Directional Space-Time Oriented Gradients for 3D Visual Pattern Analysis

Ehsan Norouznezhad[1,2], Mehrtash T. Harandi[1,2], Abbas Bigdeli[1,2], Mahsa Baktash[1,2], Adam Postula[1,2], and Brian C. Lovell[1,2]

[1] NICTA, P.O. Box 6020, St. Lucia, QLD 4067, Australia
[2] The University of Queensland, School of ITEE, QLD 4072, Australia

Abstract. Various visual tasks such as the recognition of human actions, gestures, facial expressions, and classification of dynamic textures require modeling and the representation of spatio-temporal information. In this paper, we propose representing space-time patterns using directional spatio-temporal oriented gradients. In the proposed approach, a 3D video patch is represented by a histogram of oriented gradients over nine symmetric spatio-temporal planes. Video comparison is achieved through a positive definite similarity kernel that is learnt by multiple kernel learning. A rich spatio-temporal descriptor with a simple trade-off between discriminatory power and invariance properties is thereby obtained. To evaluate the proposed approach, we consider three challenging visual recognition tasks, namely the classification of dynamic textures, human gestures and human actions. Our evaluations indicate that the proposed approach attains significant classification improvements in recognition accuracy in comparison to state-of-the-art methods such as LBP-TOP, 3D-SIFT, HOG3D, tensor canonical correlation analysis, and dynamical fractal analysis.

1 Introduction

The goal of visual pattern recognition is to detect the presence of a particular object or pattern in a given image or video. This usually involves representing patterns in a suitable feature space to achieve robustness against a broad range of environmental changes like photometric variations, occlusions, background clutter, geometric transformations, and variations in view angle.

The study of space-time patterns such as human actions, gestures, facial expressions, and dynamic textures has attracted growing attention, mainly due to the wide range of applications in real world [1]. One of the major theme of research in space-time pattern classification is shaped around devising robust spatio-temporal local descriptors [1,2]. Broadly speaking, spatio-temporal local descriptors can be categorized into three main classes.

The first and the largest category includes the direct extension of 2D local descriptors to 3D. The underlying idea is to replace the rectangular regions in 2D descriptors with 3D patches and recast the functions/procedures from the spatial domain (ie. (x, y)) into the spatio-temporal space (ie. (x, y, t)). Examples include 3D-SIFT by Scovanner et al. [3], HOG3D by Klaser et al. [4], Volume Local Binary Patterns (VLBP) by Zhao et al. [5] and Extended-SURF (ESURF) by Willems et al. [6].

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part III, LNCS 7574, pp. 736–749, 2012.

In the second category, the static (ie. spatial) and dynamic (ie. temporal) properties of a given 3D patch are independently captured and then merged to form the descriptor. The HOG/HOF descriptor proposed by Laptev et al. [7] is an example of this school of thought. Other examples of this category include combining appearance and motion descriptors for recognizing human actions as proposed by Schindler et al. [8] and Huang et al. [9].

In the third category, no direct extension from 2D to 3D is considered. Instead, a set of 2D local descriptors are extracted from three orthogonal spatio-temporal planes within a 3D patch. Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) proposed by Zhao et al. [5] was one of the very first attempts in this direction and has been successfully applied in classification of various space-time patterns such as dynamic textures, facial expressions [5] and human actions [10]. In LBP-TOP, a video patch is encoded by computing LBP features over the three orthogonal XY, XT, and YT planes. Other examples of encoding video patches by three orthogonal planes include extensions of the Weber's local descriptor (WLD) [11] and Local Phase Quantization (LQP) [12] to WLD-TOP [10] and LQP-TOP [13] respectively.

Despite considerable success, two major shortcomings are associated with the notion of three orthogonal planes (TOP). First, TOP is not able to optimally capture the dynamical properties of a given 3D patch. This is mainly due to the fact that dynamical information is only encoded by two planes (XT and YT). Second, to create the video descriptor, the descriptors associated with the XY, XT and YT planes are simply concatenated together. This is restrictive and may lead to inappropriate modeling in more complex recognition scenarios.

With the aim of addressing the aforementioned shortcomings, in this paper we extend the TOP approach to Nine Symmetry Planes (NSP). Unlike the TOP approach, NSP benefits from six more diagonal planes to model the dynamical information of a video patch. In our proposal, a video patch is described by a Histogram of Oriented Gradients (HOG) using gradient information in all nine symmetric planes. Moreover, to overcome the second shortcoming, we propose to learn the optimal fusion of sub-descriptors by a positive definite similarity kernel. This can be done through the notion of Multiple Kernel Learning (MKL) and is subtly different from previous studies on feature fusion [1]. As such, our proposed method, the *HOG-NSP*, is free from concatenation at any level.

1.1 Contributions

The three main novelties in this work are **(i)** we propose to encode a video patch through nine spatio-temporal symmetry planes, and **(ii)** we apply a positive definite similarity kernel learnt by multiple kernel learning to compare two videos, and **(iii)** the proposed descriptor has been analyzed and contrasted against state-of-the-art methods in three visual recognition tasks, namely the classification of dynamic textures, hand gestures and human actions.

