




# Frame-Level Covariance Descriptor for Action Recognition

Wilson Moreno<sup>(✉)</sup> , Gustavo Garzón, and Fabio Martínez

Grupo de Investigación en Ingeniería Biomédica (GIIB), Motion Analysis and Computer Vision (MACV), Universidad Industrial de Santander (UIS), Bucaramanga, Colombia

widamo456@gmail.com, gustavo.garzon@saber.uis.edu.co, famarcar@uis.edu.co

**Abstract.** Activity recognition is a fundamental task in areas such as video-surveillance, gesture recognition, robotics, multimedia applications among much others. Such task remains as an open problem because the variability of many factors such as the appearance of actors, illumination changes in real scenarios and the dynamic developed for each action. Despite favorable results in recent works for several academic datasets, the proposed methodologies require a huge number of training samples and the output descriptor result in a high dimensional array that difficult the implementation in real conditions. This work proposes a spatio-temporal descriptor that model human activities by using a fast regional covariance representation for each frame. At each frame, a set of motion and geometrical map measures are quantified into a pyramidal regional structure to describe the instantaneous action. Such low-level primitive maps are codified into a integral covariance that allows a fast and compact description of local correlation among features. The set of pyramidal-frame-covariances along the video sequence represent a manifold that coexist in a positive Riemannian space. Then, a set of means are approximated in Riemannian space for each regional covariance sequence to represent a very compact action descriptor. The proposed action descriptor is mapped to a Euclidean space to perform an automatic classification using a Support vector Machine. The proposed approach was evaluated in two different public datasets: (1) in UT-Interaction with a k-fold cross-validation scheme was achieved a 70.8% of accuracy with a descriptor size of just 10 features per video sequence and (2) in UCF Sports achieve an accuracy of 71.7% using 13 features.

**Keywords:** Spatio-temporal covariance · Human activity recognition  
Motion analysis · Low-level primitives

## 1 Introduction

Action recognition is a fundamental area in computer vision with widely applications such as surveillance applications, sport analysis, smart vehicles, HCI

systems among others [14]. However, the proper characterization of activities implies several challenges that include the modeling of complex variability of illumination, object representation, motion changes, among much others. Likewise, traditional approaches involve a high computational cost due to exhaustive quantification of features that lead to increased descriptor dimensionality.

In the state of the art have been proposed multiple strategies to recognize actions that coarsely can be classified as global and local recognition methods. Global representation methods have focus on characterization and quantification of extensive regions of interest or even complete video sequences. Seminal strategies proposed subtraction-based methods and human silhouette tracking methods for addressing pedestrian detection applications [6]. Moreover, Wang *et al.* [21] proposed a descriptor for activity recognition based on the extraction of binary human silhouettes using the  $R$ -Transform to represent low-level features. This strategy is robust to occluded frames, disjoint silhouettes and holes over shapes. Souvenir and Babbs [19] further extended this work by considering image contours, which improved activity characterization at the cost of decreasing computational efficiency. Additional strategies have proposed silhouettes characterization from multiple cameras, but requiring exhaustive calibration for the acquisition devices [4]. These methods quantify postural movement based on frame-level silhouettes but are sensitive to noise, partial occlusion, view point variability and dependent of proper capture of the silhouettes [14].

On the other hand, a wide variety of methods based on interest point detection and local patches have been proposed that relatively avoid invariance to appearance, perspective, and are robust to partial occlusions [14]. For instance, Laptev and Lindeberg [10] capture multiple interest points at different scales into the spatio-temporal domain that allows local structure detection for event representation in video sequences. Laptev in [9] uses local geometry characterization at multiple scales to compute salient cuboids in video-sequences. The salient points are mapped to a Support Vector Machine (SVM) to automatically classify actions. Gowayed *et al.* [7] proposed a method for action recognition using the position of joints with respect to a human skeleton. This method describe 3d human joints trajectories based on Histograms of oriented displacements (HOD). Robertson and Reid [15] proposed to combine trajectory features (i.e. position and speed) as a set of local motion descriptors for human action recognition. Liu *et al.* [11] proposed a method based on regularized multi-task learning that implicitly codify local visual characteristics and human body structure as small information blocks, which are represented as a pyramid-shaped bag of words (PPBoW). This approach achieve competitive results because the robustness to appearance and geometry but remains dependent to appearance. Nevertheless, these approaches evidenced a high computational cost which limit the development of online applications. Also, in these approaches the accuracy namely requires high-dimensional descriptors to achieve a correct action prediction.

The present investigation develops a spatio-temporal covariance descriptor that model and characterize human activities occurring in video sequences. For so doing, the proposed approach compute a dense optical flow representation

along the video sequence, allowing to quantify large displacements. Then, a set of kinematic primitives are computed at each frame to highlight main motion that represent the present action. The kinematic primitives at each frame are coded in regional covariance matrices that are computed from a coarse to fine representation by iteratively splitting each frame. The set of computed covariance at each frame represent a manifold in Riemannian space that is the signature of each particular action codified in the video. Then, the final descriptor is computed as the Riemannian mean of covariance matrices that represent the action along the video. The rest of the paper is organized as follows: the proposed approach is described in section two, while the evaluation and result of the proposed approach with respect to the state of the art is reported in Sect. 3. Finally in Sect. 4 is discuss the advantages and limitations of the approach and also are presented several conclusion of the work.

