Representing Visual Appearance by Video Brownian Covariance Descriptor for Human Action Recognition

Piotr Bilinski, Michal Koperski, Slawomir Bak, François Bremond

Representing Visual Appearance by Video Brownian Covariance Descriptor for Human Action Recognition

Piotr Bilinski, Michal Koperski, Slawomir Bak, Francois Bremond
INRIA
2004 Route des Lucioles, BP 93, 06902, Sophia Antipolis, France
{piotr.bilinski, michal.koperski, slawomir.bak, francois.bremond}@inria.fr

Abstract

This paper addresses a problem of recognizing human actions in video sequences. Recent studies have shown that methods which use bag-of-features and space-time features achieve high recognition accuracy. Such methods extract both appearance-based and motion-based features. This paper focuses only on appearance features. We propose to model relationships between different pixel-level appearance features such as intensity and gradient using Brownian covariance, which is a natural extension of classical covariance measure. While classical covariance can model only linear relationships, Brownian covariance models all kinds of possible relationships. We propose a method to compute Brownian covariance on space-time volume of a video sequence. We show that proposed Video Brownian Covariance (VBC) descriptor carries complementary information to the Histogram of Oriented Gradients (HOG) descriptor. The fusion of these two descriptors gives a significant improvement in performance on three challenging action recognition datasets.

1. Introduction

The automatic recognition of human actions in videos has become one of the most active research topic in computer vision. It plays a key role in many important applications, such as smart surveillance systems, intelligent houses, robots, human-computer interfaces, video data indexing and retrieval, sport event analysis, and virtual reality. Action recognition is a very challenging research problem, mainly due to: intra-class variations, occlusions, background clutter, and viewpoint changes.

Over the last decade, many different action recognition methods have been proposed [4, 7, 8, 17], most of which are based on local spatio-temporal features and bag-of-features approach. The local spatio-temporal features have shown to achieve good accuracy in action recognition over various datasets [2, 18]. They are able to capture appearance and motion. They are robust to viewpoint and scale changes. Moreover, they are easy to implement and quick to calculate.

The most popular local spatio-temporal descriptors for action recognition are: HOG, HOF, and MBH descriptor. The Histogram of Oriented Gradients (HOG) descriptor [8] encodes visual appearance and shape information, whereas the Histogram of Optical Flow (HOF) descriptor [8] and the Motion Boundary Histograms (MBH) descriptor [17] encode motion information.

Wang *et al*. [17] have shown that motion information from HOF descriptor is enough to achieve satisfactory classification performance, but it is not enough to fully describe an action. For example, motion information might distinguish eating a banana from peeling a banana, but motion information might have difficulty distinguishing eating snack chips from eating a banana; conversely, appearance information might be more useful in the second than the first task [10]. Therefore, it is very important to encode also appearance, which can describe an object involved in a particular action. Recent research [18] also have shown that action recognition model can benefit from complementary appearance information, *i.e*. from the HOG descriptor.

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

In image processing, a novel trend has emerged that ignores explicit values of given features, focusing instead on their pairwise relations. The most known example of such an approach is covariance descriptor [16].

In this paper we focus on pixel level features such as an intensity, image gradient and image second derivative. We propose to find relationships between mentioned features using Brownian covariance, which is an extension of

covariance measure. As covariance measures only linear relationship, it might capture insufficient information in a complex environment such as an action recognition.

The described Brownian covariance can catch all kind of relations between two image patches. We propose a new descriptor called Video Brownian Covariance (VBC), which is an extension of Brownian covariance for space-time video volumes. Using the proposed descriptor we represent relations between appearance features extracted from a video sequence.

The main contributions of this paper are:

- We propose a new descriptor (called Video Brownian Covariance), which captures appearance information by employing Brownian covariance. This descriptor captures all kinds of possible relationships between pixel level features.
- To apply Brownian Covariance to videos, we propose to extract local spatio-temporal video volumes using dense trajectories; however, the VBC descriptor is flexible and can by used together with any local spatio-temporal video volume detector. Then, we represent videos using Fisher vectors and we apply Support Vector Machines to classify videos into action categories.
- We confirm our method by experiments on URADL [10], MSRDailyActivity3D [21], and challenging HMDB51 [6] datasets.
- We show that the information provided by the VBC descriptor is complementary to the HOG descriptor. Their combination gives a significant improvement in action recognition accuracy.

