t$ and $y-t$ respectively. This inclusion of motion information in the LBP-Flow representation enriches the spatial textual characteristics with a binary representation of the dynamic texture motion, introducing redundancies in the resulting descriptor, which render it more robust. By this procedure, we obtain the LBP descriptors for appearance and motion, namely LBP_{xy}, LBP_{xt} and LBP_{yt} respectively.

The dimensionality of the resulting LBP_{xy}, LBP_{xt} and LBP_{yt} is then reduced by using a variation of the original LBP descriptor, the uniform quantized LBP descriptor [18], which uses 58 bins to describe a 3×3 area around each interest point instead of the 256 bins commonly used. To achieve this, uniform quantized LBP takes into account that there is one quantized pattern for each LBP, with exactly one transition from 0 to 1, and one from 1 to 0 when scanned counter-clockwise. Thus, the uniform quantized LBP represents the same pattern with a descriptor whose dimension is equal to $1/4$ of the original LBP [18]. The LBP-Flow descriptor is constructed by accumulating LBP_{xy}, LBP_{xt} and LBP_{yt} over $W_{LBP} = 30$ video frames and concatenating them into a single vector of 5220 bins whose dimensions are much lower than those of the latest SoA LBP [18], which reaches dimensions of 2^{14} and 2^{26}. The representation framework of the LBP-Flow is shown in Figure 2.

3.3. Fisher encoding

LBP-Flow includes both spatial and spatiotemporal information, which introduce redundancies that are expected to increase its robustness, however it remains a low-level local representation, so it may still contain noise-induced artefacts. To eliminate this noise, we build a visual vocabulary for LBP-Flow using Gaussian Mixture Model (GMM) clustering followed by Fisher encoding [14]. The LBP-Flow is represented as $\bar{x}_i \in R^D; i = \{1, ..., N\}$, and the pre-trained visual vocabulary is defined by the corresponding mean, covariance and prior probability weights for L Gaussian models $\{\mu_j,\ \Sigma_j, \pi_j;\ j \in R^L\}$ of $\bar{x}_i$. Fisher encoding is then based on:

$$f_{1j} = \frac{1}{N\sqrt{\pi_j}} \sum_{i=1}^{N} q_{ij} \Sigma_j^{-1/2} (\bar{x}_i - \bar{\mu}_j),$$

$$f_{2j} = \frac{1}{N\sqrt{2\pi_j}} \sum_{i=1}^{N} q_{ij} [(\bar{x}_i - \bar{\mu}_j) \Sigma_j^{-1} (\bar{x}_i - \bar{\mu}_j) - 1],$$

where q_{ij} is the Gaussian soft assignment of descriptor x_i to the j-th Gaussian. These distances are concatenated to form the final Fisher vector $F_X = [f_{11}, f_{21}, ..., f_{1L}, f_{2L}]$, which characterizes the dynamic texture in each video. The vectors are then normalized using a Hellinger kernel $K(g, h)$, which leads to SoA results [14]. For two Fisher vectors $g \in R^{2KD}$ and $h \in R^{2KD}$, $K(g, h)$ is computed by:

$$K(g, h) = \sum_{j=1}^{L} sign(g_j) sign(h_j) \sqrt{\|g_j\| \cdot \|h_j\|},$$

where L is the number of Gaussians, D the dimensionality of the descriptor and $2LD$ the final Fisher vector size.

3.4. Neural Network classification

The proposed Neural Network (NN) architecture is inspired from the successful results that were presented in [4] and consists of three layers: the dimensionality reduction layer, the hidden layer and the classification layer. While the proposed NN framework uses only two Fully Connected layers, the statistical power that Fisher vectors encapsulate in their scheme passes in the NN as well and leads to a highly discriminative vector.

Two recognition tasks were taken into account, depending on the nature of the classification problem: A binary one for discriminating crisis events from others (i.e. water, fire and smoke recognition) and a multi-class problem to discriminate among various dynamic textures and moving objects. The first one is specifically tailored for serious real case scenarios that require a fast and accurate recognition of a crisis event, while the second one is performed in order to prove the discrimination power of our framework in more general scenarios.