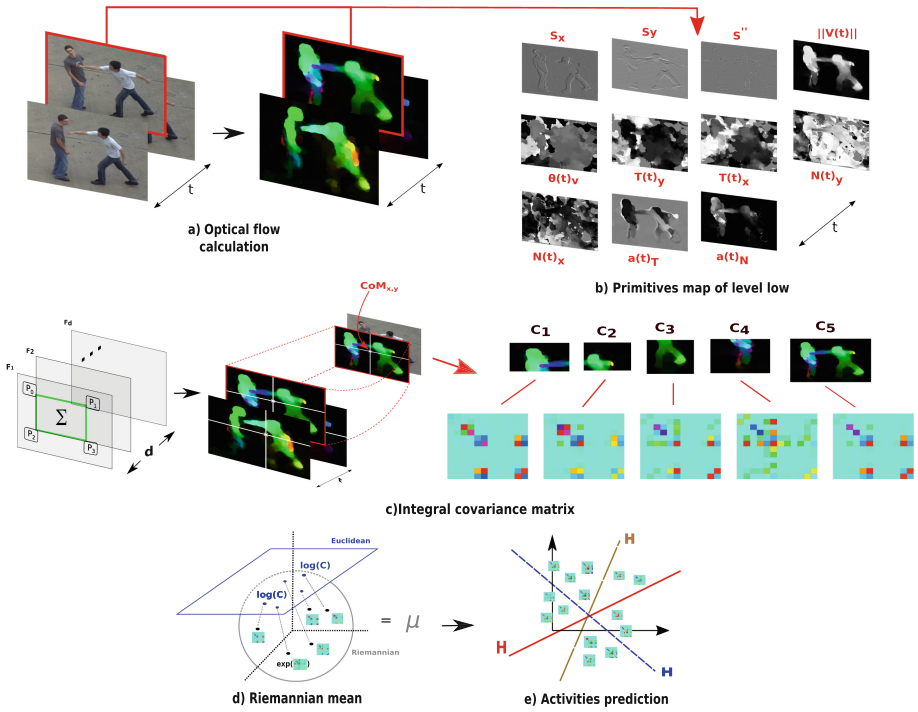
## 2 Proposed Method

In this work is introduced a compact covariance descriptor that coded spatio-temporal features mainly computed from a dense optical flow to represent activities in videos. The proposed approach starts by computing a dense velocity field that allows to code large displacements along the video. Then a set of kinematic primitives are calculated frame-wise along the sequence. Hereafter an efficient integral covariance is herein implemented to represent multiple frame regions and coded the different kinematic primitives. A special mean approximation is computed over the set of frame-covariances that form a manifold video representation in Riemannian space. Finally, the proposed descriptor is mapped to a Euclidean space to be tested on a classification algorithm to obtain an automatic activity classification. The pipeline of the proposed approach is depicted in Fig. 1.

### 2.1 Kinematic Primitive Maps

The herein proposed approach quantify apparent motion map primitives as low-level representation of each video. In this work, was firstly computed a dense optical flow to code a set of velocity field maps for each frame. Also, additional local motion patterns were computed from the optical flow representation. The proposed approach has an intrinsic advantage to admit any local measure represented as a feature map. The kinematic feature maps are described as follows:

This approach is flexible to admit any dense flow representation that codifies the apparent velocity at each frame. In this work was implemented a robust variational strategy that allows to describe large displacements along the sequence by considering several local and regional restrictions [1]. The computation of large displacements allows to properly describe many salient patterns of typical actions that implies large body displacements in short periods of time. First, the optical flow approach consider classical restriction of color  $E_{color}(w)$  and gradient  $E_{gradient}(w)$  between consecutive frames. As classical dense approximations,



**Fig. 1.** Pipeline of the proposed method: In (a) is calculated large displacement optical flow. In (b) is calculated low-level motion and appearance primitives. In (c) is calculated regional covariances for each frame. This regional descriptor is constructed by five covariance matrices. Regions are calculated using center of mass position w.r.t the optical flow. In (d) a covariance descriptor is calculated for each sequence by estimating the average of covariance matrices. (e) Finally, the proposed covariance descriptor was validated with academic datasets.

such optical flow consider that the color and shape representation of the object is the same in very short intervals of time. Also, this dense flow use additional restrictions such as:

- **Smooth:** This restriction  $E_{smooth}(w)$  quantify the minimal difference between velocity vectors inside a region. The assumption is that velocity patterns must be similar in a certain neighborhood, taking into account local dispersion of vector field.
- **Non-local regions:** This restriction  $E_{desc}(w_1)$  allows to search for large local displacements between consecutive frames by comparing matching-regions calculated from feature vectors. An non-local matching of some interest points between frames is achieved by a SIFT points representation.