The remainder of the paper is organized as follows: We elaborate on the NSP approach in Section 2. In Section 3 we compare the performance of the proposed method

[1] While in previous studies, MKL is employed to fuse features of different kind such as color, texture, and motion, to author's knowledge, this is the first paper that adapts MKL to fuse sub-descriptors of the same kind.

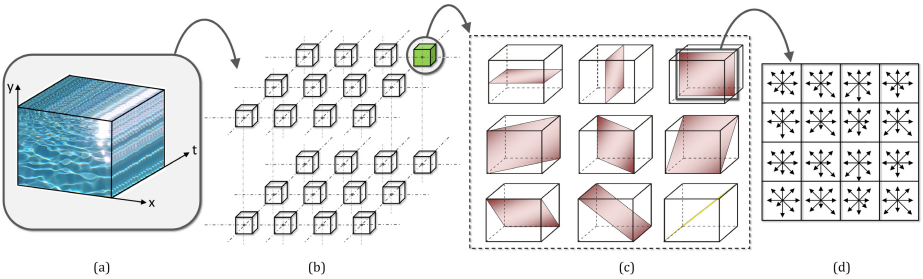

Fig. 1. The HOG-NSP feature extraction procedure: **(a)** input space-time pattern as a 3D volume; **(b)** extract cuboid patches on dense spatio-temporal grid; **(c)** nine symmetry planes within the patch are considered for feature extraction with each plane corresponding to a specific space-time direction; **(d)** extract oriented gradients on each plane and represent as locally normalized histogram

with previous approaches on the aforementioned visual classification tasks. Finally, the main findings and possible future directions are summarised in Section 4.

2 Proposed Approach

The proposed methods consists of three main components:

1. **Space-Time Plane Descriptor.** A video is decomposed into a set of 3D cuboid patches. Each cuboid patch is then represented by histogram of oriented gradients over nine spatio-temporal symmetry planes. The final descriptor of a 3D cuboid patch, is obtained by the Bag of features (Bof) framework. As such, a dictionary is trained for each spatio-temporal plane and used to encode 3D cuboid patches.
2. **Video sub-descriptors.** A video is described by a set of sub-descriptor using a spatio-temporal pyramid. More specifically, individual representations of a given video will be created in each channel of spatio-temporal pyramid. The total number of sub-descriptors is equal to the number of plane-descriptors multiplied by the number of spatio-temporal grids.
3. **Similarity-Based Classification.** Video comparison is achieved through a positive definite similarity kernel that is learnt by multiple kernel learning. The similarity kernel takes into account all video sub-descriptors and can be seen as a feature fusion.

Each of the components is explained in detail in the following sections.

2.1 Space-Time Plane Descriptor

To describe a video sequence, HOG-NSP decomposes the video into a set of cuboid patches of size $w_s \times h_s \times l_s$ with possible overlapping using a fixed and dense grid structure (see Fig. 1). We note that HOG-NSP can be used in conjunction with interest

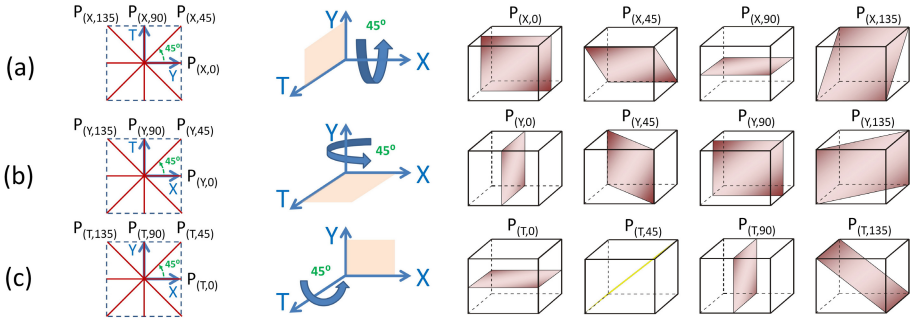

Fig. 2. The nine symmetry planes: **(a)** the first row illustrates the XY and its rotations around the X axis; **(b)** the middle row demonstrates the YT plane and its rotations around the Y axis; **(c)** the bottom row shows the XT plane and its rotations around the T axis. It should be noted that we are left with nine planes since some planes are repeated twice ($P_{(X,0)} = P_{(Y,0)}$, $P_{(T,90)} = P_{(Y,90)}$ and $P_{(T,0)} = P_{(X,90)}$).

point detectors [14] to represent a video sparsely. However, in this study we confine ourselves to dense grid-based representation. This is mainly driven by the recent success of dense representation in various recognition scenarios [15,1].