2. Related Work

Over the last decade, methods based on local spatio-temporal features and bag-of-features approach have become very popular.

There are several popular techniques proposed for extraction of local spatio-temporal video volumes. Laptev and Lindeberg [7] have proposed Harris3D point detector. Dollar *et al*. [4] have proposed Cuboid detector. Willems *et al*. [20] have proposed Hessian detector. Messing *et al*. [10] have proposed to track Harris feature points using KLT tracker. Wang *et al*. [17] have proposed to use dense sampling and to track such detected points using optical flow and median filtering.

The most popular local spatio-temporal descriptors are: HOG [3, 8], HOF [8], and MBH [17] descriptor.

Covariance based features have been introduced by Tuzel *et al*. for object detection and texture classification [16]. They have been successfully applied for object tracking [14], shape modeling [19], and face recognition [11]. Moreover, covariance based features have been also applied for action recognition [5, 22]. However, covariance measures only linear relationship between features. Second disadvantage is that covariance matrices do not lie on the Euclidean space, and therefore, we cannot directly use them with state-of-the-art machine learning algorithms. Bak *et al*. [1] have proposed Brownian descriptor for person re-identification. The authors look for relationships between pixel locations and intensity gradients. However, their approach is not suitable for action recognition, as it focuses on exact appearance matching.

In contrast, our novel Video Brownian Covariance (VBC) descriptor is able to capture appearance information based on pixel color, 1st and 2nd intensity derivatives. In addition, it handles a dynamic nature of a video as Brownian covariance is calculated on each patch in space-time video volume.

3. Brownian Covariance

The classical covariance descriptor measures only information on linear dependence between features. This might not be enough to capture the complex structure of many objects. Covariance descriptor may produce a diagonal matrix, which is not a sufficient condition for statistical independence; actually, non-monotone dependence may exists. This indicates information loss when using the covariance descriptor.

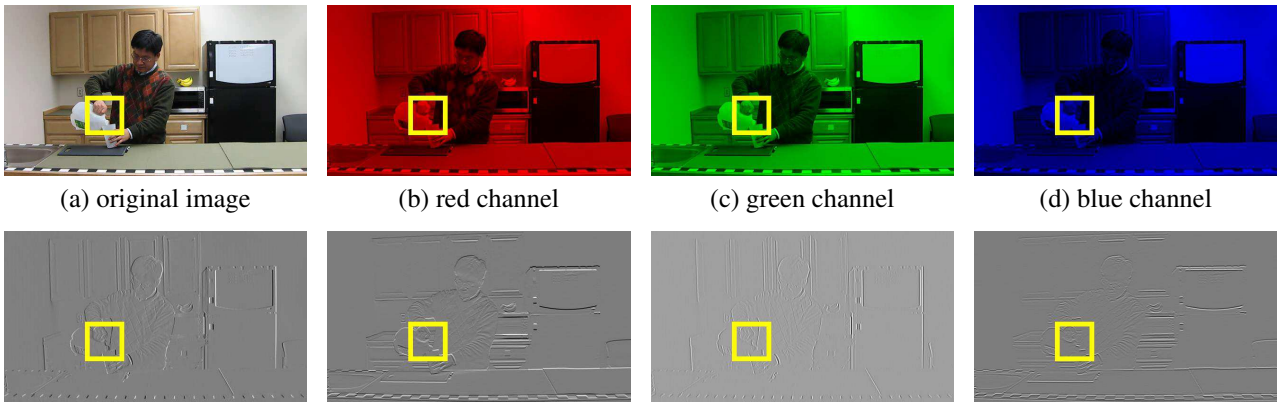
To solve the above issues we propose a novel descriptor for video description based on Brownian covariance [1, 15]. The classical covariance measures only the degree of linear relationship between features, whereas Brownian covariance measures the degree of *all kinds of possible relationships* between features.

3.1. Brownian Covariance

Brownian descriptor is based the theory in mathematical statistics related to the Brownian motion [15]. The descriptor is based on the *distance covariance* statistics that measures the dependence between random vectors in the arbitrary dimension. The mathematical notations and formulas provided in this section are in accordance with [15].

3.1.1 Distance Covariance $\mathcal{V}^2$

Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ be random vectors, where p and q are natural numbers. f_X and f_Y denote the characteristic functions of X and Y, respectively, and their joint characteristic function is denoted as $f_{X,Y}$. In terms of characteristic functions, X and Y are independent if and only if $f_{X,Y} = f_X f_Y$. Thus, a natural way of measuring the dependence between X and Y is to find a suitable norm to measure the distance between $f_{X,Y}$ and $f_X f_Y$.