Finally, the sum of all energy equations allows to estimate the best dense optical field for all video sequences. Therefore a complete model is defined as a unique optimization problem, by minimizing:  $E(w) = E_{color}(w) + \gamma E_{gradient}(w) +$

$\alpha E_{smooth}(w) + \beta E_{Match}(w, w_1) + E_{desc}(w_1)$ , where  $\{\gamma, \alpha, \beta\}$  which represent regularization constants with values between  $[0, 1]$ . The herein implemented flow can handle object deformations, motion discontinuities, occlusion and arbitrarily large displacements along the video sequences.

From dense motion field are first recovered the speed  $\|V(t)\|$  and angle  $\theta_V(t)$  maps. These kinematics represent first order primitives, from which are computed other spatio-temporal and higher order representations. For instance, the derivative of speed magnitude represents motion rapidness which corresponds to an scalar value that relates tracked distance and time  $S_{\|V(t)\|}$  and the derivative of speed angle represents direction of variation in speed  $S_{\theta_v(t)}$  for each frame.

Unit tangent speed  $T(t)$  is also calculated as  $T(t) = \frac{V(t)}{\|V(t)\|}$  and the unit normal speed  $N(t)$  is computed as  $N(t) = \frac{T'(t)}{\|T'(t)\|}$ . For each pixel inside optical flow frames there are two orthogonal vectors  $N(t)$  and  $T(t)$  for each  $t$  and they expand over an osculating plane  $\rho(t)$  which contain unit tangent vectors. Furthermore, acceleration corresponds to the derivative of motion speed with respect to time. This can be expressed in terms of unit tangent speed and unit normal speed. Osculating plane contains this primitive only if  $T(t)$  and  $N(t)$  exist. Acceleration is given by  $a(t) = a_T(t)T(t) + a_N N(t)$ , where tangential  $a_T$  and normal  $a_N(t)$  acceleration coefficients are given by:  $a_T(t) = \frac{d}{dt} \|V(t)\|$ ,  $a_N(t) = \|V(t)\| \left\| T'(t) \right\|$ .

Tangential acceleration represents the derivative of motion speed whereas normal acceleration represents the derivative of speed direction with respect to time. Both quantities allow us to obtain the magnitude of acceleration as:  $\|a\|^2 = (a_T)^2 + (a_N)^2$ . To enrich motion representation, the velocity neighborhood relationship was captured by computing the first derivative of motion field but with respect to the  $(x, y)$  axis as:  $\frac{\partial \|V(t)\|}{\partial x \partial y}$ . Additionally, the kinematic motion representation maps can be complemented by using shape and appearance maps at each frame. In this paper for some experiments was added first and second order derivatives over each frame. Such maps represent the borders i.e., local geometry captured at each frame.

## 2.2 Integral Covariance Coding

A compact correlation of independent kinematic maps is carried out by computing the covariance that represent the particular activities. Covariance matrix constitutes a natural and compact method for combining multiple correlated features, that can be expressed as:

$$C_R(i, j) = \frac{1}{n-1} \left[ \underbrace{\sum_{k=1}^n z_k(i)z_k(j)}_Q - \frac{1}{n} \underbrace{\sum_{k=1}^n z_k(i)}_P \underbrace{\sum_{k=1}^n z_k(j)}_P \right], \quad (1)$$

where  $z_{k=1\dots n}$  is a vector that coded the kinematic maps with  $n$  samples for  $(i, j) = 1 \dots d$  features and  $\frac{1}{n} \sum_{k=1}^n z_k(i)$  represents the  $\mu$  expected value.

Covariance matrices are  $d \times d$  symmetric and positives, representing a phenomenon by  $\frac{d^2+d}{2}$  different values, with the main diagonal as the feature variances. This matrix has been widely exploited in different applications of object identification, tracking, and classification [12]. However, the multiple computations with  $n$  samples require a high computational cost, because the interactions between each pair of  $d$  features. To cope such limitation, a fast regional alternative proposed in [20] was herein implemented to compute covariance regions by using integral image representation (as illustrated in Fig. 1(c)). These integral images are intermediate representations and are typically used for fast calculation of sums over a given region.

In this work each frame  $F$  is characterized with  $d$  kinematic features and then regionally coded by using a integral covariance representation. In such case, the sum of each characteristic dimension  $z(i)_{k=1\dots n}$ , is represented as a first-order tensor  $P \in \mathbb{R}^{W \times H \times d}$ , computed as  $P(x', y', i) = \sum_{x < x', y < y'} F(x, y, i)$ , where  $F$  is a frame  $F \in \mathbb{R}^{(W \times H \times d)}$  with  $i = 1 \dots d$ . Then, tensor  $P$  is a  $d$ -sized vector containing the sum of each kinematic map independently, with dimension,  $P_{x,y} = [P(x, y, 1) \dots P(x, y, d)]^T$ .