Within each cuboid patch, nine space-time directions are considered for feature extraction. Each space-time direction corresponds to one symmetry planes inside the cuboid patch as shown in Fig. 1. The planes can be obtained systematically by rotating XY, XT and YT planes. For example, by rotating the XY plane around the X axis by $45°$, $90°$ and $135°$, three orthogonal planes to the YT plane can be attained. The rotated planes are shown at the top row of Fig. 2 and labeled $P_{(X,45)}$, $P_{(X,90)}$, $P_{(X,135)}$ respectively. Similarly, three orthogonal planes to XT plane and three orthogonal planes to XY plane are constructed by rotating XY and XT planes as shown in the middle and bottom rows of Fig. 2. Since some planes are repeated twice ($P_{(X,0)} = P_{(Y,0)}$, $P_{(T,90)} = P_{(Y,90)}$ and $P_{(T,0)} = P_{(X,90)}$), we are left with nine distinct planes.

Having nine symmetry planes at our disposal, a cuboid patch is then represented by normalized gradients histograms over the spatio-temporal planes. This is accomplished by dividing each plane into smaller blocks of size $M \times N$ and computing histogram of oriented gradients within each block. For normalization, we pursue a similar normalization strategy as Lowe [16]. The descriptor in each plane is normalized using the l_2-norm with predefined cut values of c. In our experiments, the cut value is set to $c = 0.25$. Furthermore sub-block and histogram parameters are set to $M = 4$, $N = 4$ and $B = 8$ which results in a 128 dimensional plane descriptor.

HOG-NSP should not be confused with other 3D video descriptors such as HOG3D or 3D-SIFT. HOG3D or 3D-SIFT represent 3D video patches by measuring and quantizing the 3D gradient orientations using a regular n-sided polyhedron. In contrast, HOG-NSP uses directional spatio-temporal oriented gradients to model static and dynamic properties of 3D patches. In other words, local oriented gradients in nine discrete space-time directions (planes) within a 3D patch are used for representation.

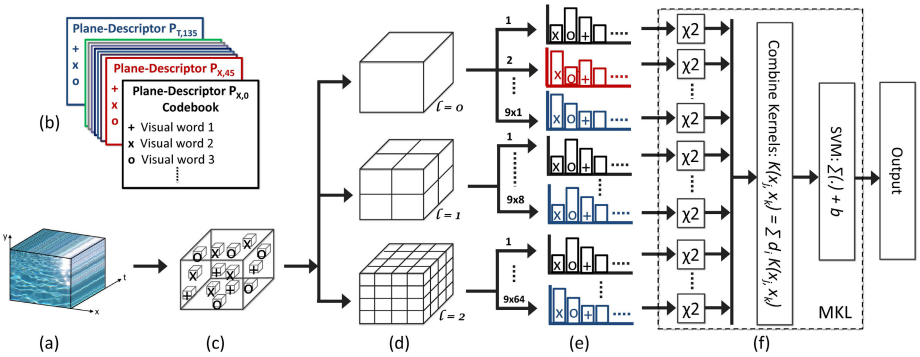

Fig. 3. The classification procedure: **(a)** input space-time pattern; **(b)** separate visual dictionaries (codebook) are learned by quantizing the feature space for each plane-descriptor using the training dataset; each codebook has a number of visual words; only three visual words are shown in this figure for illustration purposes **(c)** HOG-NSP features are extracted for a given video; **(d)** Spatio-temporal pyramids are created for representing videos; for illustrative purposes, only three levels ($L = 0, 1, 2$) are shown; **(e)** individual representations of a given video will be created in each channel; the number of channels is equal to the number of plane-descriptors (i.e. 9) multiplied by the number of spatio-temporal grids (i.e. $1, 8$ and 64 for $L = 0, 1, 2$ respectively); **(f)** the representation for each channel is assigned a kernel and multiple kernel learning (MKL) is used to find the optimal weights for the kernels; SVM is used for classification.

2.2 Bag-of-Features

Given a set of spatio-temporal local descriptors, the next step is to represent the video based on the extracted descriptors. We use bag-of-features (BoF) framework for this purpose. In BoF framework, the first step is to construct a visual dictionary called codebook. This is achieved by clustering the descriptors that are extracted from the videos in the training set using k-means. This results in nine separate visual vocabularies for the nine plane-descriptors. In our experiments, the number of visual words in the each dictionary is set to 800 by default which have shown empirically to give good results for the datasets under test. Given a video, the extracted plane-descriptors are assigned to their closest visual word, using the Euclidean distance. The occurrences of each of the visual words of each plane-descriptor can be used to represent a given video.

2.3 Encoding Global Structural Information

The bag-of-features (BoF) approach represents a video as an orderless collection of local patches. It does not preserve any information regarding the global spatio-temporal distributions of local patches. To overcome this problem, Lazebink et al. [17] proposed Spatial Pyramid Matching (SPM) method. SPM encodes the coarse layout of local patches which has proven to be very effective for object and scene classification [17]. In their proposed approach, the given image is repeatedly subdivided and distributions of local features is computed and compared at increasingly finer resolutions.

