Video Covariance Matrix Logarithm for Human Action Recognition in Videos

Piotr Bilinski, Francois Bremond

Video Covariance Matrix Logarithm for Human Action Recognition in Videos

Piotr Bilinski and Francois Bremond
INRIA Sophia Antipolis, STARS team
2004 Route des Lucioles, BP93, 06902 Sophia Antipolis, France
{Piotr.Bilinski,Francois.Bremond}@inria.fr

Abstract

In this paper, we propose a new local spatio-temporal descriptor for videos and we propose a new approach for action recognition in videos based on the introduced descriptor. The new descriptor is called the Video Covariance Matrix Logarithm (VCML). The VCML descriptor is based on a covariance matrix representation, and it models relationships between different low-level features, such as intensity and gradient. We apply the VCML descriptor to encode appearance information of local spatio-temporal video volumes, which are extracted by the Dense Trajectories. Then, we present an extensive evaluation of the proposed VCML descriptor with the Fisher vector encoding and the Support Vector Machines on four challenging action recognition datasets. We show that the VCML descriptor achieves better results than the state-of-the-art appearance descriptors. Moreover, we present that the VCML descriptor carries complementary information to the HOG descriptor and their fusion gives a significant improvement in action recognition accuracy. Finally, we show that the VCML descriptor improves action recognition accuracy in comparison to the state-of-the-art Dense Trajectories, and that the proposed approach achieves superior performance to the state-of-the-art methods.

1 Introduction

Various evaluations of local spatio-temporal descriptors [Wang *et al.*, 2009; 2011] have shown that both motion and appearance descriptors are necessary to achieve good results for action recognition in videos. The existing evaluations typically consider Histogram of Oriented Gradients (HOG) descriptor to represent appearance information and Trajectory shape, Histogram of Oriented Flow (HOF), and Motion Boundary Histogram (MBH) descriptors to represent motion information. The same evaluations have shown that motion descriptors typically work better than appearance descriptors, and there are two possible reasons for that: simply, the motion information is more important for action recognition or the existing appearance based descriptors are not discriminative enough. As still image based human action recognition [Guo and Lai, 2014] has shown to achieve good results and they do not use temporal information, we believe that more discriminative appearance descriptors could be proposed. Therefore, in this paper we focus primary on modeling appearance information for action recognition.

The above descriptors, *i.e*. HOG, HOF, and MBH, are based on a 1-dimensional histogram representation of individual features, and they directly model values of given features. However, the joint statistics between individual features are ignored by these descriptors, whereas such information may be informative. Therefore, these descriptors might not be discriminative enough to recognize similar actions.

In image processing, a novel trend has emerged that ignores explicit values of given features, focusing on their pairwise relations instead. A relation between features is well explained in the covariance, which is a measure of how much random variables change together. Covariance provides a measure of the strength of the correlation between features.

2 Covariance and the Related Work

Covariance based features have been introduced to Computer Vision for object detection and texture classification [Tuzel *et al.*, 2006]. They have also been successfully applied for object tracking [Porikli *et al.*, 2006], shape modeling [Wang *et al.*, 2007] and face recognition [Pang *et al.*, 2008]. Moreover, they have been applied for single hand gesture recognition [Harandi *et al.*, 2012] and person re-identification [Bak and Bremond, 2013]. The above techniques are very successful in many topics but, as they use figure-centric representations based on grid representations, they cannot be applied for action recognition in realistic and unconstrained scenarios, where multiple people, pose and camera viewpoint variations, background clutter, occlusions, *etc*. occur. Covariance based features have also been applied **for action recognition**. [Guo *et al.*, 2013] have modeled a whole video sequence using a covariance based representation and they have applied a sparse linear representation framework to recognize actions. One of the main drawbacks of the proposed approach is that it requires precise segmentation, which is very difficult to obtain in real world videos. [Sanin *et al.*, 2013] have also modeled a video using covariance representation, and they have applied the weighted Riemannian locality preserving projection and boosting. One of the main limitations is that the co-

variance is calculated over the whole video volume cropped by the people localization results. In real world videos, it is very difficult to obtain the precise localization of people in every video frame. Moreover, videos often contain multiple people, therefore, the authors have used manual people annotations for the action recognition experiment. Instead of modeling a whole video sequence using a single covariance based representation, [Yuan *et al.*, 2009] have applied covariance based features for local spatio-temporal interest points [Dollar *et al.*, 2005]. As input features for covariance calculation, they have applied the position of interest points, a gradient, and an optical flow. Then, each video sequence is represented by an occurrence histogram of covariance based features, and the Earth Mover's Distance (with the L2 norm as the ground distance) is applied to match pairs of video sequences. Finally, the Nearest Neighbor is used for classification. One of the main limitations of this approach (as well as previously mentioned descriptors) is the lack of structural information in a descriptor; a given spatio-temporal video volume is modeled using a single covariance based representation. Moreover, the authors have computed video representations with different sizes of histograms, and as the result they have not taken the advantage of powerful metrics developed to match histograms (*e.g.* χ^2 distance and histogram intersection distance).