Also, the sum of the product of features  $z_k(i)z_k(j)_{i,j=1\dots n}$  (first part of Eq. 1) can be expressed as integral images as a second-order tensor  $Q \in \mathbb{R}^{W \times H \times d \times d}$  as:  $Q(x', y', i, j) = \sum_{x < x', y < y'} F(x, y, i)F(x, y, j)$  with  $\{i, j\} = 1 \dots d$ . Tensor  $Q$  is a  $d * d$  symmetric matrix that contains the sum of the products of any pair of features, expressed as:

$$Q_{x,y} = \begin{pmatrix} Q(x, y, 1, 1) & \dots & Q(x, y, 1, d) \\ \vdots & \ddots & \vdots \\ Q(x, y, d, 1) & \dots & Q(x, y, d, d) \end{pmatrix} \quad (2)$$

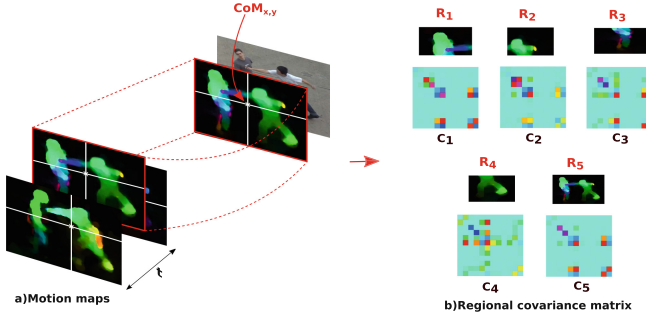
The computation of this integral tensor require  $\frac{d^2+d}{2}$  iterations. Then, any rectangular region  $R$ , bounded by upper-left and lower-right corners can be computed with a computational cost of  $O(d^2)$ .

The Eq. 1 can be re-written in terms of integral tensors as:

$$C_{R(x',y';x'',y'')} = \frac{1}{n-1} [(Q_{x'',y''} + Q_{x',y'} - Q_{x'',y'} - Q_{x',y''}) - \frac{1}{n}(P_{x'',y''} + P_{x',y'} - P_{x'',y'} - P_{x',y''}) (P_{x'',y''} + P_{x',y'} - P_{x'',y'} - P_{x',y''})^T], \quad (3)$$

where  $n = (x'' - x')(y'' - y')$ . Such expression implies faster computations for any regional covariances in the entire frame with few arithmetic operations. The computational advantages of integral covariances allow us to enrich the description of each frame by computing the covariance matrix at different spatial regions. A total of five covariance matrices were obtained to represent each frame  $F$  in the video sequence (see in Fig. 2). The first covariance correspond to whole frame. The remaining four correspond to frame sub-regions split with respect to the position of center of mass (CoM) and given by the appearance motion field. This CoM is computed as  $CoM_{x,y} = \frac{1}{M} \sum_{y=1}^n \sum_{x=1}^n \|V_{y,x}(t)\| r_{y,x}(t)$ , where  $\|V(t)\|$  represents the speed,  $\{x, y\}$  is the frame positions and  $M =$

$\sum_{y=1}^n \sum_{x=1}^n \|V_{y,x}(t)\|$ . The CoM can be interpreted as the spatial position with the biggest amount of motion for any given frame. The four sub-regions are calculated by dividing the frame with respect to the CoM. In Fig. 2 is represented this computation.



**Fig. 2.** Covariance descriptor: (a) Selection of activities in video sequences and calculating large displacement optical flow and position of center of mass. (c) Calculating covariance matrices for each region.

### 2.3 Riemannian Mean Video Sequence

The set of covariance matrices computed at each frame  $c_1, c_2, c_3, \dots, c_n$  form a manifold that represent the action video. Because covariance properties, such manifold is defined in a spherical Riemannian space and not in the classical Euclidean space [13], limiting the use of classic machine learning and vision algorithms. The projection of a covariance matrix  $c_i$  into a Euclidean space is approximated by computing  $\log(p) = \Sigma DIAG(\log(\lambda_i))\Sigma^T$ , where  $\Sigma$  are the eigenvectors of the matrix and  $\lambda$  are the respective eigenvalues. In the same way, any projection from euclidean space to Riemannian space is approximated as  $\exp(p) = \Sigma DIAG(\exp(\lambda_i))\Sigma^T$ .

Hence, a video descriptor is herein proposed as a representative covariance matrix that has minimal distance with respect to the set  $c_1, c_2, c_3, \dots, c_n$ , computed for each region along time. This representative covariance is then computed as the intrinsic mean covariance in Riemannian space, as shown in Algorithm 1 [5].