The SPM approach was later extended to spatio-temporal domain and was utilized for representing space-time patterns in videos [7,18]. In our proposed approach, Spatio-Temporal Pyramid Matching (STPM) is employed to encode global structural information into the representation. A sequence of increasingly finer binary partitions are constructed at resolution $0, , L$, such that the partition at each level l has 2^l cells along each dimension. Fig. 3 illustrates the STPM approach for $L = 2$. Combining the spatio-temporal grids at $L = 0, 1, 2$ will result in in 73 spatio-temporal grids.

2.4 Similarity-Based Classification

Up to this point, we have elaborated on how a video sequence v_i can be represented by a set of local descriptors $\{\boldsymbol{x}_1^i, \boldsymbol{x}_2^i, \cdots, \boldsymbol{x}_n^i\}, \boldsymbol{x}_j^i \in \mathbb{R}^{n_{dic}}$. One can create a unified video descriptor by concatenating the local descriptors into a large feature vector and use it for classification. Nevertheless simple concatenation is not suitable for HOG-NSP due to the curse of dimensionality. More specifically, in HOG-NSP the size of each local descriptor is equal to the length of the associated dictionary. Each video descriptor constitutes of $9 \times n_{STPM}$ local descriptors. As a result by concatenating local descriptors, the size of video descriptor would become $9 \times n_{STPM} \times n_{dic}$ which is restrictive in many practical applications. To remedy this problem, we propose to compare two video sequences based on the notion of similarity classification and Multiple Kernel Learning (MKL) [19,20]. More specifically, the similarity of two video v_s and v_t, each represented by a set of local descriptors $\{\boldsymbol{x}_1^s, \boldsymbol{x}_2^s, \cdots, \boldsymbol{x}_n^s\}$ and $\{\boldsymbol{x}_1^t, \boldsymbol{x}_2^t, \cdots, \boldsymbol{x}_n^t\}$ is defined as:

$$K\left(v_s, v_t\right) = \sum_{i=1}^{n} d_i \exp\left(-\sigma \chi^2(\boldsymbol{x}_i^s, \boldsymbol{x}_i^t)\right) \tag{1}$$

where $d_i \geq 0, \sum_{i=1}^{n} d_i = 1$ are kernel combination weights and $k(\cdot, \cdot)$ is the chi-squared distance defined as:

$$\chi(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{n_{dic}} \frac{\left(x(j) - y(j)\right)^2}{x(j) + y(j)} \tag{2}$$

The kernel combination weights, (d_i), can be learnt through a convex optimization problem as discussed for example in [20]. Having a positive definite similarity kernel at our disposal, Support Vector Machines can be utilized to classify video sequences.

2.5 Properties of HOG-NSP Descriptor

The HOG-NSP enjoys a number of properties that are worth emphasizing. Firstly, while local orientated gradients in space characterizes the appearance, it represent motion and apearance in space-time domain. Therefore HOG-NSP encodes both static and dynamic properties of video pattern by spatial and spatio-temporal planes respectively. Secondly, thanks to the inherent properties of the oriented gradient descriptor and the normalization scheme, the proposed descriptor is robust against photometric variations. Lastly, the approach is robust against geometric variations (eg. rotation, shift or scale). This

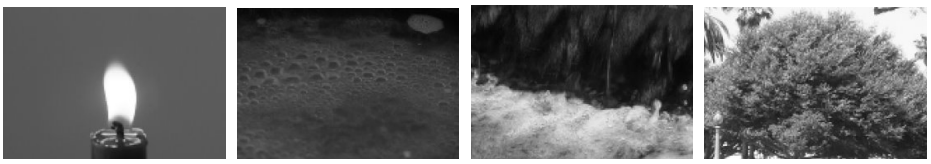

Fig. 4. Representative examples from the UCLA dynamic texture database [24] used for evaluation

is achieved through locally measuring of the static and dynamic properties of a given pattern. Furthermore our experiments demonstrate that utilizing a multi-plane approach for feature extraction and reducing the angle between spatio-temporal planes (the angle between spatio-temporal planes is reduced to $45°$) has further increased the robustness of the proposed descriptor.

3 Experiments

To evaluate and contrast the performance of the proposed HOG-NSP approach, three sets of experiments are considered in this section. First, we appraise HOG-NSP against state-of-the-art methods such as tensor canonical correlation analysis [21], dynamical fractal analysis [22] and Kinematic features with multiple instance learning [23] in classifying dynamic textures, hand gestures and human actions. Second, we compare the performance of HOG-NSP descriptor against spatio-temporal local descriptors such as 3D-SIFT and LBP-TOP. For the latter case we only consider the recognition of dynamic texture. Finally, the computational complexity of HOG-NSP descriptor is compared against state-of-the-art spatio-temporal local descriptors.