The state-of-the-art covariance based methods have significant limitations, they are applicable only for simple scenarios (single person, static camera, lack of occlusions, *etc.*), and they receive significantly lower accuracy than the remaining state-of-the-art techniques, *e.g.* [Yuan *et al.*, 2009] evaluate their descriptor only on the simplest datasets, *i.e.* Weizmann (2005) and KTH (2004) and receive 90% and 79.7% accuracy, respectively. All the recent action recognition techniques obtain 100% or nearly 100% on these datasets for the last few years, and therefore these datasets are no longer in use. Although the idea of covariance was used for action recognition in the past, there are many ways to apply covariance, and the real challenge is to find efficient representation which gets a high performance and can be applied for realistic and unconstrained scenarios.

As opposed to the existing techniques, we introduce a new covariance based local spatio-temporal descriptor for videos and we propose a new approach for action recognition based on the introduced descriptor. Our descriptor applies the covariance measure in a new way, different than the state-of-the-art techniques, and it can be applied to realistic and unconstrained scenarios. Moreover, we show that our descriptor is complementary to the histogram based features of the (Improved) Dense Trajectories, which have shown to achieve the best results on various action recognition challenges and complex datasets. Our technique outperforms the (Improved) Dense Trajectories and the current state-of-the-art. In comparison, none of the state-of-the-art covariance based descriptors achieves accuracy close to the (Improved) Dense Trajectories; moreover, they cannot be applied on realistic and unconstrained videos. The **key distinctions** of our method are: it does not require segmentation (any video applicable), it uses spatio-temporal structural information (5% improvement), it is used to represent texture along the trajectories (never before), and it outperforms existing covariance based descriptors for action recognition in videos.

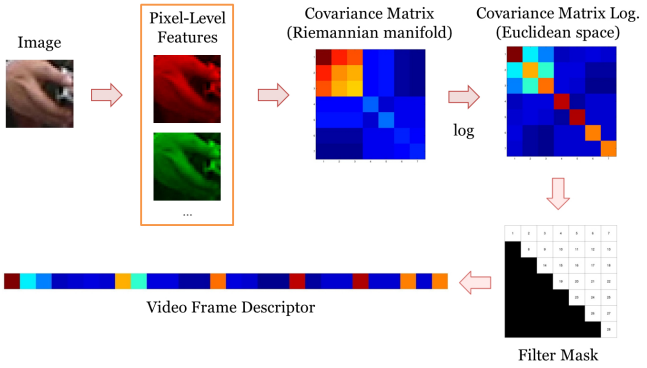

Figure 1: Overview of the video frame descriptor calculation. Firstly, we extract pixel-level features and we represent them using the covariance matrix. Then, we map the covariance matrix from the Riemannian manifold to the Euclidean space. We use the symmetric property of the covariance matrix and we apply a filter mask extracting the upper triangular part of the matrix. Finally, we represent the result in a vector called the video frame descriptor.

The remainder of the paper is organized as follows. In Section 3, we propose the Video Covariance Matrix Logarithm descriptor. Section 4 presents our action recognition framework. In Section 5, we present experimental results, comparison, and analysis. Finally, we conclude in Section 6.

3 Video Covariance Matrix Logarithm

In this section, we propose a new descriptor to encode a local spatio-temporal video volume. The new descriptor is called the Video Covariance Matrix Logarithm (VCML). It is based on a covariance matrix representation and it models relationships between low-level features.

In Section 3.1, we propose a video frame descriptor, and in Section 3.2, we present low-level, *i.e.* pixel-level, features that we use to compute the video frame descriptor. Similarly to the most popular and powerful action recognition local spatio-temporal descriptors, *i.e.* HOG, HOF, and MBH descriptors, we base our descriptor on the representation of individual frames. Finally, in Section 3.3, we propose a video volume descriptor, which is an extension of the video frame descriptor to the spatio-temporal domain of videos.

3.1 Video Frame Descriptor

We are given a single video frame t of spatial size $n_x \times n_y$ pixels, and our goal is to create its discriminative and compact representation. The overview of the calculation process is presented in Figure 1.

