A compact and recursive Riemannian motion descriptor for untrimmed activity recognition

Fabio Martinez Carrillo, Michèle Gouiffès, Gustavo Garzon Villamizar, Antoine Manzanera

A compact and recursive Riemannian motion descriptor for untrimmed activity recognition

Fabio Martínez[1], Michèle Gouiffès[2], Gustavo Garzón[1], Antoine Manzanera[3]

1 - Biomedical Imaging, Vision and Learning Laboratory (BIVL^{2}ab). Universidad Industrial de Santander (UIS), Colombia. E-mail: {famarcar, gustavo.garzon}@saber.uis.edu.co

2 - LIMSI, CNRS, Université Paris-Saclay, FRANCE. E-mail: michele.gouiffes@limsi.fr

3 - U2IS/Robotics & Autonomous Systems, ENSTA Paris, Institut Polytechnique de Paris, FRANCE. E-mail: antoine.manzanera@ensta-paris.fr

Abstract – A very low dimension frame-level motion descriptor is herein proposed with the capability to represent incomplete dynamics, thus allowing online action prediction. At each frame, a set of local trajectory kinematic cues are spatially pooled using a covariance matrix. The set of frame-level covariance matrices forms a Riemannian manifold that describes motion patterns. A set of statistic measures are computed over this manifold to characterize the sequence dynamics, either globally, or instantaneously from a motion history. Regarding the Riemannian metrics, two different versions are proposed: (1) by considering tangent projections with respect to updated recursive statistics, and (2) by mapping the covariance onto a linear matrix using as reference the identity matrix. The proposed approach was evaluated for two different tasks: (1) for action classification on complete video sequences and (2) for online action recognition, in which the activity is predicted at each frame. The method was evaluated using two public datasets: KTH and UT-Interaction. For action classification, the method achieved an average accuracy of 92.27 and 81.67%, for KTH and UT-interaction, respectively. In partial recognition task, the proposed method achieved similar classification rate as for the whole sequence by using only the 40% and 70% on KTH and UT sequences, respectively.

Keywords - Activity recognition, Motion descriptor, Motion analysis, Motion trajectories

1. **Introduction**

In computer vision, many applications require to describe or characterize the dynamics or the activities occurring in a video or in a region of interest [22,23]. Motion descriptors can either be hand-crafted or learned from a large database to address a specific application. In each case, the choice of the optimal duration of the spatio-temporal volume to be described is an open problem. In addition, most approaches require the reading and processing of the whole sequence to provide a descriptor, instead of producing one result at each frame. In addition, a good descriptor has to be discriminant, while being compact and invariant to different phenomena such as illumination changes, partial occlusions, strong variability within the classes, appearance variations, and large pose variations.

In the domain of action recognition, scenarios are more and more complex, but it is generally assumed that the activity is present from the beginning to the end of the input sequence, i.e., its temporal support interval is a priori delimited. Despite significant efforts, the proposed methods are very sensitive to action phase shift, which induces considerable limitations in real-time applications. For instance, surveillance or online video indexing systems require any-time prediction capability on untrimmed videos.

The main contribution of this work is a frame-level recognition approach that is able to recognize activities in a continuous video stream. First, a set of kinematic cues are computed from a semi-dense field of trajectories, which represent atomic activity motions, that are relatively invariant to appearance. At each frame, the set of computed cues are spatially aggregated within a covariance matrix. Along the sequence, the set of frame level covariance matrices forms a special Riemannian manifold, whose geometry represents the dynamic of the activities. Hence, a set of recursive statistics is computed along the manifold to capture the main action dynamic with the potential capability to be updated at each frame. Such statistics are used to outline the temporal deformations of the manifold, and allow to predict the ongoing activity in the current frame. The statistics computed along the manifold are the mean and variance as well as the forgetting version of the maximum and minimum of the covariance matrix at each frame. Finally, the updated version of the motion descriptor is mapped to a SVM and a predicted label of the activity is returned for the current frame.

2. **Related work**

2.1 **Motion features**

Motion descriptors computed along dense point trajectories have been successfully used to represent activities and interpret video sequences [7, 25]. The most popular descriptors based on these trajectories are formed by local features such as HOF (Histograms of Optical Flow), MBH (Motion Boundary Histograms) and HOG (Histograms of Oriented Gradients), which are integrated within space-time volumes centered around each trajectory. In [7] and [24] local descriptors are also computed around each trajectory, where

motion trajectories are improved by correcting the camera motion. A major limitation of such descriptors is that the spatio-temporal volumes are empirically cut off to a fixed temporal length (for example 15 frames in [25]), which could result restrictive to represent a large variety of activities including long or non-periodic actions. Additionally, such works are based on the classical Bag-of-Features methodology that namely requires the complete computation of spatio-temporal volumes to effectively compute the occurrence histograms and then assign an activity signature to the video.

Other works have focused on dynamic characterization of dense beams of trajectories using for instance a regional motion characterization has been proposed by computing chaotic invariant features using a sparse coding method [27]. Likewise, in [10], a trajectory clustering is proposed to analyze and identify dominant motion regions. Such regions are used as reference to filter out camera motion and to coarsely segment regions respectively related to background and to foreground. Then, local patch descriptors are computed along the foreground trajectories and code-words are defined for their representation.