3.1 Dynamic Texture Recognition

Dynamic textures (DTs) are video sequences that exhibit some form of stationary patterns in time [24]. Examples of dynamic textures such as fire, smoke, river, clouds and windblown vegetation can be found commonly in real world. To assess HOG-NSP, we have selected two publicly available datasets, namely UCLA [24] and DynTex++ [25]. The UCLA dynamic texture dataset has been widely used for benchmarking DT classification methods and it contains 50 classes of different dynamic textures. Each texture has four video sequences (ie. 200 DT sequences in total) recorded from various viewpoints. The dataset is originally recorded in color with a resolution of 220×320 pixels. In our tests, the videos were cropped to 48×48 pixels and converted to grayscale.

For a thorough evaluation, we have considered four different test protocols [22]. The first test protocol consists of 50 classes of dynamic textures, each having 4 sequences. In the second protocol, the video sequences are grouped into 9 classes by combining classes which refer to similar dynamic texture recorded from different viewpoints. The DT classes in the second protocol are boiling water (8), fire (8), flowers (12), sea (12), smoke (4), fountains (20), water (12), waterfall (16) and plants (108), where the numbers denotes the number of sequences in each class. In the third test

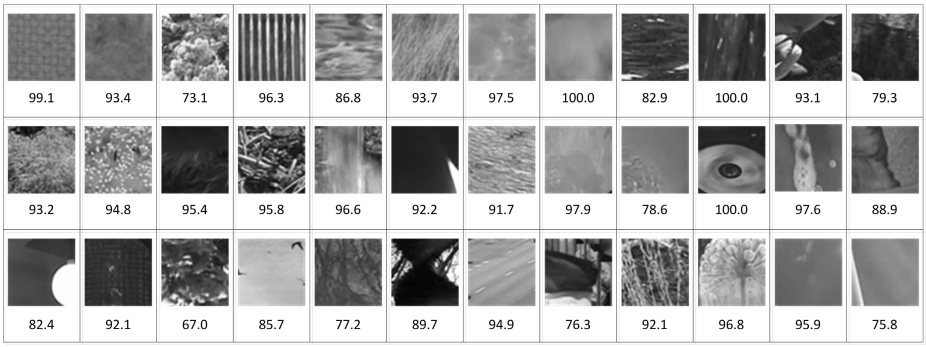

Fig. 5. Classification rate (%) on each class of the DynTex++ dataset

protocol [26], the number of classes is reduced to 8 by removing the class of plants. Finally the fourth test protocol, shift-invariant-recognition (SIR) was proposed by [27] to evaluate shift-invariance properties of descriptors. Representative examples of UCLA dataset are shown in Fig. 4

The second dataset is the Dyntex++ dataset. This dataset is compiled from a Dyntex dataset [28] and is proposed in [25]. It contains 3600 dynamic textures, grouped into 36 classes. Each sequence has a size of $50 \times 50 \times 50$ pixels. We follow the test protocol proposed in [25]. In all our experiments the data was randomly split into two equal size training and test sets. The random split was repeated 20 times and the average classification accuracy is reported here.

For representing dynamic texture, spatio-temporal pyramid with $l = 0$ was used which is the standard bag-of-features (BoF) representation. This is due to the fact that in our experiments we did not observe significant improvement by using higher levels of pyramid. Furthermore the size of codebook and cuboid patch was empirically set to 800 and $16 \times 16 \times 16$ pixels with 50% overlap respectively.

The proposed approach is compared against the following methods: Bag-of-dynamical systems (BDS) [26], Distance Learning using DL-PEGASOS [25], Space-time oriented structures [27], and dynamic fractal spectrum (DFS) descriptors [22]. BDS models dynamic textures using linear dynamical systems (LDS) in the framework of Bag-of-features. Space-time oriented structures and DFS utilize 3D oriented filters for representing dynamic textures. DFS exploit dynamic fractal analysis for modelling dynamic textures. DL-PEGASOS uses maximum margin distance learning (MMDL) approach based on Pegasos algorithm for dynamic texture recognition.

The classification accuracies are presented in Table 1 and 2 for UCLA and Dyntex++ dataset respectively. In UCLA dataset, the proposed approach achieves the highest performance with 98.1% and 78.2% in 9-class and SIR test protocols respectively. In 8-class test protocol, the proposed approach achieves 98.7% which is very close to state-of-the-art (99%). Moreover in 50-class test protocol the proposed approach achieves the second highest performance with 97.2% as compared to DFS method [22] that achieves perfect recognition accuracy. In Dyntex++ dataset, the proposed approach achieves the highest classification rate with 90.1% in comparison to DL-Pegasos [25] and DFS [22].