Firstly, we calculate low-level (*i.e.* pixel-level) features, *e.g.* intensities in red, green, and blue channels (see Section 3.2). For each pixel of a given video frame, we extract d low-level features. Therefore, we represent a video frame t by a set $\{f_{(x,y,t)}\}_{1 \leq x \leq n_x, 1 \leq y \leq n_y}$ of d-dimensional

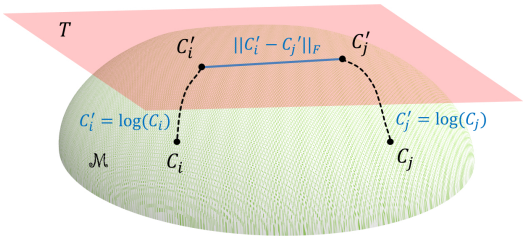

Figure 2: Two sample covariance matrices C_i and C_j are projected from a two-dimensional manifold $\mathcal{M}$ to the tangent space $\mathcal{T}$ via matrix logarithm operation $\log(\cdot)$. Then, the difference between the two covariance matrices can be calculated as the subtraction of the projected matrices, followed by the Frobenius norm $\|\cdot\|_F$.

feature vectors ($f_{(x,y,t)} \in \mathcal{R}^d$). Such a frame representation is typically of high dimension ($n_x \times n_y \times d$), and thus it is necessary to transform it into a more compact representation. For simplicity, we denote the set $\{f_{(x,y,t)}\}_{1 \le x \le n_x, 1 \le y \le n_y}$ as $\{f_{(k,t)}\}_{k=1...n}$, where n is the number of pixels in each video frame ($n = n_x \times n_y$).

We propose to represent each video frame t via covariance matrix (also known as dispersion matrix or variance-covariance matrix). The covariance matrix encodes the variance within each feature and the covariance between different features. The covariance matrix is defined as:

$$C_t = \frac{1}{n-1} \sum_{k=1}^{n} (f_{(k,t)} - \mu_t)(f_{(k,t)} - \mu_t)^T, \qquad (1)$$

where μ_t is the mean of the feature vectors, *i.e.* $\mu_t = \frac{1}{n} \sum_{k=1}^{n} f_{(k,t)}$. Therefore, we transform a video frame representation of size $n_x \times n_y \times d$ into a tensor C_t of size $d \times d$.

Covariance matrices are symmetric and positive semidefinite (nonnegative definite) matrices, and they can be represented as a connected Riemannian manifold. The tensor space of the covariance matrices is a manifold, that is not a vector space with the usual additive structure. Since the Euclidean norm does not correctly capture the distance between two covariance matrices, we need to apply a Riemannian metric in order to use the covariance matrix based descriptors with a local feature encoding technique. There are two popular distance metrics for covariance matrices, which are defined on the Riemannian manifold: Affine-Invariant Riemannian Metric [Forstner and Moonen, 1999] and Log-Euclidean Riemannian Metric [Arsigny *et al.*, 2006], and they both provide results very similar to each other [Arsigny *et al.*, 2006]. Our goal is to use the covariance matrix based features with a local feature encoding technique. We need to create a codebook and the codebook is typically created using a clustering algorithm. Since covariance matrices do not form a Euclidean vector space, standard clustering algorithms cannot be used effectively. Clustering on the Riemannian manifold is time-consuming and it is still an open research problem. Therefore, we use the Log-Euclidean Riemannian metric, which defines a distance between two covariance matrices C_i and C_j as:

$$dist(C_i, C_j) = \|\log(C_i) - \log(C_j)\|_F, \qquad (2)$$

where $\log(\cdot)$ is the matrix logarithm, and $\|\cdot\|_F$ is the Frobenius norm of a matrix. According to this metric, we can map covariance matrices from the Riemannian manifold to the Euclidean space using the matrix logarithm operation (see Figure 2). We apply the Singular Value Decomposition, which decomposes the covariance matrix into 3 matrices:

$$C_t = U \Sigma U^T, \qquad (3)$$

where U is the orthonormal matrix of size $d \times d$, and Σ is the square diagonal matrix with nonnegative real numbers, eigenvalues $(\lambda_1, \lambda_2, ..., \lambda_d)$, on the diagonal. Then, the new representation of the covariance matrix is:

$$C_t^{(log)} = \log(C_t) = U \Sigma' U^T, \qquad (4)$$

where Σ' is the square diagonal matrix of size $d \times d$ with logarithm values of the eigenvalues on the diagonal $(\log(\lambda_1), \log(\lambda_2), ..., \log(\lambda_d))$.

The covariance matrix is a symmetric matrix, and thus it is determined by $\frac{d(d+1)}{2}$ values, forming the upper or lower triangular part of the covariance matrix. To represent a single video frame and create its compact representation, we apply a filter mask extracting all the entries on and above (below) the diagonal of the covariance matrix. We represent these values in a form of a vector V_t:

$$V_t = \text{triu}(C_t^{(log)}), \qquad (5)$$

where $\text{triu}(\cdot)$ is the filter mask operation.

Therefore, we transform a video frame representation of size $n_x \times n_y \times d$ into a compact vector V_t of size $\frac{d(d+1)}{2}$. The obtained feature vector V_t is called the video frame descriptor.

3.2 Low-Level Features

In this section, we present the extraction of low-level features in a single video frame. As mentioned before, we focus on the representation of the appearance information.