For binary classification, we found that just 6 neural nodes in the first and second layer of the network are sufficient to create an accurate recognition scheme. A sigmoid kernel is then used as a final layer to classify crisis events from others. For multi-class recognition 128 neural nodes are used in the first layer, 64 features in the second layer

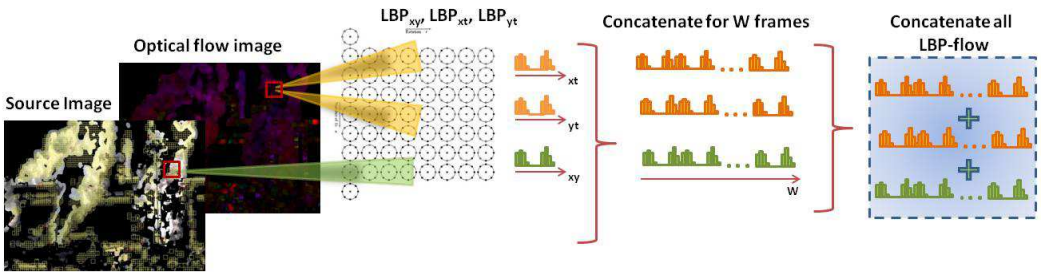

Figure 2. LBP-flow computation

and a softmax layer for classification purposes. This results in a much more complicated network, which is expected, as it requires more attributes than the binary one to discriminate among the classes. In both cases cross entropy loss is computed in order to measure the error in the output, while all the units in the hidden layers are rectified linear units (ReLu). Dropout has also been considered to reduce overfitting but did not improve our results and was eventually omitted from the scheme. To train each network the efficient Adam optimizer [9], a gradient-based optimization algorithm based on adaptive estimates of lower-order moments, is used with all parameters set at default.

4. Experiments

Experiments and comparisons with the SoA took place on three challenging benchmark datasets, namely Dyntex [15], videoWaterDatabase [11] and MovingVistas [16]. All datasets were split into $1/3$ for testing and $2/3$ for training, creating 3 different train/test splits to evaluate the performance of our algorithm. The dimension of the initial vector LPB-Flow was reduced via PCA from 5220 to 80 and 32 cluster centers where chosen to represent the visual vocabulary. $W_{AA} = 10$ and $W_{LBP} = 30$ were selected after empirical cross validation performed on the Dyntex dataset [15]. SVM classification is used when we need to measure the recognition accuracy of LBP-Flow, while SoftMax and Sigmoid is used when we apply the hybrid NN scheme.

Dyntex is one of the earliest and most renowned benchmark datasets for dynamic textures, containing a wide range of videos, such as fires, smoke, clouds, flags, waves, vegetation among others. In our first experiment, we split the dataset into water and non-water scenes to test the water classification accuracy of our descriptor. As we can see in Table 1, both schemes of LBP-Flow surpass VLBP and VLBP-TOP, but our accuracy is a little lower than that reported by Mettes in [11]. However, Mettes et al [11] do not specify which 80 videos were chosen from Dyntex to evaluate their descriptor, so a direct comparison with their results cannot be provided.

Method	Score
LBP-Flow	92.74%
LBP-Flow+NN	94.35%
LBP-Fourier [11]	**100%**
VLBP [19]	90.0%
LBP-TOP [18]	87.5%

Table 1. Comparison with SoA water/non-water binary classification on the Dyntex dataset.