2.2 **Decision and learning strategies**

Currently, deep learning approaches have emerged on AR to automatically learn discriminative appearance and motion features, and perform prediction of relevant events on video sequences [8, 17]. For instance, trajectory-pooled deep convolutional descriptors have integrated convolutional feature maps learned from appearance and motion streams ConvNet along trajectories [26]. In such case, the trajectory locations are used as spatial support of the local ConvNet responses. Although these learning approaches achieve high accuracy rates on realistic datasets, they require huge quantity of training data to optimise ConvNet filter parameters. Additionally, the descriptor formed by ConvNet regions around trajectories is extremely large, which is prohibitive in real-time applications. In spite of growing importance of such convolutional descriptors, the design of such architectures remain empirical, with dependence of samples amount to achieve a proper representation. Specifically for video analysis, the coding of temporal information remains an open problem. For instance, it has been observed in [19] that the consecutive application of independent spatial (2D) and temporal (1D) filters achieved better performances than 3D spatio-temporal filters.

Compact descriptors based on special manifolds have taken advantage of the data topology to carry out the video representation. For instance, optical flow features, like velocity, gradient and divergence, and shape / appearance features were coded and embedded in covariance matrices by using a linear sparse representation [6]. Nevertheless, such works are limited to off-line recognition where the global statistics of the whole video have to be collected to eventually form the activity descriptor.

2.3 **Online action recognition**

Regarding the online activity prediction, Gaidon et al. [5] proposed an incremental action representation by computing sequences of actoms that consider the temporal evolution of the activities. A main limitation of this approach is the supervised learning from annotated atomic action units to exploit such relationship. A rank learning machine is proposed in [2] to analyze the video-wide temporal evolution, under the assumption that temporal ordering of the activities is preserved. This ranking strategy captures the appearance patterns to model evolution of actions during time. Nevertheless, this representation can fail because of appearance pattern dependency, or in sequences with important action occlusions.

Exhaustive learning strategies have been also proposed to deal with recognition of partial sequences. For instance, Varol et. al proposed long-term temporal convolutions (LTC-CNN) to represent actions [20]. This architecture learns from (3D) kernels applied on several time length intervals. This approach however requires fixed video divisions, being inadequate for untrimmed sequences. Also, in [21], a differential recurrent network is proposed to recover salient motions to represent the dynamic evolution of actions. This approach requires large size descriptors at each frame with additional requirements for dimensionality reduction, which results prohibitive in online applications. In contrast, the proposed approach describes actions from a very compact per-frame representation and provides a result at each time during the video sequence.

3. **Covariance manifold video representation**

Despite the evidence that the human visual system can recognize activities in a reactive and instantaneous manner, most of the proposed approaches are designed to find spatio-temporal patterns over complete sequences. Then, a major challenge for online recognition is to capture temporal evolution of the activities from a proper dynamic representation. The pipeline of the proposed approach is illustrated in Figure 1.

3.1 **Local trajectory motion cues**

The proposed approach starts by computing a set of kinematic measures over dense trajectories, that serves as a low-level video representation. A set of improved dense motion trajectories are computed as reference to build the video descriptor [26]. Instead of using trajectories as central axis of neighboring block size descriptors [24, 26], we directly exploit the dynamic information of trajectories. Each trajectory represents a particle traveling from time t_1 to t_n, following a sequence of coordinates $\Gamma(t) = \{(x_t, y_t) \in \mathbb{R}^2\}_{t=t_1}^{t_n}$ estimated from optical flow. Then, a set of kinematic trajectory features (KTF) is computed along each trajectory. Herein, it was considered as KTF:

- the velocity v_t , depicted by its direction $\theta_t = \arg v_t$ and modulus $s_t = \| v_t \|$.
- the normal acceleration a_t^N, representing the norm of the acceleration component toward the curvature of the trajectory.
- the tangential acceleration a_t^T , representing the norm of the acceleration component along the trajectory.
- the curvature κ_t of the trajectory.

These features can be used separately or jointly, and the proposed framework is flexible to include any kind of local features. The resulting set of d kinematics forms the local KTF descriptor associated to a trajectory Γ at time t. A set of KTF is then computed over active trajectories estimated from a dense grid of points and expressed as a spatial function $K_t(x) = \{s_t(x), \theta_t(x), \dots, \kappa_t(x)\}$, where each kinematic component $k_t^i(x)$ is a scalar image only defined at active points where trajectory information exists at time t (See an example in Figure 1-(b)).

3.2 **Frame covariance representation**

The covariance representation allows to compactly describe actions by measuring the correlation degree within the set of KTF (K), at each frame. Specifically, at each time t is computed a KTF covariance matrix C_t, expressed for any pair of kinematics features as:

$$C_t(i,j) = \frac{1}{n}\sum_{k=1}^{n} k_t^i(x_k)k_t^j(x_k) - \frac{1}{n^2}\sum_{k=1}^{n} k_t^i(x_k)\sum_{k=1}^{n} k_t^j(x_k) \quad (1)$$

where n is the number of active trajectories at time t, and k_t^i, $i = 1, \dots, d$ is a particular kinematics. The covariance matrix $C_t \in \mathbb{R}^{d\times d}$ is symmetric $(C_t = C_t^T)$ and positive $(\det(C_t) > 0)$. It allows to describe and summarize complex KTF patterns, the diagonal being the partial variance of each kinematic feature, and the rest the covariances between features.