To compute intrinsic mean, the log and exp operations are defined with respect to the computed  $\mu$ , that is expressed as:  $\exp_\mu(X) = \mu^{\frac{1}{2}} \exp(\mu^{-\frac{1}{2}} X \mu^{-\frac{1}{2}}) \mu^{\frac{1}{2}}$  and  $\log_\mu(p) = \mu^{\frac{1}{2}} \log(\mu^{-\frac{1}{2}} p \mu^{-\frac{1}{2}}) \mu^{\frac{1}{2}}$ , respectively. It holds that  $\exp(\log(\mu)) = \mu$  and the inverse matrix  $\mu^{\frac{1}{2}} = \exp(\frac{1}{2}(\log \mu))$ . The stop criteria  $\|X_i\|$ , is measured in the iterative matrix result as  $\|X_i\| = \sum_{i=1}^N (\log(\sigma_i))^2$ , where  $\sigma$  are the respective eigenvalues. An error threshold must be defined as:  $(0.01 < \epsilon < 0.1)$ . A final video descriptor is then formed by the concatenation of the set of regional covariance means as  $V_d = \{\mu_{cR1}, \mu_{cR2}, \dots, \mu_{cRn}\}$ .

---

**Algorithm 1.** Gradient descent algorithm to compute the intrinsic mean computation from a set of regional frame-covariances

---

**Output:**  $\mu \in C(n)$

```

1: for Each regional covariance sequences  $j$  do
2:    $c_1^j, \dots, c_N^j \in C(n)$ 
3:    $\mu = c_1^j$ 
4:    $\tau = 1 \rightarrow$  initial step size
5:   Do
6:      $X_i = \frac{1}{N} \sum_{k=1}^N \log_{\mu_i}(c_k^j)$ 
7:      $\mu_{i+1} = \exp_{\mu_i}(\tau X_i)$ 
8:
9:     if ( $\|X_i\| > \|X_{i-1}\|$ ) then
10:        $\tau = \tau/2$ 
11:        $X_i = X_{i-1}$ 
12:     end if
13:
14:   While ( $\|X_i\| > \epsilon$ )
15: end for

```

---

### 2.4 Support Vector Machine Training

Finally, action classification of covariance descriptor was achieved by using a Support Vector Machine (SVM). The SVM strategy has been widely used for action recognition problems because the proper trade-off between accuracy and computational cost. Since, the action descriptor  $V_d$  herein proposed is formed by a set of mean covariance matrices, it is necessary to project to euclidean space as  $\log(V_d) = \{\log(\mu_{c_{R1}}), \log(\mu_{c_{R2}}), \dots, \log(\mu_{c_{Rn}})\}$  by using a spectral decomposition  $\log(\mu_{c_{Ri}}) = \Sigma \log(\lambda) \Sigma^T$ , as explained in Sect. 2.3.

Once the covariance descriptor is projected to euclidean space a *One against one SVM multiclass classification* was herein implemented, with a Radial Basis Function (RBF) kernel [3]. Here, the classes represent the actions and optimal hyperplanes separated by a classical max-margin formulation. For  $k$  motion classes, a majority voting strategy is applied on the outputs of the  $\frac{k(k-1)}{2}$  binary classifiers. A  $(\gamma, C)$ -parameter sensitivity analysis was performed with a grid-search using a cross-validation scheme and selecting the parameters with the largest number of true positives.

### 2.5 Data

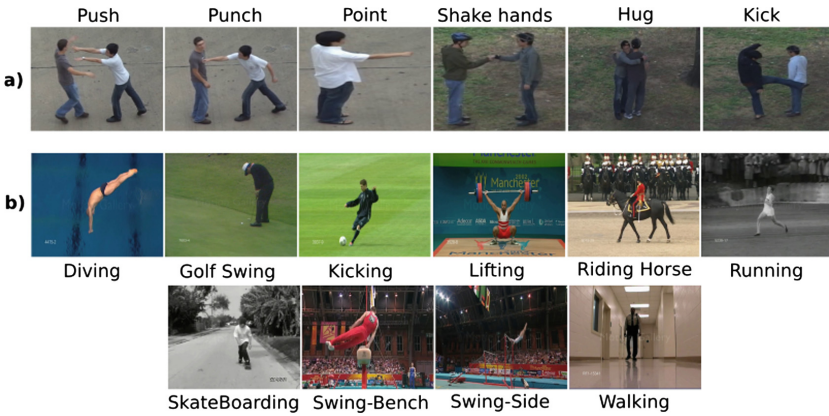
To evaluate the performance of the proposed strategy, two different public datasets were considered, described as follows:

- **UT-Interaction** (High-level Human Interaction Recognition Challenge) exhibits complex human activities in real-world scenarios [16]. This dataset contains 6 human interaction classes: *shake-hands (sh)*, *point (po)*, *hug (hg)*,



*push (ps)*, *kick (ki)* and *punch (pn)*, with a spatial resolution of  $720 \times 480$  with 30 fps. The dataset is split in two groups of 60 videos (see Fig. 3). The first group was captured in a relative static background while the second one report some jitter of the camera, and also there are some human motions in background. Following criteria of evaluation proposed by authors of the dataset, we used a cross-validation strategy using a  $K$ -fold scheme.