(a) original image (b) red channel (c) green channel (d) blue channel

(e) 1st deriv. axis-x (f) 1st deriv. axis-y (g) 2nd deriv. axis-x (h) 2nd deriv. axis-y

Figure 1. Low-level appearance features extracted in a video frame. Yellow rectangle indicates a sample patch.

Distance covariance $\mathcal{V}^2$ [15] is a new measure of dependence between random vectors and can be defined as:

$$\mathcal{V}^2(X,Y) = ||f_{X,Y}(t,s) - f_X(t)f_Y(s)||^2 \quad (1)$$

$$= \frac{1}{c_p c_q}\int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s)-f_X(t)f_Y(s)|^2}{|t|_p^{1+p}|s|_q^{1+q}}dtds, \quad (2)$$

where c_p and c_q are constants determining norm function in $\mathbb{R}^p \times \mathbb{R}^q$, $t \in X$, $s \in Y$.

This measure is analogous to classical covariance, but with the important property that $\mathcal{V}^2(X,Y) = 0$ if and only if X and Y are independent.

The paper [15] provides us the following definition of a sample distance covariance $\mathcal{V}_n^2$. For a random sample $(\mathbf{X},\mathbf{Y}) = \{(X_k,Y_k) : k = 1\dots n\}$ of n i.i.d random vectors (X,Y) from their joint distribution, compute the Euclidean distance matrices $(a_{kl}) = (|X_k - X_l|_p)$ and $(b_{kl}) = (|Y_k - Y_l|_q)$. Define:

$$A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}, \quad k,l = 1,\dots,n, \quad (3)$$

where:

$$\bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \;\; \bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}, \;\; \bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}. \quad (4)$$

Similarly, we define $B_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$. Having these simple linear functions of the pairwise distances between sample elements of X and Y distributions, we define distance covariance as:

$$\mathcal{V}_n^2(X,Y) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}. \quad (5)$$

Although, the relation of equations (1) and (5) is not straightforward, THEOREM 2 from [15] justifies it:

If $E|X|_p < \infty$ *and* $E|Y|_q < \infty$, *then almost surely*

$$\lim_{n\to\infty} \mathcal{V}_n(X,Y) = \mathcal{V}(X,Y). \quad (6)$$

Standardization. Similarly to covariance, which has its standardized counterpart ρ, $\mathcal{V}_n^2$ has its standardized version referred to as *distance correlation* $\mathcal{R}_n^2$, defined as:

$$\mathcal{R}_n^2(X,Y) = \begin{cases} \frac{\mathcal{V}_n^2(X,Y)}{\sqrt{\mathcal{V}_n^2(X)\mathcal{V}_n^2(Y)}}, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) > 0; \\ 0, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) = 0, \end{cases} \quad (7)$$

where:

$$\mathcal{V}_n^2(X) = \mathcal{V}_n^2(X,X) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}^2. \quad (8)$$

4. Video Brownian Covariance Descriptor

In this section, we present our Video Brownian Covariance (VBC) descriptor. Our descriptor encodes a space-time video volume of size $S \times S$ pixels and t frames.

4.1. Image Low-Level Appearance Features

For each frame of a video volume, we compute seven low-level appearance features. For every pixel we extract intensities in red, green, and blue channels, first and second order derivatives of grey scale intensity image along x and y axis. Thus, every pixel at time t can be expressed in the following vector form:

$$f_t = \left[R, G, B, \frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}, \frac{\partial^2 I}{\partial x^2}, \frac{\partial^2 I}{\partial y^2}\right], \quad (9)$$

where I is a gray scale intensity image. The examples of the extracted low-level appearance features are presented in Figure 1.

4.2. Video Brownian Covariance Descriptor

For each frame of the video volume we compute Brownian covariance between all mentioned appearance features. We use 7 low level appearance features, thus each frame is represented by a vector of $m = 28$ unique pairwise relations

between features. We define the Video Brownian Covariance (VBC) descriptor D as an average descriptor among all the Brownian covariance descriptors extracted from all video frames:

$$D = [d_1, d_2, ..., d_m], \quad d_i = \frac{1}{t}\sum_{j=1}^{t} b_{i,j}, \tag{10}$$

where $b_{i,j}$ is the i-th value of the Brownian covariance descriptor in the j-th video volume frame.