Table 1. Classification rate (%) for dynamic texture recognition on the UCLA dataset using BDS [26], DL-PEGASOS [25], Space-time oriented structures (STOS) [27], and DFS descriptors [22] and the proposed approach (HOG-NSP descriptor in STPM framework using MKL for sub-descriptors fusion)

Method	50-class	9-class	8-class	SIR
BDS [26]	N/A	N/A	80	N/A
DL-PEGASOS [25]	N/A	95.6	**99**	N/A
STOS [27]	81	N/A	N/A	42.3
DFS [22]	89.5, **100**	97.5	**99**	73.8
HOG-NSP	97.2	**98.1**	98.7	**78.2**

Table 2. Classification rate (%) for dynamic texture recognition on the Dyntex++ dynamic texture database using DL-PEGASOS [25] , DFS [22], and the proposed approach

Method	DL-PEGASOS [25]	DFS [22]	HOG-NSP
Accuracy	63.7	89.9	90.1

3.2 Gesture Recognition

For gesture recognition, we used the Cambridge hand gesture dataset [21]. It consists of 900 image sequences of 9 gesture classes. The gesture classes are defined by three primitive hand shapes and three primitive motions. Each class has 100 image sequences which are performed by 2 subjects, captured under 5 illuminations and 10 arbitrary motions. Each sequence was recorded at $30 fps$ with a resolution of 320×240, in front of a fixed camera having roughly isolated gestures in space and time. Representative examples of this dataset are shown in Fig 6.

Following the test protocol in [21], sequences are divided into five sets where the fifth set with normal illumination is used for training and the remaining sets are used for testing. As per [21], we report the classification rate for all the four illumination sets. The proposed method is compared against the original Canonical Correlation Analysis (CCA) [29], Tensor Canonical Correlation Analysis (TCCA) [21], and Product Manifolds (PM) [30]. Canonical correlation analysis (CCA) is a standard method for measuring the similarity between subspaces. TCCA, as the name implies, is the extension of canonical correlation analysis to multiway data arrays or tensors. In the PM method, a tensor is characterised as a point on a product manifold and classification is performed on this space.

For the hand gesture test, HOG-NSP with three levels of spatio-temporal pyramid attains the maximum performance. As such each video sequence is described by $(1 + 8 + 64) \times 9$ spatio-temporal sub-descriptor. The size of codebook and cuboid patch is

Fig. 6. Examples of actions in the Cambridge hand-gesture video dataset [21]

Table 3. Average correct classification rate (%) for gesture recognition on the Cambridge hand-gesture database using CCA [29], TCCA [21], Product Manifold [30] and the proposed approach

Method	Set1	Set2	Set3	Set4	Overall
CCA [29]	63	61	65	69	65 ± 3.2
TCCA [21]	81	81	78	86	82 ± 3.5
PM [30]	89	86	89	87	88 ± 2.1
HOG-NSP	**91**	**87**	**90**	**89**	$\mathbf{90 \pm 1.3}$

the same as previous experiment. The results, presented in Table 3, demonstrate that the proposed approach achieves the highest performance with overall classification rate of 90%. We note that the performance gap between CCA, TCCA and HOG-NSP is quite significant. The proposed approach also achieves the lowest variations with 1.3% standard deviations. This is an indication of robustness in presence of photometric variations.

3.3 Action Recognition

In order to evaluate the performance of proposed approach for action recognition, we used two publicly available datasets, namely the KTH [31] and Weizmann [32] datasets. Both datasets consist of videos of a single person, performing various actions in front of a static camera with uncluttered homogenous backgrounds. In the case of KTH dataset, there are six different human actions classes (see figure 7 for examples), performed by 25 subjects in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes, and indoors. In total, there are 2391 sequences. The Weizmann actions dataset consists of ten different types of human action classes performed by 9 subjects. There are 93 sequences in total. For fair comparison, the test protocols presented in [31] and [32] were used for KTH and Weizmann datasets respectively. For both datasets, the classification performance is presented as average accuracy over all classes. The proposed method is compared against six other state-of-the-art approaches, the HOG/HOF [7], 3D-SIFT [3], HOG3D [4], LBP-TOP [10], Kinematic features with multiple instance learning (Abbreviated as KF-MIL) [23], and unsupervised feature learning using independent subspace analysis (Abbreviated as UFL-ISA) [33]. For the human actions experiment, the levels of spatio-temporal pyramids, size of the codebook and cuboid patch was empirically set to three, 1000 and $18 \times 18 \times 18$ pixels with 50% overlap respectively. The results presented in table 4 demonstrate that the proposed approach achieves the highest performance in both datasets.

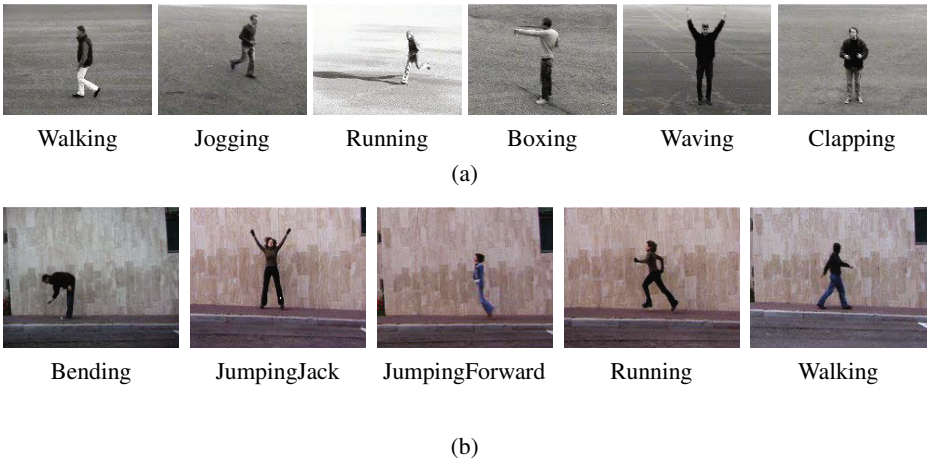

Fig. 7. Sample frames from (a) KTH (b) Weizmann dataset

Table 4. Comparison of mean classification accuracy on the KTH and Weizmann datasets