3.3 **A global intrinsic video-covariance descriptor**

To get global action description, it is necessary to temporally take statistic measures over frame-covariance matrices that represent a video sequence. This set of covariance matrices, that are symmetric and positive-definite, form a convex half-cone subset on $\mathbb{R}^{d^2}$. This subset is not a vector space, but a Riemannian symmetric space and therefore the use of Euclidean metrics are not suitable to compute statistics [4, 13]. For instance, some Euclidean-based statistic over such symmetric matrices could result on covariance estimations with negative eigenvalues. Hence, the sequence of KTF covariance matrices $C = (C_1, C_2, C_3, \dots, C_N)$ computed along the video, lies within a Riemannian manifold $\mathcal{M}$, whose geometry can be associated to a particular action. Each KTF covariance C_t corresponds to a point on a curved Riemannian manifold $\mathcal{M}$, as illustrated by Figure 1. So, a video descriptor could be defined as a set of measures over such manifold to summarize a particular action or gesture.

To take measures over such manifold [13], each covariance point should be projected to a tangent plane (which is a vector space), from a logarithmic operation $(\mathcal{M} \xrightarrow{\log(C_t)} \mathbb{T}_{C_t}(\mathcal{M}))$. In a same way, a projected covariance could be mapped from Euclidean space to original Riemannian manifold, following exponential operation $(\mathbb{T}_{C_t}(\mathcal{M}) \xleftarrow{\exp(C_t)} \mathcal{M})$. Hence, intrinsic manifold statistics such as mean and variance of manifold points can be defined from such mapping projections. Particularly, the mean of $\mathcal{M}$ in a set of covariance matrices C can be iteratively found by considering it as an optimization problem where the mean μ is the point (matrix) with minimum distance ρ among the sample covariance matrices [4]:

$$\mu(C) = \arg\min_{\mu\in\mathcal{M}} \left[\frac{1}{2N}\sum_{t=1}^{N} \rho(\mu, C_t)^2\right]$$

Regarding the variance, according to Fréchet definition, such measure can be expressed as the expected value of the square distance from the mean [4]. The $\log_{\mu_{t+1}}(C_t)$ operation is the geodesic distance between a particular covariance C_t and the expected value μ_{t+1}. Then Euclidean norm of such geodesic over all covariance matrices constitutes an approximation of the variance (see expression in Algorithm 1). Finally, the global video descriptor can be summarized as mean and variance of Riemannian manifold $\mathcal{M}$, computed as described in Algorithm 1.

Algorithm 1 Global video descriptor from intrinsic Riemannian measures

Input: $C = (C_1, C_2, C_3, \dots, C_N)$
1: start with: $\mu_0 = C_1$
2: **repeat**
3: $\quad X_k = \frac{1}{N}\sum_{i=1}^{N} \log_{\mu_t}(C_i)$
4: $\quad \mu_{k+1}(C) = exp_{\mu_k}(X_k)$
5: **until** $\|X_k\| < \varepsilon$
6: $\sigma^2(C) = \sum_{i=1}^{N} \left\|\log_{\mu_{k+1}}(C_i)\right\|^2$
Output: $[\mu_{k+1}(C), \sigma^2(C)]$

The stop criterion, in line 5 of Algorithm 1, is defined as $\|X_k\| = \sum_{i=1}^{N}(\log(\lambda_i))^2$ where λ_i are the respective eigenvalues. The log-map and exponential projection with respect to μ_k are defined as:

$$\Omega_{\mu_k}(C_i) = \mu_k^{\frac{1}{2}}\Omega(\mu_k^{-\frac{1}{2}}C_i\mu_k^{-\frac{1}{2}})\mu_k^{\frac{1}{2}}$$

where $\Omega = \{log, exp\}$. It is also expected that the variance inherits such error since it affects the base of the log-map. A final video descriptor is then formed by the concatenation of intrinsic measures such as the mean and the variance $V_d = \{\mu(C), \sigma^2(C)\}$.

4. **Recursive covariance metrics**

One of the principal goals of this work is to compute a reactive (on-the-fly) representation of the activities at each time of the sequence. From analysis of global Riemannian measures, recursive and partial measures were herein considered to update motion history over partial manifold while the sequence is run. For doing so, two different strategies were considered, described in the next subsections.

4.1 **Recursive measures by Proximity Mapping**

Under the assumption that consecutive covariance matrices are closer within the manifold, we can project covariance points