In the second experiment that we performed in Dyntex we performed multi-class scene recognition so as to examine the applicability of our descriptor and our hybrid architecture in surveillance scenarios involving other kinds of dynamic textures. Table 2 shows that our descriptor achieves high accuracy in fire (82.1% for fire and 91.7% for smoke) and water-related (100% for CalmWater, 77.8% for Fountains, 88% for HomeWater and 100% for Sea) discrimination. In the lower-right edge of Table 2, it can be seen that water-related classes are confused for each other, rather than other categories of dynamic textures, thus enhancing our claim that LBP-Flow will robustly detect water-related dynamic textures. Comparisons with SoA multi-class recognition [6] in Table 3 show that our algorithm leads to more accurate results, especially for water and fire-related scenarios, than the more sophisticated but computationally expensive HOGP descriptor.

In recognition problems, feature extraction is usually the most time-consuming step. However, for the Dyntex database, the extraction of LBP-Flow features required only about 2.65 fps, due to the simple structure of LBP-flow. Its low computational cost makes it appropriate for security and surveillance applications, where videos are often recorded at $7-8$ fps, making our approach nearly real-time.

VideoWaterDatabase introduced in [11] contains 256 high definition videos, where the presence of water needs to be detected. This dataset contains water samples in fountains, waves, river and non-water video segments, as well as other dynamic texutres, like fires, clouds, flags. The patterns between the two classes are quite similar and very dif-

	Animals	Clouds	Fire	Flag	Lights	Rotation	Smoke	Underwater	Vegetation	CalmWater	Fountains	HomeWater	Sea
Animals	81,82%							18,18%					
Clouds		100,00%											
Fire			78,57%	3,57%	3,57%			7,14%				7,14%	
Flag			16,67%	75,00%			8,33%						
Lights			8,33%		91,67%								
Rotation						84,62%						15,38%	
Smoke			8,33%	FIRE			91,67%						
Underwater			2,63%					97,37%					
Vegetation								7,14%	92,86%	WATER			
CalmWater										100,00%			
Fountains											77,78%	22,22%	
HomeWater			12,00%							4,00%		84,00%	
Sea													100,00%
Average Accuracy	88,87%												

	Animals	Clouds	Fire	Flag	Lights	Rotation	Smoke	Underwater	Vegetation	CalmWater	Fountains	Home Water	Sea
Animals	45,45%							45,45%		9,09%			
Clouds		66,67%	33,33%										
Fire			82,14%		3,57%			3,57%				10,71%	
Flag			25,00%	66,67%								8,33%	
Lights			41,67%		50,00%							8,33%	
Rotation						76,92%					7,69%	15,38%	
Smoke							91,67%	8,33%					
Underwater	2,63%		2,63%	FIRE				94,74%		WATER			
Vegetation			7,14%					7,14%	78,57%			7,14%	
CalmWater			5,00%							85,00%			10,00%
Fountains			11,11%					11,11%		11,11%	55,56%	11,11%	
HomeWater			12,00%									88,00%	
Sea								4,17%					95,83%
Average Accuracy	75,17%												

Table 2. Multi-class classification accuracy of LBP-Flow over all Dyntex classes, when the hybrid(left) and the shallow(right) schemes are used.

	HOGP [6]	LBP-Flow	LBP-Flow+NN
Fire	-	**82.1%**	78.6%
Smoke	83.0%	91.7%	**91.7%**
Vegetation	81.0	78.6%	**92.9%**
Flags	56.0%	66.7%	**75.0%**
Fountain	**88.0%**	55.6%	77.8%
Calm Water	81.0%	85.0%	**100%**
Sea	81.0%	95.8%	**100%**
HomeWater	-	**88.0%**	84.0%
All	-	75.2%	**88.8%**

Table 3. Comparison with SoA on smoke and water related classes

Method	Score
LBP-Flow	98.3%
LBP-Flow+NN	**98.8%**
LBP-Fourier [11]	96.5%
VLBP [19]	93.5%
LBP-TOP [18]	93.3%

Table 4. Recognition accuracy on the VideoWaterDatabase benchmark dataset.

ficult to model. Comparisons with other dynamic texture modeling methods based on LBP are provided in Table 4, where it can be seen that our hybrid descriptor leads to the higher accuracy than the SoA, providing robust recognition results on challenging video datasets like VideoWaterDatabase. For the VideoWaterDatabase videos, the extraction of LBP-Flow features required about 2.05 fps, showing that our method is indeed computationally efficient, even for such challenging and complex datasets.