Method	3D-SIFT [3]	HOG3D [4]	HOG/HOF [7]	LBP-TOP [10]	KF-MIL [23]	UFL-ISA [33]	HOG-NSP
KTH	N/A	91.4 %	91.8 %	86.25 %	87.7 %	93.9 %	**96.4**%
Weizmann	82.6 %	84.3 %	N/A	N/A	95.75 %	N/A	**95.9** %

3.4 Comparison with Local Descriptors

In order to contrast the performance of the proposed spatio-temporal local descriptor against previous video descriptors, we performed another set of experiments. For a fair comparison, the MKL framework is removed from HOG-NSP and sub-descriptors are uniformly concatenated. The proposed descriptor is evaluated against three state-of-the-art spatio-temporal local descriptors, namely 3D-SIFT [3], VLBP and LBP-TOP [5]. For all descriptors, cuboid patches of $16 \times 16 \times 16$ pixels with 50% overlap were used . Inline with evaluation framework presented in [1], the size of codebook was selected to be 4000. For classification, we used Nearest-Neighbour (NN) classifier for all descriptors. The performance of each descriptor was evaluated on both UCLA and Dyntex++ datasets. For UCLA dataset, three different test protocols, the 50-class, 8-class and shift-invariant-recognition (SIR) were used. The results presented in Table 5 demonstrate that the proposed HOG-NSP descriptor significantly outperforms the other three descriptors by a notable margin. Moreover, the shift-invariant test indicates that the HOG-NSP enjoys a better level of robustness in presence of geometric variations. Comparing HOG-NSP, HOG-TOP and LBP-TOP that utilize spatio-temporal planes for feature extraction against direct extension methods like 3D-SIFT and VLBP, illustrates that the former approaches achieve higher accuracy.

Table 5. Classification rate (%) for dynamic texture recognition on UCLA and Dyntex++ dataset

Method	UCLA			Dyntex++
	50-class	8-class	SIR	
3D-SIFT [3]	77.3	86.2	44.7	63.7
VLBP [5]	79.8	91.4	40.6	61.1
LBP-TOP [5]	86.1	96.8	60.3	71.2
HOG-TOP	83.7	95.4	62.7	72.8
HOG-NSP	**88.5**	**97.5**	**66.3**	**78.7**

3.5 Computational Complexity

As final evaluation, we compare the computational complexity of HOG-NSP descriptor against other state-of-the-art 3D video descriptors. To this end, we measured the average time required for computating HOG-NSP, LBP-TOP [5], 3D-SIFT [3] and HOG3D [4] on a set of videos from KTH actions dataset [31]. Experiments were performed on an Intel 3.00GHz processor using Matlab. Moreover, all videos were downsampled to 160×120 and a grid of non-overlapping $16 \times 16 \times 16$ cuboid patches were used to compute each descriptor.[2]. Table 6 presents the average runtime for all studied descriptors. The proposed HOG-NSP descriptor is more than 8 times faster than HOG3D and 3D-SIFT. Moreover, the complexity of HOG-NSP is roughly similar as LBP-TOP.

Table 6. Average run-time (sec) for computing HOG-NSP, LBP-TOP [5], 3D-SIFT [3] and HOG3D [4] descriptors on KTH dataset

Descriptor	HOG-NSP	LBP-TOP [5]	3D-SIFT [3]	HOG3D [4]
Average run-time (sec)	8.6	7.4	74.7	73.3

4 Main Findings and Future Directions

In this paper, we proposed a novel approach for space-time pattern recognition. The main contributions of this work are the proposed 3D video descriptors and the application of Multiple Kernel Learning (MKL) for optimal fusion of sub-descriptors. In our proposal, a local 3D video patch is described by histogram of oriented gradients on nine discrete space-time planes which are the symmetry planes of the 3D patch. This results in a rich descriptor that encodes both static and dynamic properties of video pattern. Furthermore, to compare two video sequences, we proposed to learn a positive-definite similarity kernel by combining local video descriptors over a spatio-temporal pyramid structure.

[2] The source codes of LBP-TOP and 3D-SIFT descriptors were downloaded from the authors' website while HOG3D was implemented in Matlab by ourselves.