4.3. Normalization

Each value of our descriptor encodes a different relationship between two appearance features and it has a different meaning. Therefore, to make the descriptor values more uniform, we apply the min-max normalization to each descriptor value separately, *i.e.*:

$$Normalized(D_i) = \frac{D_i - \Theta_i^{min}}{\Theta_i^{max} - \Theta_i^{min}}, \tag{11}$$

where D_i is the i-th value of the descriptor D, Θ_i^{min} is the minimum value among all the i-th values of the training descriptors, and Θ_i^{max} is the maximum value among all the i-th values of the training descriptors.

5. Approach Overview

In the first step of our approach, we extract local spatio-temporal patches from a video sequence, and we represent each patch by our Video Brownian Covariance (VBC) and the Histogram of Oriented Gradients (HOG) descriptors (Section 5.1). Then, we apply Fisher vectors on the extracted descriptors to represent videos (Section 5.2). Finally, we apply Support Vector Machines for action classification (Section 5.3).

5.1. VBC and HOG Features Extraction

We use dense trajectories [17] to extract local spatio-temporal patches [1]. By extracting dense trajectories, we provide a good coverage of a video and we ensure extraction of meaningful features. We limit the length of trajectories to $t = 15$ frames. Short trajectories are more robust than long trajectories, in particular in the presence of fast irregular motions and when the trajectories are drifting. Moreover, short trajectories are necessary for the recognition of short actions like smiling or doing a hive five.

Similarly to [17], we extract a space-time volume (*i.e.* a patch) of size $S \times S$ pixels and t frames around each trajectory. The volume is subdivided into 3 temporal cells of $l = 5$ frames. For each cell we compute a descriptor, and we concatenate the descriptor of each cell to create a final trajectory descriptor. As a descriptor we use our Video Brownian Covariance (VBC) descriptor and the HOG descriptor. Therefore, each trajectory is represented by two appearance descriptors.

[1]The dense trajectories were selected based on their use in the recent literature. However, our approach can be used together with any other algorithm extracting local spatio-temporal patches.

5.2. 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 $\mathbb{X}$, based on Fisher vectors [12, 13]. We model features with a generative model and compute the gradient of their likelihood with respect to the parameters of the model, *i.e.* $\Delta_\lambda \log p(\mathbb{X}|\lambda)$. We describe how the set of features deviates from an average distribution of features, modeled by a parametric generative model. Firstly, during the preliminary learning stage, we fit a M-centroid Gaussian Mixture Model (GMM) to our training features, which can be regarded as a soft visual vocabulary:

$$p(x_i|\lambda) = \sum_{j=1}^{M} w_j g(x_i|\mu_j, \Sigma_j), \tag{12}$$

$$\mathbf{s.t.} \quad \forall_j : w_j \geq 0, \quad \sum_{j=1}^{M} w_j = 1, \tag{13}$$

$$g(x_i|\mu_j, \Sigma_j) = \frac{1}{(2\pi)^{D/2}|\Sigma_j|^{1/2}} e^{-\frac{1}{2}(x_i-\mu_j)^T \Sigma_j^{-1}(x_i-\mu_j)}, \tag{14}$$

where $x_i \in \mathbb{X}$ is a D-dimensional feature vector, $\{g(x_i|\mu_j, \Sigma_j)\}_{j=1}^{M}$ are the component Gaussian densities and $\lambda = \{w_j, \mu_j, \Sigma_j\}_{j=1}^{M}$ are the parameters of the model, respectively the mixture weights $w_j \in \mathbb{R}_+$, the mean vector $\mu_j \in \mathbb{R}^D$ and the positive definite covariance matrices $\Sigma_j \in \mathbb{R}^{D\times D}$ of each Gaussian component. We learn the parameters λ using the Expectation Maximization restricting the covariance of the distribution to be diagonal. To estimate the GMM parameters, we randomly sample a subset of $100,000$ features from the training set and we set the number of Gaussians to $M = 128$. To increase the precision, we initialize GMM ten times and we keep the codebook with the lowest error. We define the soft assignment of descriptor x_i to the Gaussian j as a posteriori probability $\gamma(j|x_i, \lambda)$ for component j:

$$\gamma(j|x_i, \lambda) = \frac{w_j g(x_i|\mu_j, \Sigma_j)}{\sum_{l=1}^{M} w_l g(x_i|\mu_l, \Sigma_l)}, \tag{15}$$