For every pixel in each frame of the given video volume, we extract seven low-level, *i.e.* pixel-level, appearance features. We extract normalized intensities in red, green, and blue channels, and first and second order derivatives of gray scale intensity image along "x" and "y" axes, as in [Tuzel *et al.*, 2006]. Thus, every pixel is represented in the form:

$$f = \left[R,\ G,\ B,\ \frac{\partial I}{\partial x},\ \frac{\partial I}{\partial y},\ \frac{\partial^2 I}{\partial x^2},\ \frac{\partial^2 I}{\partial y^2} \right], \qquad (6)$$

where R, G, and B are the red, green, and blue intensity channels, and I is the corresponding gray scale intensity image. An example of the extracted seven low-level appearance features is presented in Figure 3.

The covariance representation based on the above features provides a rotation invariant representation of a video frame. However, the relationships between these low-level features and the spatial positions of these features may be informative and useful for action recognition. Therefore, we also use the extended set of low-level features, where every pixel is represented in the following form:

$$f' = \left[X,\ Y,\ R,\ G,\ B,\ \frac{\partial I}{\partial x},\ \frac{\partial I}{\partial y},\ \frac{\partial^2 I}{\partial x^2},\ \frac{\partial^2 I}{\partial y^2} \right], \qquad (7)$$

where X and Y represent the spatial position of a pixel in a video frame, and the remaining features are as before.

Figure 3: Seven low-level appearance features extracted in a sample video frame from the URADL dataset.

3.3 Video Volume Descriptor

We are given a spatio-temporal video volume of size $n_x \times n_y \times n_t$, of spatial size $n_x \times n_y$ pixels, and of temporal size n_t video frames, and our goal is to create its discriminative and compact representation.

Firstly, we use the spatio-temporal grid to encode structural information of the video volume. Thus, we treat an input spatio-temporal video volume as a cuboid, and we divide it into a spatio-temporal grid, where each cell of the grid is of size $g_x \times g_y \times g_t$, of spatial size $g_x \times g_y$ pixels, and of temporal size g_t video frames.

For each video frame in each cell of the grid, we compute a separate video frame descriptor V_t, as explained in Section 3.1. Then, to create a compact cell representation, we describe each cell of the grid as a mean vector of all video frame representations calculated inside this cell:

$$V_{cell} = \frac{1}{g_t} \sum_{t=1}^{g_t} \text{triu}(C_t^{(log)}). \tag{8}$$

Finally, we define the Video Covariance Matrix Logarithm (VCML) descriptor D as the concatenation of all the descriptors from all cells of the grid:

$$D = [\, V_{cell_1},\, V_{cell_2},\, ...,\, V_{cell_m}\,]^T, \tag{9}$$

where m is the number of cells of the spatio-temporal grid.

Note that we can significantly speed-up the covariance matrix calculation process using Integral Images [Tuzel *et al.*, 2006] and extending the idea from images to the spatio-temporal domain of videos.

4 Approach Overview

We present our action recognition framework based on the introduced VCML descriptor. Firstly, we extract local spatio-temporal video volumes from a given video sequence (Section 4.1). Then, we represent each local video volume by the proposed VCML and complementary descriptors (Section 4.2). Then, we use the extracted descriptors to create video sequence representations (Section 4.3). Finally, we classify videos into action categories (Section 4.4).

4.1 Local Spatio-Temporal Video Volumes

In order to extract local spatio-temporal video volumes, we compute the Dense Trajectories in a video sequence [Wang *et al.*, 2011; Wang and Schmid, 2013]; we apply a dense sampling to extract interest points and we track these interest points using a dense optical flow field. Then, we extract local spatio-temporal video volumes around the detected trajectories. By extracting dense trajectories, we provide a good coverage of a video sequence and we ensure extraction of meaningful features. The Dense Trajectories were selected based on their use in the recent literature. However, the VCML descriptor can be used to represent local spatio-temporal video volumes extracted by any other algorithm, *e.g.* by the Spatio-Temporal Interest Points detector [Laptev, 2005].

4.2 Local Spatio-Temporal Video Features

Then, we use the proposed Video Covariance Matrix Logarithm descriptor (Section 3) to represent the extracted spatio-temporal video volumes. Moreover, we calculate the Trajectory shape, Histogram of Oriented Gradients, Histogram of Optical Flow, and Motion Boundary Histogram descriptors for each local spatio-temporal video volume, as these descriptors carry complementary information about the visual appearance and visual motion.

4.3 Action Representation

Once the descriptors are extracted, we use them to create video representations. We encode a video sequence using first and second order statistics of a distribution of a feature set, based on the idea of the Fisher vector encoding [Perronnin *et al.*, 2010], which has shown to achieve excellent results as a global descriptor both for image classification and retrieval. We model features with a generative model and compute the gradient of their likelihood w.r.t. the parameters of the model. We describe how the set of features deviates from an average distribution of features, modeled by a parametric generative model. Each video is represented by a $2DK$-dimensional Fisher vector for each descriptor type, where D is the descriptor size and K is the number of Gaussians. Then, we apply the power normalization, which can be seen as explicitly applying non-linear additive kernel (the Hellinger's kernel), and then the obtained vector is further normalized by the L2 norm, as in [Perronnin *et al.*, 2010]. The above representation is calculated for each descriptor type, and finally, to combine different descriptor types, we concatenate their normalized Fisher vectors.