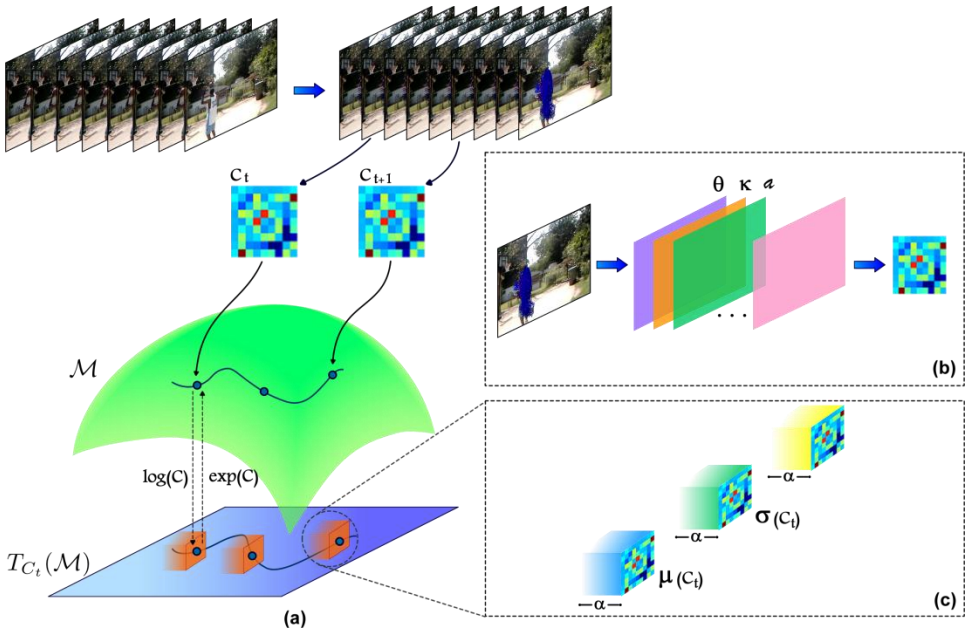

Fig. 1 Online action recognition from a Riemannian video representation. In (a) the method starts by computing point trajectories, along which the local motion cues are calculated. Then, at each frame, this kinematic information is spatially aggregated in a covariance matrix (see plot (b)). The set of frame-level covariances then forms a Riemannian manifold. To achieve a frame level recognition, a set of time-recursive statistics are computed in the manifold space (see plot (c)).

on Euclidean space, taking as reference the previous covariance. In such case, the geodesics between manifold points, corresponding to successive frames, will exhibit a small value. Hence, a recursive version of mean starts by assuming $\mu_0 = C_1$ as the covariance computed in first frame where KTF are available. The propagation and updating of mean is then achieved by log-projecting covariance C_t with respect to historical mean μ_{t-1}. This projection computes the geodesic distance between the mean and any new point in the covariance space. An α coefficient is introduced to weight the contribution of each new C_t into the recursive mean covariance. In such case, larger α values give more importance to current C_t observations, and temporal mean is significantly affected. After that, a back-projection to the manifold is achieved by computing the respective exponential operation. The resulting mean is then expressed as:

$$\mu_t = exp_{\mu_{t-1}}(\log_{\mu_{t-1}}(\alpha C_t))$$

where $log_{\mu_{t-1}}$ and $exp_{\mu_{t-1}}$ follow the rule expressed in equation 3.3. This formulation allows to perform incremental measures within the manifold $\mathcal{M}$, as action is developed throughout the sequence. At each time, if we measure the square distance between the updated mean and each new covariance C_t, we obtain a recursive estimation of the variance. To initialize such statistics, we assume the initial variance as $v_1 = exp_{C_1}(\|\log_{C_1}(C_2)\|_2)$.

This way, a recursive variance can be expressed into a interpolation scheme, as follows:

$$v_t = exp_{v_{t-1}}\left[\log_{v_{t-1}}(\alpha\|\log_{\mu_t}(C_t)\|_2)\right] \quad (2)$$

This recursive measure assumes local tangent planes regarding close recursive points. The recursive variance and mean can constitute a untrimmed and online action recognition description to represent partial dynamics coded into consecutive KFT covariances. It should be noted that each recursive statistics herein computed remains in the Riemannian manifold, as illustrated in Figure 1-(a).

4.2 **Recursive measures from Identity Mapping**

An alternative to compute recursive measures is to project manifold points to a tangent plane with respect to the identity. Hence, the projection to Euclidean and Riemannian space from identity, could be computed as $\Omega(C_t) = \Sigma_t \Omega(\lambda_t)\Sigma_t^T$, where Σ are the eigenvectors of the matrix and λ are the respective eigenvalues of matrix C [14]. Again, Ω represents either function log or exp and the recursive statistics remain on Euclidean space.

Then, an initial recursive mean from this approximation is defined as $\log(\mu_0) = \log(C_1)$. So, a progressive version of the mean is achieved by projecting each new C_t w.r.t the identity and updating previous mean μ_{t-1}. In this case too, an α value allows to weight the importance of mean KTF covariance history. Then, the recursive mean can be expressed as follows:

$$\log(\mu_t) = \log(\mu_{t-1}) + \alpha(\log(C_t) - \log(\mu_{t-1})) \quad (3)$$

Accordingly, we can estimate a recursive variance by projecting each of the points to an identity Euclidean space, and defining a recursive square distance regarding the expected value. This recursive variance can be expressed as:

$$\log(v_t) = \log(v_{t-1}) + \alpha((\log(C_t) - \log(\mu_t))^2 - \log(v_{t-1})) \quad (4)$$

An additional advantage of such general Euclidean projection with respect to the identity is the easy and intuitive extension to other statistics and measures. For instance, we can define non-linear and recursive operators, such as the forgetting minimal $\log(\min_t)$ and maximal value $\log(\max_t)$, expressed as:

$$\log(\max_t) = \alpha \log(C_t) + (1-\alpha) \max(\log(C_t), \log(\max_{t-1}))$$

$$\log(\min_t) = \alpha \log(C_t) + (1-\alpha) \min(\log(C_t), \log(\min_{t-1}))$$

All the recursive measures described above (or a subset of them) constitute a video descriptor that is able to operate at frame level and allows to obtain a video representation at each time of the sequence. Such estimation can be integrated to obtain a more robust description of the partial activities along the sequence.