We extensively evaluated the performance of the proposed approach on three challenging visual recognition tasks, namely classification of dynamic textures, hand gestures and human actions. The experiments indicate that the proposed approach achieves superior or comparable performance to state-of-the-arts methods. Furthermore the proposed HOG-NSP descriptor outperforms state-of-the-art 3D video descriptors such as such as LBP-TOP, 3D-SIFT and HOG3D, with much lower computational load compared to HOG3D or 3D-SIFT. We also acknowledge that the proposed approach has achieved the highest reported performance on DynTex++, a challenging dataset for evaluating dynamic texture methods. In this study, the symmetry spatio-temporal planes are encoded using oriented gradient features. However, our proposed approach is quite general and is not restricted to gradient features. As such, future avenues of research include exploring integration of other 2D local descriptors inside the HOG-NSP machinery. We also plan to appraise HOG-NSP on other space-time classification tasks such as facial expression recognition.

Acknowledgements. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy, as well as by the Australian Research Council through the ICT Centre of Excellence program.

References

1. Wang, H., Ullah, M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition (2009)
2. de Campos, T., Barnard, M., Mikolajczyk, K., Kittler, J., Yan, F., Christmas, W., Windridge, D.: An evaluation of bags-of-words and spatio-temporal shapes for action recognition. In: WACV, pp. 344–351 (2011)
3. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: International Conference on Multimedia, pp. 357–360 (2007)
4. Kläser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC, pp. 995–1004 (2008)
5. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. PAMI 29(6), 915–928 (2007)
6. Willems, G., Tuytelaars, T., Van Gool, L.: An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)
7. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR, pp. 1–8 (2008)
8. Schindler, K., Van Gool, L.: Action snippets: How many frames does human action recognition require? In: CVPR, pp. 1–8 (2008)
9. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: ICCV, pp. 1–8 (2007)
10. Mattivi, R., Shao, L.: Human Action Recognition Using LBP-TOP as Sparse Spatio-Temporal Feature Descriptor. In: Jiang, X., Petkov, N. (eds.) CAIP 2009. LNCS, vol. 5702, pp. 740–747. Springer, Heidelberg (2009)
11. Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., Gao, W.: Wld: A robust local image descriptor. PAMI 32(9), 1705–1720 (2010)

12. Ojansivu, V., Heikkilä, J.: Blur Insensitive Texture Classification Using Local Phase Quantization. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) ICISP 2008. LNCS, vol. 5099, pp. 236–243. Springer, Heidelberg (2008)
13. Päivärinta, J., Rahtu, E., Heikkilä, J.: Volume Local Phase Quantization for Blur-Insensitive Dynamic Texture Classification. In: Heyden, A., Kahl, F. (eds.) SCIA 2011. LNCS, vol. 6688, pp. 360–369. Springer, Heidelberg (2011)
14. Laptev, I.: On space-time interest points. IJCV 64(2), 107–123 (2005)
15. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR, vol. 2, pp. 524–531 (2005)
16. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
17. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR, vol. 2, pp. 2169–2178 (2006)
18. Choi, J., Jeon, W., Lee, S.: Spatio-temporal pyramid matching for sports videos. In: ACM Int. Conf. on Multimedia Information Retrieval, pp. 291–297 (2008)
19. Chen, Y., Garcia, E.K., Gupta, M.R., Rahimi, A., Cazzanti, L.: Similarity-based classification: Concepts and algorithms. JMLR 10, 747–776 (2009)
20. Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y.: SimpleMKL. JMLR 9, 2491–2521 (2008)
21. Kim, T., Cipolla, R.: Canonical correlation analysis of video volume tensors for action categorization and detection. PAMI 31(8), 1415–1428 (2009)
22. Xu, Y., Quan, Y., Ling, H., Ji, H.: Dynamic texture classification using dynamic fractal analysis. In: ICCV (2011)
23. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and multiple instance learning. PAMI 32(2), 288–303 (2010)
24. Doretto, G., Chiuso, A., Wu, Y., Soatto, S.: Dynamic textures. IJCV 51(2), 91–109 (2003)
25. Ghanem, B., Ahuja, N.: Maximum Margin Distance Learning for Dynamic Texture Recognition. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 223–236. Springer, Heidelberg (2010)
26. Ravichandran, A., Chaudhry, R., Vidal, R.: View-invariant dynamic texture recognition using a bag of dynamical systems. In: CVPR, pp. 1651–1657 (2009)
27. Derpanis, K., Wildes, R.: Dynamic texture recognition based on distributions of spacetime oriented structure. In: CVPR, pp. 191–198 (2010)
28. Péteri, R., Fazekas, S., Huiskes, M.: Dyntex: A comprehensive database of dynamic textures. Pattern Recognition Letters 31(12), 1627–1632 (2010)
29. Kim, T., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. PAMI 29(6), 1005–1018 (2007)
30. Lui, Y., Beveridge, J., Kirby, M.: Action classification on product manifolds. In: CVPR, pp. 833–839 (2010)
31. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: ICPR, vol. 3, pp. 32–36 (2004)
32. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. PAMI 29(12), 2247–2253 (2007)
33. Le, Q., Zou, W., Yeung, S., Ng, A.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR, pp. 3361–3368 (2011)