4.4 Action Recognition

We use linear Support Vector Machines (SVMs) [Chang and Lin, 2011] for action classification. Linear SVMs have shown to provide very good and promising results with high-dimensional data such as Fisher vectors. Typically, if the number of features is large, there is no need to map data to a higher dimensional space and the non-linear mapping does not improve the accuracy [Hsu *et al.*, 2003]. Moreover, linear SVMs have shown to be efficient both in training and prediction steps. We implement the one-vs-all approach for multi-class classification.

5 Experiments

We present an evaluation, comparison and analysis of the proposed VCML descriptor and action recognition approach on 4 state-of-the-art action recognition datasets: URADL, MSR Daily Activity 3D, UCF50 and HMDB51.

5.1 Datasets

This section briefly presents the four datasets used in our experiments. In all the experiments we follow the recommended evaluation protocols provided by the authors of the dataset.

The **URADL** dataset [Messing *et al.*, 2009] contains 10 types of human activities of daily living, selected to be useful for an assisted cognition task. Each action is performed 3 times by 5 different people and actions are difficult to separate on the basis of a single source of information (eating a banana vs. eating snack chips, answering a phone vs. dialing a phone). We use the leave-one-person-out cross-validation evaluation scheme to report the performance.

The **MSR Daily Activity 3D** dataset [Wang *et al.*, 2012] consists of 16 daily living activities. Each action is performed by 10 subjects and each subject performs each action in standing and sitting position, what adds an additional intra-class variation. We use the leave-one-person-out cross-validation evaluation scheme to report the performance.

The **UCF50** dataset [Reddy and Shah, 2012] contains 50 action categories and 6618 video sequences. The dataset contains real word videos collected from YouTube ranging from general sports to daily life exercises. Videos are divided into 25 folds and we follow the recommended 25-folds cross-validation to report the performance.

The **HMDB51** dataset [Kuehne *et al.*, 2011] contains 51 action categories and 6766 video sequences. The dataset is collected from a variety of sources ranging from digitized movies to YouTube videos. We use 3 train-test splits provided by the authors of this dataset and we report average accuracy over the 3 splits. Note that we use the original videos and not the stabilized ones.

The **datasets differ** in the number of actions, number of training samples per action, inter and intra class variations, motion blur, pose and camera view point variations, background clutter, occlusions, illumination conditions, *etc*. which affect the action recognition accuracy.

5.2 Implementation Details

To estimate the GMM parameters for the Fisher vector encoding, we randomly sample a subset of 100k features from the training set. We consider 6 various codebook sizes $K = \{2^i\}_{i=4}^{9}$ for the URADL and the MSR datasets and 4 codebook sizes $K = \{2^i\}_{i=4}^{7}$ for the UCF50 and the HMDB51 datasets. We set the number of Gaussians (*i.e*. the codebook size) using the leave-one-person-out cross-validation for the URADL and the MSR datasets, leave-one-fold-out cross-validation for the UCF50, and 5-folds cross-validation for the HMDB51. To increase precision, we initialize the GMM ten times and we keep the codebook with the lowest error. To report the results, we use the mean class accuracy metric. When Fisher Vectors are applied with PCA together, we reduce a descriptor dimensionality by a factor of two, as in [Perronnin *et al.*, 2010].

	1×1×1	1×1×3	2×2×1	2×2×3
HOG	71.33	74.67	79.33	83.33
VCML7	76.67	79.33	80.67	81.33
VCML9	79.33	79.33	**84.00**	**84.00**

Table 1: Evaluation of the HOG and the VCML descriptors using various spatio-temporal grids on the URADL dataset.

	VCML7	VCML9
MSR Daily Activity 3D	56.88	59.38
HMDB51	24.68	27.10

Table 2: Evaluation of the VCML descriptors on the MSR Daily Activity 3D and the HMDB51 datasets.

5.3 Importance of Spatio-Temporal Grid

Firstly, we evaluate the influence of the spatio-temporal grid on the performance of the HOG and the VCML descriptors (using 7 low-level features (VCML7 descriptor) and 9 low-level features (VCML9 descriptor), see Section 3.2). The results using the URADL dataset are presented in Table 1. The obtained results present the superior performance of the VCML9 descriptor over the HOG and the VCML7 descriptors and they clearly show that the spatio-temporal grid significantly improves action recognition accuracy. The spatial grid increases the accuracy more than the temporal grid, but the best results are achieved by the spatio-temporal grid (2×2×3), which is used in the following evaluations. The results confirm the importance of using the structural information in a descriptor, which is missing in the existing state-of-the-art action recognition covariance based descriptors.