- **UCF Sports** (University of Central Florida sports dataset) consists of a set 150 sequences of different sports [18], such as: *Diving (dv)*, *Golf Swing (gs)*, *kicking (ki)*, *Lifting (lf)*, *Riding Horse (rd)*, *Running (ru)*, *SkateBoarding (sk)*, *Swing-Bench (sw)*, *Swing-Side (ss)* and *Walking (wl)*. The sequences were recorded with a spatial resolution of  $720 \times 480$  and namely 10 fps. The dataset represents a natural pool of actions featured in a wide range of scenes and viewpoints (see Fig. 3). Some of sequences are formed by several split sequences which difficult to track the coherence in developed actions. For evaluation purposes, a multi-class classifier of the proposed descriptor was run under a leave-one-out scheme.

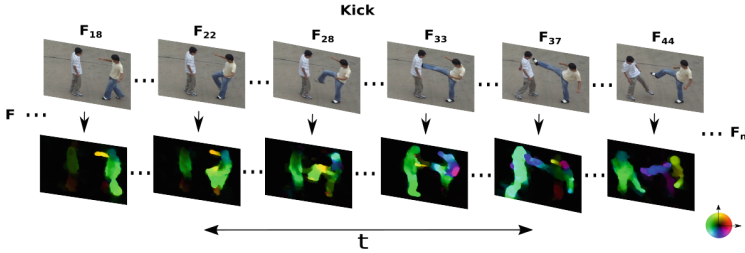


**Fig. 3.** Example of video sequences in both datasets: (a) UT-Interaction dataset and, (b) UCF Sport dataset. The Figure illustrates the shape and dynamic variability for both datasets as well as the non-controlled scenarios that difficult the action characterization

### 3 Evaluation and Results

A first evaluation of the proposed approach was carried in UT-Interaction dataset by computing several kinematic primitives at each frame of the sequences. The kinematic primitives considered were  $V(t)$ ,  $\|V(t)\|$ ,  $\theta_V(t)$ ,  $T(t)$ ,  $N(t)$ ,  $a_T(t)$  and  $a_N(t)$ . Such kinematic maps codified five integral covariances per each frame. Then, five Riemannian means were estimated for each regional covariance region for the entire sequence, resulting in a total descriptor of 275 values per video

sequence. Table 1 shows the confusion matrices obtained for the dataset with relative static background (left) and for data with background motions and camera jitters (right). In Fig. 4 is illustrated the optical flow herein computed in a typical video sequence of UT-Interaction. In most of the cases the optical flow achieves a proper description of local gesture activities along the video-sequence, capturing large displacement in actions such as punching or kicking.



**Fig. 4.** A large displacement optical flow computed over a UT-Interaction sequence. A color map represent the computed field on each frame. Colors represent displacements while the intensity represent the norm of each vector.

**Table 1.** Confusion matrix for the UT-Interaction dataset by coding several kinematic primitives. The results are expressed in percentage (%). In the left is shown the performance of the proposed approach for relative static background videos, while in the right is presented the performance with videos that present jitter camera motions and other activities in background.

Category	sh	hg	ki	po	pn	ps
sh	90	10	0	0	0	0
hg	0	100	0	0	0	0
ki	0	0	80	0	10	10
po	0	0	0	100	0	0
pn	20	0	20	0	40	20
ps	0	0	20	0	10	70

Category	sh	hg	ki	po	pn	ps
sh	60	30	10	0	0	0
hg	20	70	0	0	0	10
ki	0	0	60	0	40	0
po	0	0	0	100	0	0
pn	0	0	30	0	40	30
ps	10	10	10	0	30	40

In average, the proposed approach achieve an average accuracy of 80.0% and 61.66% for both UT-Interaction datasets. Because the flexibility of the proposed approach to code any feature maps computed over the frame, a set of appearance was included in a new experiment. In such case, simple gradient features of first  $S_{\|I(t)\|}$  and second order  $S''_{\|I(t)\|}$  were integrated in the proposed strategy. On average, we obtained an accuracy of 75.0% and 55.0% for UT-Interaction in categories 1 and 2, respectively. The addition of appearance to our proposed descriptor shows no improvement on classification and recognition of human activities because shape among actions in this dataset tend to be similar. For

instance, the actions shake-hands and pointing have similar shape information along the action developing. Also hugging and pushing along the sequences share shape similarities that can lead to misclassification. Also, the interaction and background variability on video sequences can limit the proper quantification of actions geometry.

Table 2 report a comparison of the proposed motion descriptor with other state of the art strategies. Some of these approaches achieve high accuracy rates but demand a complete processing of the video to compute the features. For instance, the propagative voting approach [22] reports a computational complexity of  $O(N_M) + O(WHT)$ , where  $N_M$  is the number of matches and  $W$ ,  $H$ ,  $T$  is the spatial (width  $\times$  height) and temporal video resolutions. Such number of matches is computed by using random projection trees, a precise strategy that results computationally expensive and prohibitive for online applications.