Moving vistas was introduced in [16] and is the last and the most challenging of all datasets, as it contains video samples of low quality using a moving camera, different viewpoints and significant illumination changes. Water-related dynamic textures appear in videos of: Boiling Water, Iceberg Collapse, Fountain, Waves, Waterfall, Whirlpool, while other types of dynamic textures include: Tornadoes, Chaotic Traffic, Landslide, Smooth Traffic, Volcanic Eruption, Avalanche and Forest Fire. In the first binary classification scenario, we achieved a highly accurate classification rate of 84.6% and 73.1% when our hybrid scheme and shallow scheme were applied respectively. This shows that LBP-Flow is more than appropriate for recognizing water-related video samples and secondly that the hybrid framework highly boosts the descriptor.

Method	Score
LBP-Flow	62.3%
LBP-Flow+NN	**63.8%**
SOE [5]	41.0%
[16]	52.0%

Table 6. Overall recognition accuracy on moving vista dataset

Multi-class recognition accuracy of LBP-FLow was estimated in our second experiment, to compare with the SoA on scene recognition in [5], [16]. The results, presented in Table 5 and Table 6, show that our hybrid scheme achieves significantly better recognition rates compared to the SoA, and at a much lower computational cost.

For the Moving vistas database, the extraction of LBP-Flow features required about 9.7 fps, due to the lowest resolution of the videos. Our method can therefore be applied to a variety of monitoring videos in real-world scenarios, ranging from low to high resolution, with near real-time performance.

5. Conclusions

In this paper we introduce a novel descriptor that extends the SoA LBP to the spatio-temporal domain, and particularly to the optical flow values, resulting in a highly discriminative and compact scheme after dimensionality reduction. Furthermore it introduces a hybrid classification scheme that leverages both shallow and depth features in order to recognise dynamic textures in an efficiently. Experiments on three recent and challenging benchmark datasets demonstrate that our approach achieves significantly higher recognition accuracy than the SoA in a computationally efficient manner. The results can be considered as an important indicator that these techniques could be used in real world applications such as the detection and monitoring of natural disasters (e.g. fires, floods) to improve the situational

	Avalanc e	Boiling Water	Chaotic Traffic	Forest Fire	Fountai n	Iceberg Collapse	Landslid e	Smoot h	Tornado	Volcano Erruptio	Waterfa ll	Waves	Whirlpoo l
Avalance	40,0%		10,0%			10,0%	10,0%	10,0%		20,0%			
Boiling Water		80,0%				10,0%	10,0%						
Chaotic Traffic			100,0%										
Forest Fire				80,0%					10,0%				10,0%
Fountain		10,0%		10,0%	60,0%			10,0%			10,0%		
Iceberg Collapse			10,0%			50,0%	10,0%	10,0%	20,0%				10,0%
Landslide					10,0%	10,0%	60,0%	10,0%					10,0%
Smooth Traffic			10,0%		10,0%			70,0%					10,0%
Tornado				20,0%			10,0%		50,0%	10,0%			10,0%
Volcano Erruption		10,0%	10,0%	10,0%					20,0%	50,0%			
Waterfall					10,0%		10,0%				70,0%	10,0%	
Waves	10,0%			10,0%		10,0%			10,0%			60,0%	
Whirlpool		10,0%				10,0%	10,0%		10,0%				60,0%
Average Accuracy	63,85%												