5.4 Comparison of VCML7 and VCML9

We evaluate the performance of the VCML descriptor with 7 low-level features (VCML7 descriptor) and 9 low-level features (VCML9 descriptor) on the URADL, MSR Daily Activity 3D and HMDB51 datasets. The results are presented in Table 1 and Table 2. The VCML9 descriptor shows superior performance to the VCML7 descriptor. This confirms that the relationships between low-level features and the spatial positions of these features are informative and useful for action recognition. Therefore, in the following evaluations we use the VCML descriptor with the 9 low-level features.

5.5 Improving HOG with VCML

Then, we evaluate the performance of the HOG and the VCML descriptors on the URADL, MSR Daily Activity 3D and HMDB51 datasets. Moreover, we evaluate the fusion of the HOG and the VCML descriptors. The experiments are performed without and with the use of the Principal Component Analysis (PCA). The results are available in Table 3. We observe that the VCML descriptor achieves better results than the HOG descriptor on the URADL and the HMDB51 datasets, and similar results on the MSR Daily Activity 3D dataset. Moreover, the fusion of the HOG and the VCML descriptors significantly improves action accuracy. This confirms that the HOG and the VCML are complementary to

Dataset	Descriptor(s)	Accuracy	Accuracy with PCA
URADL	HOG	83.33	86.67
	VCML	84.00	88.00
	VCML + HOG	88.00	92.67
MSR	HOG	60.94	59.69
	VCML	59.38	54.38
	VCML + HOG	63.44	63.13
HMDB51	HOG	25.64	33.14
	VCML	27.10	36.34
	VCML + HOG	37.10	40.52

Table 3: Evaluation results of the HOG and VCML descriptors with and without the PCA. Moreover, we present the fusion of the descriptors.

Dataset	Descriptor(s)	Accuracy	Accuracy with PCA
URADL	DT	94.00	92.67
	VCML + DT	94.00	94.00
MSR	DT	76.25	75.31
	VCML + DT	78.13	76.25
HMDB51	DT	47.02	50.65
	VCML + DT	50.92	52.85

Table 4: Evaluation results of the Dense Trajectories (DT) descriptors (*i.e*. Trajectory shape, HOG, HOF, and MBH) and the fusion of the DT and the VCML, with and without the PCA.

each other, as the former directly models low-level features and the latter models relations between low-level features.

5.6 Improving DT with VCML

Then, we evaluate the Dense Trajectories (DT) representation [Wang *et al.*, 2011] (*i.e*. Trajectory shape, HOG, HOF and MBH) on the URADL, MSR Daily Activity 3D and HMDB51 datasets. As before, using all the time the same evaluation framework, for each descriptor we compute a separate video representation, which are then concatenated. Moreover, we evaluate the fusion of the VCML and the Dense Trajectories (DT) representations. The experiments are performed without and with the use of the PCA. The results are available in Table 4 and they clearly show that the fusion of the VCML and DT improves action recognition accuracy in comparison to the DT representation alone. This is natural as the VCML and the HOG capture complementary information about the visual appearance, and the Trajectory shape, HOF and MBH capture information about the visual motion.

Descriptor(s)	URADL	MSR	HMDB51
HOG	83.33	60.94	25.64
HOG + PCA	86.67	59.69	33.14
HOG3D	–	–	33.3
VCML	88	59.38	36.34
VCML + HOG	**92.67**	**63.44**	**40.52**

Table 5: Comparison of local spatio-temporal appearance descriptors on the URADL, MSR Daily Activity 3D and HMDB51 datasets.

5.7 Improving IDT with VCML

More recently, [Wang and Schmid, 2013] have proposed the Improved Dense Trajectories (IDT), which improve the Dense Trajectories (DT) [Wang *et al.*, 2011] taking into account camera motion to correct them. In Table 6 and Table 7 we present the accuracy of the DT and the IDT on the UCF50 and the HMDB51 datasets. Moreover, we compare the proposed action recognition approach (the DT and the IDT represented by the VCML representation) on all the four datasets. The IDT show superior performance to the DT, and the VCML with the IDT show superior performance to the VCML with the DT. The IDT remove background trajectories, which is particular useful for the UCF50 and HMDB51 datasets containing real world videos with significant camera motion. The MSR dataset does not contain camera motion but it contains people moving in the background. Their trajectories are removed by using the results of the human detector, which is applied by the IDT. The smallest improvement is on the URADL dataset (no camera motion, no moving people in the background).

5.8 Comparison with the State-of-The-Art

We compare VCML and the fusion of VCML and HOG with the state-of-the-art appearance descriptors (*i.e*. HOG [Wang *et al.*, 2011], HOG + PCA, HOG3D [Klaser *et al.*, 2008; Shi *et al.*, 2013]). The results are presented in Table 5 and they show the superior performance of the VCML with HOG over other local spatio-temporal appearance descriptors.