5. **Evaluation and results**

In this work we are interested in measuring the capability of the proposed covariance manifold to represent human activities in videos. Two tasks are considered: global action classification and per-frame action recognition. The proposed approach is evaluated in terms of recognition accuracy but also in terms of complexity of the algorithm and size of the descriptor. In both schemes, the video descriptor is constituted by measures over the Riemannian manifold, namely, the temporal mean and variance of spatial covariances. Each of the video-descriptors is mapped to a previously trained classifier. The Support Vector Machines (SVM) is chosen as action classifier, because of its efficiency in terms of inference time, which is compatible with our fast online approach. Another determining advantage of SVM is its flexibility: higher dimensional boundaries are conveniently obtained by mapping the samples to a feature space using a non-linear kernel function. For the following experiments, a Radial Basis Function kernel (RBF) produced desirable results with acceptable processing time. It should be noted that the SVM works into a Euclidean space and therefore any covariance measure is mapped as $\log(C_t)$. To assess the relevance of the different kinematic features and temporal statistics of our descriptor, different combinations of KTS and statistics were evaluated.

Scheme 1: Global recognition (detailed in 5.2). For action classification, a unique prediction was considered for whole video sequence. The SVM is trained with intrinsic global statistics computed from complete sequences. Then for each new video a set of KTF is computed and coded as per-frame covariance matrices, forming a temporal video manifold. A set of intrinsic statistics, namely, mean and variance are taken from such manifold to represent the action, which thereafter is mapped to the SVM.

Scheme 2: Online recognition (detailed in 5.4). Regarding the evaluation of action recognition, we compute and update recursive statistics from new KTF covariance at each frame. The SVM was trained also with recursive estimations, from training videos, at different sections of video sequences. For testing, the video descriptor is available at each frame for every video. Hence, this recursive video descriptor was mapped to the trained SVM to obtain an online prediction. The set of proposed recursive metrics depends on the temporal scale α that represents the memory depth: larger α scales consider larger intervals of time. Different α values are evaluated: $\alpha \in \{2^{-5},\ 2^{-6},\ 2^{-7},\ 2^{-10}\}$.

5.1 **Data**

The proposed approach has been evaluated on two well-known public human action datasets. Here is a brief description of these datasets:

- KTH [11] contains six human action classes: walking, jogging, running, boxing, waving and clapping. Each action is performed by 25 subjects in four different scenarios with different scales, clothes and scene variations. This dataset contains a total of 2391 video sequences. Each video has a spatial resolution of 160×120 pixels and a frame rate of 25 fps. The proposed approach was evaluated following the original experimental setup which specifies training, validation and test groups of files as well as a five-fold cross validation suggested in [16].

- UT-Interaction [15] contains six different human interactions between different people: shake-hands, point, hug, push, kick and punch. The dataset is made of two subsets of 60 sequences, with all 6 actions but one with static camera (UT-set1) and the other one (UT-set2) with some jitters camera motions. Each video has a spatial resolution of 720×480 and a frame rate of 30 fps. A ten-fold leave-one-out cross-validation was performed, as described in [15].

5.2 **Global Action evaluation**

5.2.1 Impact of the kinematics

The first evaluation aims to evaluate the relevance of individual kinematics in the action recognition task, and also to define the sets of kinematic features that provide a good trade-off between accuracy and computation costs. Then, resulting covariance of each kinematic feature was computed together with spatial (x,y) location and time activation t. Figure 2 illustrates the performance achieved by the proposed approach using KTH dataset, first for each type of combination of isolated kinematic features. In general, single features correlated with spatial coordinates report an interesting classification accuracy, as shown by Figure 2(a). The addition of temporal trajectory activation further improves the performance (Figure 2(b)). Interestingly, using only the velocity angle to form the vector $[\theta, x, y, t]^T$, the proposed approach reaches a performance of 80%. Figure 2(c) finally displays the performance when adding successively the best kinematic feature into the covariance descriptor. The best result is achieved with a per-frame covariance that integrates all kinematics, obtaining an average score of 92.27%.

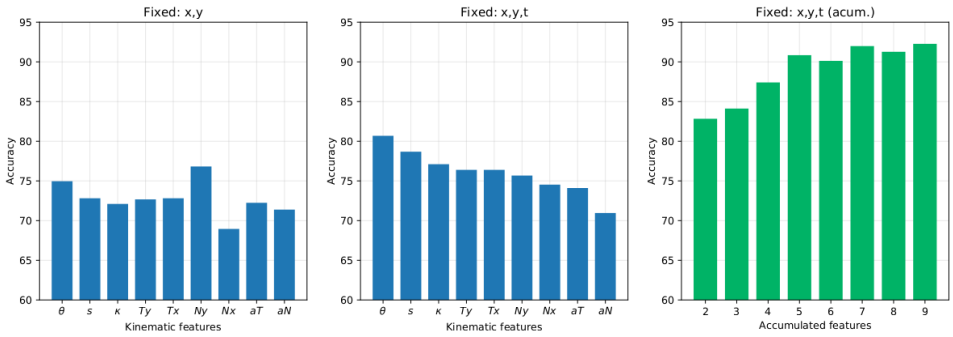

Fig. 2 Impact of the kinematic features on the classification accuracy, using the KTH datasets. (a) evaluation of a single kinematic feature f included in the descriptor vector $F_a = [f\ x\ y]^T$; (b) same experiment with temporal variable $F_b = [f\ x\ y\ t]^T$; (c) evaluation of the multiple descriptor when increasing the number of kinematic features (ordered from better to worse as individual feature) $F_c = [f_1\ \ldots\ f_n\ x\ y\ t]^T$.