**Table 2.** Average accuracy for different reported state of the art strategies. Although the propagation voting achieves better results in terms of accuracy, the match of features using random projection trees is computationally expensive. The Xiaofei *et al.* work integrates BoW occurrence histogram with HoG, representing again a high computational time to obtain an action representation. In contrast, the proposed approach produces a compact descriptor that takes into account different time interval depths by using the same source of primitives, i.e., a dense optical flow.

Approaches	UT-interaction set 1	UT-interaction set 2
Propagative voting [22]	93.3	91.7
Daysy [2]	71.67	56.67
Laptev [9] + SVM	68	65
Slimani [17]	40	66
Xiaofei [8]	83	-
Proposed approach	80	61.66

In summary, the proposed approach achieves a relevant dynamic characterization of the different human interaction activities but only considering kinematic information. The activities recorded in UT-Interaction are often the result of combinations of complex motion patterns that may occur during a short time interval. In such cases some misclassifications can be obtained in several sample activities that share several local motion patterns along the sequence. For instance, interactions like *hand shaking*, *pointing* or *pushing*, share similar limb movements during certain temporal intervals.

This strategy was also evaluated in UCF Sports achieving an average accuracy of 61.5% using kinematic primitives, such as:  $V(t)$ ,  $\|V(t)\|$ ,  $\theta_V(t)$ ,  $T(t)$ ,  $N(t)$ ,  $a_T(t)$  and  $a_N(t)$ . For this specific dataset with the integration of gradient features ( $S_{\|I(t)\|}$ ,  $S'_{\|I(t)\|}$ ) the proposed approach achieve an average accuracy of 71.7%. An overall increase of 10% was obtained because characteristic human postures in certain sports can help to distinguish activities.

**Table 3.** Confusion matrix for UCF Sports dataset with motion and appearance primitives. The results are expressed in percentage (%).

Category	dv	gs	ki	lf	rd	ru	sk	sw	ss	wl
dv	83.3	0	0	0	0	0	0	8.3	0	8.3
gs	0	68.7	0	0	0	0	0	0	0	31.2
ki	5.2	5.2	36.8	0	0	26.3	5.2	5.2	0	15.7
lf	0	0	0	83.3	0	0	16.6	0	0	0
rd	0	0	0	0	80.0	20.0	0	0	0	0
ru	0	0	25	0	0	58.3	0	0	0	16.6
sk	9.0	9.0	9.0	0	0	0	63.6	9.0	0	0
sw	0	0	5.2	0	0	0	0	94.7	0	0
ss	0	9.0	0	0	0	9.0	9.0	0	72.7	0
wl	0	4.5	4.5	0	4.5	0	0	4.5	0	81.8

In Table 3 is shown the confusion matrix for the proposed descriptor integrating kinematic and shape features. The proposed descriptor was implemented with a total of 455 scalar variables which result very compact and appropriate in applications that demand efficiency in time to obtain results. Some mistakes are reported in confusion matrix among actions such as running and walking because the close dynamic description of such actions. Also, rapid motion on kicking was misclassified with other actions. In UCF Sports dataset the integration of shape information result relevant because the typical gestures of some sports can help with the signature of such actions.

The proposed descriptor was also evaluated in terms of runtime execution at the different stages herein considered. Table 4 summarizes the average of execution time for the proposed descriptor at a frame level. The experiments for execution time were developed on a computer machine with the following hardware features: Intel Xeon(R) CPU E5-1650 v3 at 3.50 GHz - 12 with a 32 GB Random Access Memory. It is worth noting that there is not significant difference for the computation of multiple covariances in the integral representation. This fact result fundamental to implement more robust descriptor without lost computational advantages. A major computational cost is reported in the computation of feature maps. However, additional fast features can be explored and included in the proposed approach.

Table 5 shows the execution time of our final video-level covariance descriptor. In such case, it is presented the execution time average for a Riemannian mean computed from all frame-covariances in each datasets. For sequences with more than 100 frames the proposed approach take in average 6s, which result efficient in different tasks. Also, the proposed approach can be computed in partial sequences for faster prediction of the interactions.

**Table 4.** Table summarizes the computational time average that take the proposed approach at each stage. The experiment were run in a standard computer with a total of 10 and 12 features for UT-interaction and UCF-sport, respectively. As observed, there is not difference to compute many regional covariances in the integral representation. In average, videos on UCF-sport take more time because the number of features considered for the analysis. In the UT-interaction only 275 values are considered in the descriptor, while for UFC a total of 455 values integrate the descriptor.

Dataset	Resolution	Primitives maps [s] (number of features)	Regional covariance [s]	Five regional covariances [s]
UT-1	$373 \times 278$	1.67 (10)	0.53	0.58
UT-2	$336 \times 254$	1.45 (10)	0.43	0.44
UCF-sport	$672 \times 428$	4,99 (13)	0.96	0.97

**Table 5.** An average computational time is herein presented for the different datasets considered in the evaluation. The Riemannian mean achieve a convergence in 6s for more than 100 frames and using 5 different covariances at each frame.