Then, we compute the gradients of the j-th component with respect to μ and σ, using the following derivations:

$$G_{\mu,j}^{\mathbb{X}} = \frac{1}{N_x\sqrt{w_j}} \sum_{l=1}^{N_x} \gamma(j|x_l, \lambda) \left(\frac{x_l - \mu_j}{\sigma_j}\right),$$

$$G_{\sigma,j}^{\mathbb{X}} = \frac{1}{N_x\sqrt{2w_j}} \sum_{l=1}^{N_x} \gamma(j|x_l, \lambda) \left(\frac{(x_l - \mu_j)^2}{\sigma_j^2} - 1\right), \tag{16}$$

where N_x is the cardinality of the set $\mathbb{X}$. Finally, we encode a set of local descriptors $\mathbb{X}$ as a concatenation of partial derivatives with respect to the mean $G^{\mathbb{X}}_{\mu,j}$ and standard deviation $G^{\mathbb{X}}_{\sigma,j}$ parameters for all M components:

$$V = [G^{\mathbb{X}}_{\mu,1}, G^{\mathbb{X}}_{\sigma,1}, ..., G^{\mathbb{X}}_{\mu,M}, G^{\mathbb{X}}_{\sigma,M}]^T. \quad (17)$$

The dimension of the Fisher vector representation is $2DM$.

5.3. Action Recognition

Linear classifier has shown to be efficient and has shown to provide good results with high dimensional video representations like Fisher vectors. Therefore, we use linear Support Vector Machines for action classification. Given a set of n instance-label pairs $(\mathbf{x}_i, y_i)_{i=1..n}$, $\mathbf{x}_i \in \mathbb{R}^k$, $y_i \in \{-1, +1\}$, we solve the following unconstrained optimization problem:

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n} \xi(\mathbf{w}; \mathbf{x}_i, y_i), \quad (18)$$

where C is a penalty parameter ($C > 0$) and $\xi(\mathbf{w}; \mathbf{x}_i, y_i)$ is a loss function $\max(1-y_i\mathbf{w}^T\mathbf{x}_i, 0)$, referred to as L1-SVM. We set the parameter C to $C = 200$, which has shown good results on a subset of training samples across various datasets. For multi-class classification, we implement the one-vs-all strategy.

6. Experiments

In this section, we present the evaluation of our approach using three state-of-the-art challenging datasets: URADL, MSRDailyActivity3D, and HMDB51.

We study the performance of appearance descriptors, *i.e*. VBC and HOG, separately and together using late fusion technique. Moreover, we study the performance of these descriptors using Principal Component Analysis (PCA). The HOG descriptor is the most popular appearance descriptor for action recognition, and therefore, it provides a good baseline for comparison.

Many authors [8, 9, 17] combine appearance descriptors with motion-based descriptors (*e.g*. HOF, MBH). However, in this paper we focus only on the appearance features and not motion features.

6.1. URADL Dataset

The URADL (University of Rochester Activities of Daily Living) dataset [10] [2] contains 10 types of human activities of daily living, selected to be useful for an assisted cognition task. The full list of activities is: answering a phone, dialing a phone, looking up a phone number in a telephone directory, writing a phone number on a whiteboard, drinking a glass of water, eating snack chips, peeling a banana, eating a banana, chopping a banana, and eating food with silverware. Each action is performed three times by five different people. In total, the dataset contains 150 video sequences recorded with 30 fps frame rate and 1280×720 pixels spatial resolution. The videos were downsampled to the 640×360 pixels spatial resolution. The dataset contains a set of challenges like: different shapes, sizes, genders and ethnicities of people, and difficulty to separate activities on the basis of a single source of information (e.g. eating a banana vs. eating snack chips, and answering a phone vs. dialing a phone). We use leave-one-person-out cross-validation evaluation scheme to report the performance of our approach on this dataset.

6.2. MSRDailyActivity3D Dataset

The MSRDailyActivity3D dataset [21] [3] consists of 16 actions such as: drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum cleaner, cheer up, sit still, toss paper, play game, lie down on sofa, walk, play guitar, stand up, sit down. 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. In total, the dataset contains 320 videos recorded with 640×360 pixels spatial resolution. The videos were downsampled to the 320×180 pixels spatial resolution. We use leave-one-person-out cross-validation evaluation scheme to report the performance of our approach on this dataset.