Figure 3 illustrates the performance achieved by the proposed approach on the two UT-interaction sets. Here also, a higher accuracy is obtained by including the temporal variable. As expected, a much better performance is achieved on UT-set1 because of the relative static camera and the plane background. In this case, kinematics of first order (s, θ, N_x) achieved the best performance of action representation with 73.3%, 71.6% and 70% respectively. Nevertheless, it should be noted that in set UT-set2, some individual features (s, N_y) both achieve a classification rate of 65%, and that the best accuracy for multiple features is achieved with only the best three kinematics, i.e., (s, N_y, a_T) for an average accuracy of 68%. Regarding UT-Interaction sequence 1, additional testing has shown that a significant improvement is obtained by excluding kinematic T_y, which generates an accuracy of 85% for features $\{\theta, s, \kappa, T_y, N_y, N_x, a_T\}$. Interestingly, when kinematic a_N is added, performance decreases to 81.6%. As for UT sequence 2, a slight improvement occurred when concatenating kinematics $\{s, N_y, a_N, T_y\}$ obtaining 68.3%.

Recursive and intrinsic statistics can be used as a global descriptor of video sequences. In the next experiments we analyze the global performance of such statistics in a classification task. First, a covariance sample is computed in each frame. Then, global statistics are computed as general descriptors of the video content. Figure 4 collects the resulting classification scores. Best classification scores are achieved by using the computation of intrinsic mean (I. Mean) for full video sequences, following the algorithm proposed by Fletcher [3]. Also, recursive mean and variance versions were herein used as global classifiers to compare the performance with intrinsic measures.

5.3 **Global classification**

Table 1 illustrates the comparison of best achieved result on KTH w.r.t baseline approaches based on local dense blocks [11] and trajectory-based descriptors [24]. The proposed approach achieved a performance of 92.27% comparable with state of the art techniques, but using a descriptor size of only 78 scalar values. In contrast, trajectory based descriptors [24] take more than 4000 scalar values, computed after an BoW occurrence histograms. In table 1 it is worth noting that the proposed descriptor is much more lightweight, which increases the possibility to carry out classification in very short time, and also to be used as a complementary descriptor to analyze complex scenarios and movements with negligible additional cost.

Approaches	**KTH**	
	Size	**Accuracy**
Laptev 2008 [12]	4000	91.80
Wang 2011 [25]	4000	94.20
Our method	78	92.27

Table 1 Comparison with state of the art methods on the KTH dataset. Comparable methods, i.e. based on pooling local motion descriptors, were chosen as baseline: local block (Laptev 2005) and trajectory based descriptors (Wang 2011).

Table 2 summarizes the achieved scores on UT-interaction dataset by the proposed approaches and some state-of-the-art strategies. In UT-Interaction the best classification scores, using the same descriptor, achieved in average 81.6% and 65%, for UT-set1 and UT-set2 respectively. After a selection of kinematic features, on UT-set1 was achieved an accuracy of 85%. Because UT-interaction sets exhibit more complex activities, with dynamic backgrounds, the temporal variance of spatial covariance was required to achieve a better performance.

The main advantages of the proposed approach are the compactness and the efficient computation of video descriptor. Best baseline results are achieved by propagative voting with a computational cost of $O(N_M) + O(WHT)$, being N_M the number of matches over a sequence with resolution W, H, T. In such case, the N_M is computed by using random projection trees, an expensive and prohibitive strategy for online applications. Other baseline approaches require relative large descriptors with a natural dependence of number of points (np), which are proportional to image dimension.

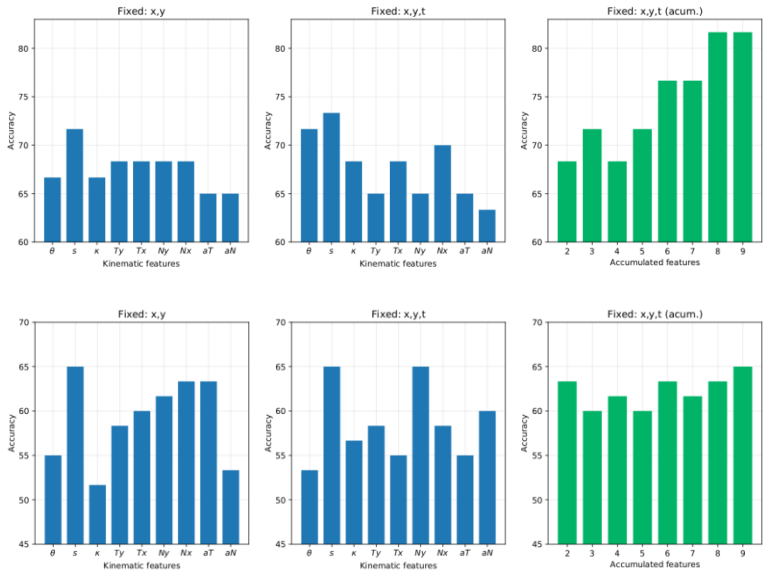

Fig. 3 Kinematic feature evaluation using both UT-interaction sets: the UT-set1 (top) and the UT-set2 (bottom). As for KTH (see in Figure 2) the evaluation was carried out including spatial (first column) temporal correlations (second column). Similarly, for UT was analyzed an incremental kinematic descriptor.