	Avalan ce	Boiling Water	Chaotic Traffic	Forest Fire	Fountai n	Iceberg Collapse	Landslid e	Smooth Traffic	Tornado	Volcan o	Waterf all	Waves	Whirlpoo l
Avalance	40,0%			10,0%		10,0%	10,0%	10,0%		20,0%			
Boiling Water		70,0%				10,0%				10,0%			10,0%
Chaotic Traffic			100,0%										
Forest Fire	10,0%			80,0%	10,0%								
Fountain				10,0%	40,0%	20,0%		10,0%			10,0%		10,0%
Iceberg Collapse						40,0%	10,0%	10,0%	20,0%	10,0%			10,0%
Landslide	10,0%				10,0%		70,0%						10,0%
Smooth Traffic		10,0%	10,0%		10,0%			60,0%					10,0%
Tornado				10,0%		20,0%			60,0%	10,0%			
Volcano Erruption	10,0%					10,0%			20,0%	50,0%	10,0%		
Waterfall					10,0%		10,0%				80,0%		
Waves	10,0%					10,0%	10,0%		20,0%			50,0%	
Whirlpool		10,0%				10,0%			10,0%				70,0%
Average Accuracy	62,31%												

Table 5. Multi-class classification accuracy of LBP-Flow over all Moving Vista classes

awareness in crisis management. Future work will include the challenging problem of spatio-temporal detection of dynamic textures in temporally unsegmented videos of a long duration.

6. Acknowledgement

This work was supported by beAWARE[1] project partially funded by the European Commission under grant agreement No 700475.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 28(12):2037–2041, 2006. 3

[2] K. Avgerinakis, A. Briassouli, and Y. Kompatsiaris. Activity detection using sequential statistical boundary detection (ssbd). *Computer Vision and Image Understanding*, 2015. 3

[3] A. B. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE transactions on pattern analysis and machine intelligence*, 30(5):909–926, 2008. 2

[4] C. R. de Souza, A. Gaidon, E. Vig, and A. M. López. Sympathy for the details: Dense trajectories and hybrid classification architectures for action recognition. In *European Conference on Computer Vision*, pages 697–716. Springer International Publishing, 2016. 1, 3, 4

[5] K. G. Derpanis, M. Lecce, K. Daniilidis, and R. P. Wildes. Dynamic scene understanding: The role of orientation features in space and time in scene classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1306–1313. IEEE, 2012. 2, 6

[6] K. Dimitropoulos, P. Barmpoutis, A. Kitsikidis, and N. Grammalidis. Classification of multidimensional time-evolving data using histograms of grassmannian points. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016. 2, 5, 6

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

[8] C. Feichtenhofer, A. Pinz, and R. Wildes. Bags of spacetime energies for dynamic scene recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2681–2688, 2014. 2

[9] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5

[10] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2929–2936. IEEE, 2009. 2

[11] P. Mettes, R. T. Tan, and R. Veltkamp. On the segmentation and classification of water in videos. In *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, volume 1, pages 283–292. IEEE, 2014. 2, 3, 5, 6

[12] A. Mumtaz, E. Coviello, G. R. Lanckriet, and A. B. Chan. Clustering dynamic textures with the hierarchical em algorithm for modeling video. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1606–1621, 2013. 2

[13] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001. 2

[14] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. *Computer Vision–ECCV 2010*, pages 143–156, 2010. 3, 4

[15] R. Péteri, S. Fazekas, and M. J. Huiskes. DynTex : a Comprehensive Database of Dynamic Textures. *Pattern Recognition Letters*, doi: 10.1016/j.patrec.2010.05.009. http://projects.cwi.nl/dyntex/. 5

[16] N. Shroff, P. Turaga, and R. Chellappa. Moving vistas: Exploiting motion for describing scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1911–1918. IEEE, 2010. 2, 5, 6

[17] X. Wang, T. X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 32–39. IEEE, 2009. 3

[18] G. Zhao, T. Ahonen, J. Matas, and M. Pietikainen. Rotation-invariant image and video description with local binary pattern features. *IEEE Transactions on Image Processing*, 21(4):1465–1477, 2012. 2, 4, 5, 6

[19] G. Zhao and M. Pietikainen. Local binary pattern descriptors for dynamic texture recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 2, pages 211–214. IEEE, 2006. 2, 5, 6

[1] http://beaware-project.eu/