Then, we compare the proposed action recognition approach with the state-of-the-art (see Table 6 and Table 7). Our approach achieves the best score on the MSR, UCF50 and HMDB51 datasets, which are among the most challenging datasets. Our approach achieves the second best score on the URADL dataset, *i.e*. [Rostamzadeh *et al.*, 2013] achieve better performance in this specific dataset, however, they use a human body part detector and their method fails in real world videos, *e.g*. UCF50 and HMDB51 datasets.

For fair comparison, we have not included the following. [Simonyan and Zisserman, 2014] who achieve better performance on the HMDB51, however, they do not follow the standard evaluation protocol and they use different and extended data for training, which is not provided by the authors of the HMDB51. We are also aware of methods using additional depth data with the MSR Daily Activity 3D, however, in this paper we only use the RGB videos as the depth data is limited by the range of the Microsoft Kinect device and is not available in typical real world videos.

5.9 Results Summary

We use 4 datasets for experiments and we receive the best results outperforming the state-of-the-art. The VCML improves HOG (15% on the HMDB51) and improves 2 types of

URADL		MSR Daily Activity 3D		UCF50	
[Benabbas *et al.*, 2010]	81.0	[Koperski *et al.*, 2014]	72.0	[Kantorov and Laptev, 2014]	82.2
[Raptis and Soatto, 2010]	82.7	JPF [Wang *et al.*, 2012]	78.0	[Shi *et al.*, 2013]	83.3
[Messing *et al.*, 2009]	89.0	[Oreifej and Liu, 2013]	80.0	[Oneata *et al.*, 2013]	90.0
[Bilinski and Bremond, 2012]	93.3	AE [Wang *et al.*, 2012]	85.7	[Wang and Schmid, 2013]	91.2
Dense Trajectories	94.0	Dense Trajectories	76.2	Dense Trajectories	84.2
Our Approach (DT)	94.0	Our Approach (DT)	78.1	Our Approach (DT)	88.1
Our Approach (IDT)	**94.7**	**Our Approach (IDT)**	**85.9**	**Our Approach (IDT)**	**92.1**

Table 6: Comparison with the state-of-the-art on the URADL, MSR Daily Activity 3D and UCF50 datasets.

HMDB51	
[Kantorov and Laptev, 2014]	46.7
[Jain *et al.*, 2013]	52.1
[Oneata *et al.*, 2013]	54.8
[Wang and Schmid, 2013]	57.2
Dense Trajectories	47.0
Our Approach (DT)	52.9
Our Approach (IDT)	**58.6**

Table 7: Comparison with the state-of-the-art on the HMDB51 dataset.

trajectory description, *i.e*. DT (5.9%) and IDT (1.4%). The combination of VCML and DT shows significant improvement on UCF50 ($\sim$4%) and HMDB51 ($\sim$6%). UCF50 and HMDB51 are the most challenging and very competitive action recognition datasets, therefore the obtained improvement over the best state-of-the-art methods can be seen as significant, especially if the result is above 92% like for UCF50 dataset.

6 Conclusions

We have proposed a new local descriptor for videos to encode local spatio-temporal video volumes. The new descriptor is called the Video Covariance Matrix Logarithm (VCML). The VCML is based on a covariance matrix representation and it models linear relationships between low-level features.

We have applied the VCML descriptor to encode appearance information of local spatio-temporal video volumes. Using the Fisher vectors and the Support Vector Machines, we have presented an extensive evaluation of the VCML descriptor on four challenging action recognition datasets. In comparison with the most popular visual appearance descriptor, *i.e*. the HOG descriptor, the VCML achieves superior results. The experiments have shown that the additional accuracy increase can be achieved by the fusion of these two descriptors. This is not surprising as the HOG and the VCML descriptors are complementary to each other. The former directly models low-level features and the latter models relations between low-level features. Finally, we have presented that the VCML descriptor improves action recognition accuracy in comparison to the state-of-the-art (Improved) Dense Trajectories and achieves superior results over the state-of-the-art.

To the best of our knowledge, this is the first time when the covariance based features are used to represent the dense trajectories. Moreover, this is the first time when they encode the structural information and they are applied with the Fisher vectors for action recognition.

Acknowledgements.

This work was supported by the CENTAUR and Dem@Care projects. However, the views and opinions expressed herein do not necessarily reflect those of the financing institutions.

References

[Arsigny *et al.*, 2006] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic resonance in medicine*, 56(2):411–421, 2006.

[Bak and Bremond, 2013] Slawomir Bak and Francois Bremond. Re-identification by Covariance Descriptors. In Shaogang Gong, Marco Cristani, Yan Shuicheng, and Chen Change Loy, editors, *Person Re-Identification*, Advances in Computer Vision and Pattern Recognition. Springer, December 2013.