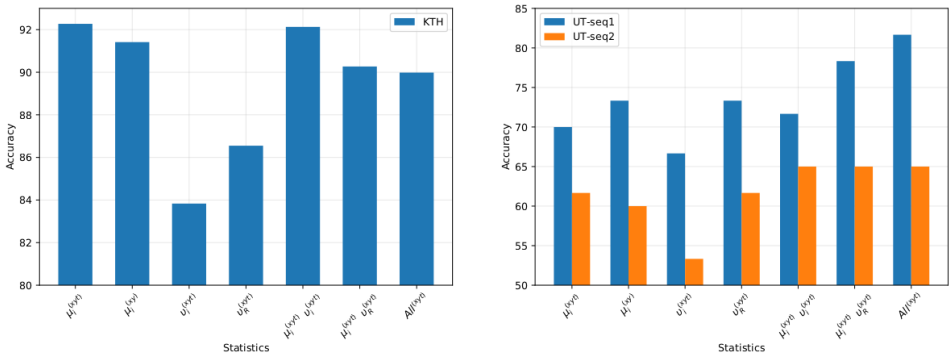

Fig. 4 Evaluation of complete (offline) video descriptors by using different statistics (Left: KTH, right: UT-Interaction). In such case it was computed for whole sequence intrinsic and recursive mean and variance statistics. This chart also illustrates the performance of combining different statistics.

Approaches	**Size**	**UT-Set 1 Accuracy**	**UT-Set 2 Accuracy**
Propagative voting [28]	162 × np	93.3	91.7
Daysy [1]	192 × np	71.67	56.67
Laptev [11] + SVM	41800 × np	68	65
Slimani 2014 [18]	22500	40	66
Xiaofei [9]	252	83	-
Proposed approach	312	81.6	65

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.

Table 3 summarizes typical times of our non-optimized implementation. Each of the steps of the proposed approach were measured and the average time is reported for each video sequence. Column (a) reports the average times to compute all motion trajectories on each video sequence and column (b) collects the times needed for the kinematic computation. Because the time for each kinematic is negligible, it was only reported time for all kinematics on each video sequence. In column (c) is reported the time consumed to build whole per-frame covariance and compute the video descriptor for each sequence. The last column reports the time to train a SVM model. The experimental setup involved used one core of an

Intel Xeon CPU E5-1650 v3 at 3.50GHz with 32Gb RAM with a Gnu C++ compiler.

5.4 **Online Action evaluation**

The recursive statistics proposed in section 4 allowed to get a description of the video activity at each frame. Table 4 and 5, and table 6 and 7 collect the classification accuracy for each configuration, on dataset KTH and UT-Interaction respectively. For these experiments, all 9 kinematic features are embedded in the feature vector. These results summarize performance of different recursive statistics for different α values for both datasets. First of all, the best results are obtained by using the mapping onto the Identity space: 89.93% versus 82.37% for KTH; 80% versus 76.66% for UT-Interaction 1; 78.33% versus 76.66% for UT-Interaction 2. In addition, the statistics calculated after mapping are faster to compute. Also, the computation of recursive statistics results very close in accuracy w.r.t the descriptor built from the intrinsic measures. Additionally, the assumption of the identity reference preserves the accuracy while remaining faster to compute w.r.t. the other statistics. Best performance was achieved with large scales, because the smoothness on statistics tends to provide a more stable action representation. The scale has a varying impact on the accuracy results. It both depends on the dataset and on the nature of the statistics. As for the scores over the descriptor formed by all-statistics shown the best performance, achieving a performance of 89.93 and 82.37% for *Identity Mapping*, i.e. the mapping in the Euclidean space according to the identity reference and the *Proximity Mapping*, respectively.

	(a)	(b)	(c)	(d)
KTH	0.971 ms	105.65 ms	1.94 s	36.06 s
UT-Interaction	47.16 ms	73.51 ms	9.08 s	3.13 s

Table 3 Average computation time for each stage for intrinsic mean method over KTH (top) and UT-Interaction set 1 (bottom) datasets using 78 and 312 scalar values respectively. Stages: (a) Reading all trajectories, (b) calculating kinematics for each sequence, (c) frame descriptor for each sequence and (d) SVM model training.

As expected, KTH and UT-Interaction (set 1) datasets showed accurate results since the camera motion and artifacts are not significant on controlled environments. Thanks to the natural flexibility of the covariance framework, we can easily introduce additional features in order to be more robust to some common issues like illumination and appearance variability.

One of the main contributions of this work is the online character of the proposed descriptor, that allows to predict actions at any time of the sequence. A final evaluation was carried out to test the performance of the proposed approach to predict partially developed actions. To that purpose, the accuracy of the classification is computed for different percentages of the ongoing action. Figure 5 illustrate the performance of the proposed approach to represent partial action along the video sequences at different percentages of the video. The different combinations
of statistics were evaluated in such online and untrimmed recognition. In each case, two temporal scales are considered, namely $\alpha = 2^{-5}$ and $\alpha = 2^{-10}$. As expected, the proposed KFT recursive covariance coding allows a robust representation of partial actions. Since KTH exhibits in many cases periodic actions, a very compact descriptor is possible. As expected, best performance is achieved by using all the recursive statistics, but using only the recursive mean already achieves competitive results. In such case, the video descriptor is summarized in only 156 scalar values, which results very useful for embedded applications.

For this dataset, and for some statistical combinations, using only 50% of video sequences achieves more than 70% of accuracy on recognition task. Thereafter, 80% of video sequences is sufficient to achieve a maximum accuracy score for the whole video sequences. In both time intervals (i.e. values of 2^{-5} and 2^{-5} for α), was observed same performance of action coding, being the more stable result the combination of all the statistics, but resulting interesting the performance of mean and min statistics alone.