6.3. HMDB51 Dataset

The HMDB51 dataset [6] [4] contains 6766 video sequences divided into 51 action categories, each containing a minimum of 101 video clips. All action categories can be divided into 5 groups: (1) general facial actions like smile, laugh, chew, talk, (2) facial actions with object manipulation like eat and drink, (3) general body movements like clap hands and dive, (4) body movements with object interaction like brush hair and ride bike, (5) body movements for human interaction like hug and shake hands. This dataset contains multi person actions and is collected from movies and public datasets such as Prelinger archive and YouTube. It is very challenging due to significant camera/background motion, huge appearance variations of people and actions, not stabilized videos, occlusions, amount of video data and changes in scale, rotation and viewpoint. We use three splits provided by the authors, to report the performance of our approach on this dataset.

[2] http://www.cs.rochester.edu/~rmessing/uradl/

[3] http://research.microsoft.com/en-us/um/people/zliu/ActionRecoRsrc/

[4] http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/

Table 1. Results from our experiments presenting the performance of HOG and VBC descriptors individually and together. The VBC descriptor is complementary to the HOG descriptor and improves the action recognition performance.

Table 2. Results without PCA.

	HOG	VBC	**Fusion**
URADL	69.33%	70%	**80.00%**
MSR	42.19%	31.56%	**45.31%**
HMDB51	16.71%	11.31%	**18.95%**

Table 3. Results with PCA.

	HOG	VBC	**Fusion**
URADL	80.67%	74%	**80.67%**
MSR	47.81%	47.81%	**54.38%**
HMDB51	26.21%	20.02%	**31.07%**

6.4. Results

The results are available in Table 2 and Table 3. We observe that the VBC descriptor carries complementary information to the HOG descriptor, as fusion of both descriptors consistently outperforms each of them. The HOG descriptor directly models values of given features, whereas the VBC descriptor focuses on their pairwise relations.

The results also show that both descriptors can benefit from PCA. The descriptors with PCA always achieve better results than without it.

7. Conclusions

We presented a novel, appearance-based descriptor for action recognition, which carries complementary information to the HOG descriptor. In contrast to the HOG (which directly models values of given features), the VBC descriptor focuses on features pairwise relations. The fusion of both descriptors gives a significant advantage in performance. In addition, we showed that further improvement can be made by applying PCA. Our novel descriptor can be applied in complex action recognition methods, which combine appearance-based and motion-based descriptors.

In further work we intend to examine the VBC descriptor using motion features.

Acknowledgment

This work is supported by the Région Provence-Alpes-Côte d'Azur and by the Dem@Care, SafEE, and PANORAMA projects. However, the views and opinions expressed herein do not necessarily reflect those of the financing institution.

References

[1] S. Bak, R. Kumar, and F. Bremond. Brownian descriptor: a Rich Meta-Feature for Appearance Matching. In *WACV*, 2014.

[2] P. Bilinski and F. Bremond. Evaluation of local descriptors for action recognition in videos. In *ICVS*, 2011.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

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

[5] K. Guo, P. Ishwar, and J. Konrad. Action recognition from video using feature covariance matrices. *IEEE Transactions on Image Processing*, 22(6):2479–2494, 2013.

[6] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *ICCV*, 2011.

[7] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.

[8] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[9] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *CVPR*, 2009.

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

[11] Y. Pang, Y. Yuan, and X. Li. Gabor-based region covariance matrices for face recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(7):989–993, 2008.

[12] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-Scale Image Retrieval with Compressed Fisher Vectors. In *CVPR*, 2010.

[13] F. Perronnin, J. Sanchez, and T. Mensink. Improving the Fisher Kernel for Large-Scale image classification. In *ECCV*, 2010.

[14] F. Porikli, O. Tuzel, and P. Meer. Covariance tracking using model update based means on riemannian manifolds. In *CVPR*, 2006.

[15] G. J. Szekely and M. L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265, 2009.

[16] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *ECCV*, 2006.

[17] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[18] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[19] X. Wang, G. Doretto, T. B. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In *ICCV*, 2007.

[20] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *ECCV*, 2008.

[21] Y. Wu. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.

[22] C. Yuan, W. Hu, X. Li, S. J. Maybank, and G. Luo. Human action recognition under log-euclidean riemannian metric. In *ACCV*, 2009.