[Benabbas *et al.*, 2010] Yassine Benabbas, Adel Lablack, Nacim Ihaddadene, and Chabane Djeraba. Action recognition using direction models of motion. In *ICPR*, 2010.

[Bilinski and Bremond, 2012] Piotr Bilinski and Francois Bremond. Contextual statistics of space-time ordered features for human action recognition. In *AVSS*, 2012.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2:27:1–27:27, 2011.

[Dollar *et al.*, 2005] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *ICCCN*, 2005.

[Forstner and Moonen, 1999] Wolfgang Forstner and Boudewijn Moonen. A metric for covariance matrices. In *Festschrift for Erik W. Grafarend on the occasion of his 60th birthday*, pages 113–128, 1999.

[Guo and Lai, 2014] Guodong Guo and Alice Lai. A survey on still image based human action recognition. *PR*, 47(10):3343 – 3361, 2014.

[Guo *et al.*, 2013] Kai Guo, Prakash Ishwar, and Janusz Konrad. Action recognition from video using feature covariance matrices. *IEEE TIP*, 22(6):2479–2494, 2013.

[Harandi *et al.*, 2012] Mehrtash Tafazzoli Harandi, Conrad Sanderson, Arnold Wiliem, and Brian C Lovell. Kernel analysis over riemannian manifolds for visual recognition of actions, pedestrians and textures. In *WACV*, 2012.

[Hsu *et al.*, 2003] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003.

[Jain *et al.*, 2013] Mihir Jain, Herve Jegou, and Patrick Bouthemy. Better exploiting motion for better action recognition. In *CVPR*, 2013.

[Kantorov and Laptev, 2014] Vadim Kantorov and Ivan Laptev. Efficient feature extraction, encoding and classification for action recognition. In *CVPR*, 2014.

[Klaser *et al.*, 2008] Alexander Klaser, Marcin Marszalek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[Koperski *et al.*, 2014] Michal Koperski, Piotr Bilinski, and Francois Bremond. 3D Trajectories for Action Recognition. In *ICIP*, 2014.

[Kuehne *et al.*, 2011] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A Large Video Database for Human Motion Recognition. In *ICCV*, 2011.

[Laptev, 2005] Ivan Laptev. On space-time interest points. *IJCV*, 64(2-3):107–123, 2005.

[Messing *et al.*, 2009] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, 2009.

[Oneata *et al.*, 2013] Dan Oneata, Jakob Verbeek, and Cordelia Schmid. Action and Event Recognition with Fisher Vectors on a Compact Feature Set. In *ICCV*, 2013.

[Oreifej and Liu, 2013] Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *CVPR*, 2013.

[Pang *et al.*, 2008] Yanwei Pang, Yuan Yuan, and Xuelong Li. Gabor-based region covariance matrices for face recognition. *IEEE TCSVD*, 18(7):989–993, 2008.

[Perronnin *et al.*, 2010] Florent Perronnin, Jorge Sanchez, and Thomas Mensink. Improving the Fisher Kernel for Large-Scale image classification. In *ECCV*, 2010.

[Porikli *et al.*, 2006] Fatih Porikli, Oncel Tuzel, and Peter Meer. Covariance tracking using model update based means on riemannian manifolds. In *CVPR*, 2006.

[Raptis and Soatto, 2010] Michalis Raptis and Stefano Soatto. Tracklet Descriptors for Action Modeling and Video Analysis. In *ECCV*, 2010.

[Reddy and Shah, 2012] Kishore K. Reddy and Mubarak Shah. Recognizing 50 Human Action Categories of Web Videos. *Machine Vision and Applications*, 2012.

[Rostamzadeh *et al.*, 2013] Negar Rostamzadeh, Gloria Zen, Ionut Mironica, Jasper Uijlings, and Nicu Sebe. Daily living activities recognition via efficient high and low level cues combination and fisher kernel representation. In *ICIAP*. 2013.

[Sanin *et al.*, 2013] Andres Sanin, Conrad Sanderson, Mehrtash Tafazzoli Harandi, and Brian C Lovell. Spatio-temporal covariance descriptors for action and gesture recognition. In *WACV*, 2013.

[Shi *et al.*, 2013] Feng Shi, Emil Petriu, and Robert Laganiere. Sampling strategies for real-time action recognition. In *CVPR*, June 2013.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[Tuzel *et al.*, 2006] Oncel Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. In *ECCV*, 2006.

[Wang and Schmid, 2013] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, pages 3551–3558. IEEE, 2013.

[Wang *et al.*, 2007] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. Shape and appearance context modeling. In *ICCV*, 2007.

[Wang *et al.*, 2009] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[Wang *et al.*, 2011] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[Wang *et al.*, 2012] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.

[Yuan *et al.*, 2009] Chunfeng Yuan, Weiming Hu, Xi Li, Stephen J. Maybank, and Guan Luo. Human action recognition under log-euclidean riemannian metric. In *ACCV*, 2009.