Datasets	Frames	Resolution	Riemannian mean time[s]
UT-1	119	$373 \times 278$	6.24
UT-2	105	$336 \times 254$	6.03
UCF-sport	67	$672 \times 428$	3.72

## 4 Conclusions and Perspectives

In this work was proposed a compact descriptor based on the computation of frame covariances over a set of kinematic and shape features. A Riemannian mean average was computed over the set of frame-covariances to obtain a compact video descriptor. The proposed method was evaluated in UT-Interaction dataset achieving in average up to 80% and up to 61.66% by using only 275 values to describe videos. Also the proposed strategy was evaluated in UCF Sport dataset achieving 71.7% with a descriptor size of 455 scalar values. Future works include the evaluation in additional public datasets and the exploration of new high-order estimations in Riemannian space for covariance matrices. Also, salient regions methods will be included to focus on most important regions along the video sequences.

**Acknowledgements.** The authors acknowledge Vicerrectoría de Investigación y Extensión of Universidad Industrial de Santander for supporting this research registered by the project: Análisis de movimientos salientes en espacios comprimidos para la caracterización eficiente de videos multiespectrales, with VIE code 2347.

## References

1. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(3), 500–513 (2011)
2. Cao, X., Zhang, H., Deng, C., Liu, Q., Liu, H.: Action recognition using 3D DAISY descriptor. *Mach. Vis. Appl.* **25**(1), 159–171 (2014)
3. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2**(3), 27 (2011)
4. Cherla, S., Kulkarni, K., Kale, A., Ramasubramanian, V.: Towards fast, view-invariant human action recognition. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2008, pp. 1–8. IEEE (2008)
5. Fletcher, P.T., Joshi, S.: Riemannian geometry for the statistical analysis of diffusion tensor data. *Sig. Process.* **87**(2), 250–262 (2007)
6. Geronimo, D., Lopez, A.M., Sappa, A.D., Graf, T.: Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(7), 1239–1258 (2010)
7. Gowayyed, M.A., Torki, M., Hussein, M.E., El-Saban, M.: Histogram of oriented displacements (HOD): describing trajectories of human joints for action recognition. In: *IJCAI* (2013)
8. Ji, X., Wang, C., Zuo, X., Wang, Y.: Multiple feature voting based human interaction recognition. *Int. J. Sig. Process. Image Process. Pattern Recognit.* **9**(1), 323–334 (2016)
9. Laptev, I.: On space-time interest points. *Int. J. Comput. Vis.* **64**(2–3), 107–123 (2005)
10. Laptev, I., Caputo, B., Schüldt, C., Lindeberg, T.: Local velocity-adapted motion events for spatio-temporal recognition. *Comput. Vis. Image Underst.* **108**(3), 207–229 (2007)
11. Liu, A.A., Xu, N., Su, Y.T., Lin, H., Hao, T., Yang, Z.X.: Single/multi-view human action recognition via regularized multi-task learning. *Neurocomputing* **151**, 544–553 (2015)
12. Ma, B., Su, Y., Jurie, F.: Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image Vis. Comput.* **32**(6), 379–390 (2014)
13. Pennec, X., Fillard, P., Ayache, N.: A Riemannian framework for tensor computing. *Int. J. Comput. Vis.* **66**(1), 41–66 (2006)
14. Poppe, R.: A survey on vision-based human action recognition. *Image Vis. Comput.* **28**(6), 976–990 (2010)
15. Robertson, N., Reid, I.: A general method for human activity recognition in video. *Comput. Vis. Image Underst.* **104**(2), 232–248 (2006)
16. Ryoo, M.S., Aggarwal, J.K.: UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA) (2010). [http://cvrc.ece.utexas.edu/SDHA2010/Human\\_Interaction.html](http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html)
17. Nour el houda Slimani, K., Benezeth, Y., Souami, F.: Human interaction recognition based on the co-occurrence of visual words. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 455–460 (2014)
18. Soomro, K., Zamir, A.R.: Action recognition in realistic sports videos. In: Moeslund, T.B., Thomas, G., Hilton, A. (eds.) *Computer Vision in Sports*. *ACVPR*, pp. 181–208. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-09396-3\\_9](https://doi.org/10.1007/978-3-319-09396-3_9)

19. Souvenir, R., Babbs, J.: Learning the viewpoint manifold for action recognition. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–7. IEEE (2008)
20. Tuzel, O., Porikli, F., Meer, P.: Region covariance: a fast descriptor for detection and classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 589–600. Springer, Heidelberg (2006). [https://doi.org/10.1007/11744047\\_45](https://doi.org/10.1007/11744047_45)
21. Wang, Y., Huang, K., Tan, T.: Human activity recognition based on R transform. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8. IEEE (2007)
22. Yu, G., Yuan, J., Liu, Z.: Propagative hough voting for human activity recognition. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 693–706. Springer, Heidelberg (2012). [https://doi.org/10.1007/978-3-642-33712-3\\_50](https://doi.org/10.1007/978-3-642-33712-3_50)