Same evaluation was carried out over the two UT-interaction sets. Figure 6 summarizes the online recognition over UT-set1, by using different combinations of recursive statistics, computed with same two α values. First of all, it should be noted that for the first 20% of the video segments a random classification is obtained. This is due to the fact that most of the time, there is no motion in the first frames of the video-sequence, and therefore the KFT recursive covariance matrices are not meaningful. For UT-set1, the best results are achieved by using short motion history memory, i.e. with $\alpha = 2^{-5}$, that achieves a progressive increasing of accuracy until 80%. Best performance is also obtained by statistics whose covariances are projected w.r.t to a common tangent plane, following the identity matrix. Such fact could be associated to stability of common tangent plane, and also less numerical error because in such strategy it is not necessary to back-project to Riemannian manifold from exponential operation. For small α values, the mean and the minimum statistic measures result very effective and compact to recognize online actions.

Figure 7 shows the online prediction performance for different recursive statistics in the more challenging UT2 dataset. It is interesting to observe that statistics projected over Riemannian manifold result more stable, with an increasing trend for complete KFT video descriptor, but also for combination of non-linear operation (min and max) with the recursive mean. For the most complete combinations, an efficient video representation is achieved after 60% of the video-sequence.

6. **Discussion and concluding remarks**

We proposed a very compact covariance-based descriptor for untrimmed video action recognition. The proposal starts with a multiple kinematic motion representation, in the purpose to

Statistic over $\mathbb{E}$ / Scale($-\log_2(\alpha)$)	**5**	**6**	**7**	**10**
$[\log(\mu_t)]$	85.02	80.22	76.69	77.68
$[\log(\upsilon_t)]$	74.43	74.85	70.19	69.35
$[\log(min_t)]$	83.05	85.02	85.16	84.74
$[\log(max_t)]$	84.46	85.59	86.15	87.14
$[\log(\mu_t), \log(\upsilon_t)]$	85.59	81.49	79.51	77.82
$[\log(\mu_t), \log(min_t)]$	87.28	88.70	87.42	86.44
$[\log(\mu_t), \log(max_t)]$	86.86	86.29	82.90	84.60
$[\log(\mu_t), \log(\upsilon_t), \log(max_t), \log(min_t)]$	88.98	88.70	88.84	**89.83**

Table 4 Experiments results using Identity Mapping on the KTH dataset. The whole 9 Kinematics are used to compute the statistics, which are combined to form different descriptors.

Statistic over $\mathcal{M}$ / Scale	**5**	**6**	**7**	**10**
$[\mu_t]$	77.28	74.39	71.72	69.52
$[\upsilon_t]$	70.10	67.32	69.87	69.87
$[\mu_t, \upsilon_t]$	77.86	75.78	73.46	72.19
$[\mu_t, \log(min_t)]$	80.64	81.22	81.34	81.34
$[\mu_t, \log(max_t)]$	78.79	79.95	78.33	78.56
$[\mu_t, \upsilon_t, \log(min_t), \log(max_t)]$	81.34	81.34	**82.73**	80.18

Table 5 Experiments results using Identity Mapping on the UT-Interaction dataset. An interesting point is combination with non-linear recursive statistics that were computed from identity matrix.

Statistic over $\mathbb{E}$ / Scale	**5**		**6**		**7**		**10**	
	Set 1	**Set 2**	**Set 1**	**Set 2**	**Set 1**	**Set 2**	**Set 1**	**Set 2**
$[\log(\mu_t)]$	76.66	60	65	50	53.33	41.66	46.66	45
$[\log(\upsilon_t)]$	63.33	53.33	50	40	45	40	50	40
$[\log(max_t)]$	61.66	55	70	60	66.66	61.66	68.33	65
$[\log(min_t)]$	56.66	60	61.66	66.66	70	70	76.66	71.66
$[\log(\mu_t), \log(\upsilon_t)]$	71.66	60	60	50	56.66	45	50	45
$[\log(\mu_t), \log(max_t)]$	73.33	70	78.33	73.33	73.33	**78.33**	70	73.33
$[\log(\mu_t), \log(min_t)]$	**80**	61.66	75	70	70	70	66.66	65
$[\log(\mu_t), \log(\upsilon_t), \log(max_t), \log(min_t)]$	76.66	66.66	75	71.66	70	75	75	75

Table 6 Experiments results using Identity Mapping on the UT-Interaction dataset. Statistics were computed at different scales that take different intervals of time and tested on both sets.

Statistic over $\mathcal{M}$ / Scale	**5**		**6**		**7**		**10**	
	Set 1	**Set 2**	**Set 1**	**Set 2**	**Set 1**	**Set 2**	**Set 1**	**Set 2**
$[\mu_t]$	**76.66**	58.33	66.66	41.66	60	35	48.33	35
$[\upsilon_t]$	51.66	41.66	43.33	40	45	41.66	46.66	31.66
$[\mu_t, \upsilon_t]$	66.66	56.66	60	41.66	53.33	40	48.33	35
$[\mu_t, \log(max_t)]$	68.33	70	75	71.66	66.66	**76.66**	66.66	66.66
$[\mu_t, \log(min_t)]$	**76.66**	65	65	65	61.66	66.66	58.33	68.33
$[\mu_t, \upsilon_t, \log(min_t), \log(max_t)]$	71.66	61.66	73.33	66.66	70	71.66	70	68.33

Table 7 Experiments results using Proximity Mapping on the UT-Interaction dataset. The whole statistics and their respective combinations were evaluated at different scales. The combination of mean with non-linear operators result the best combination on this approach.

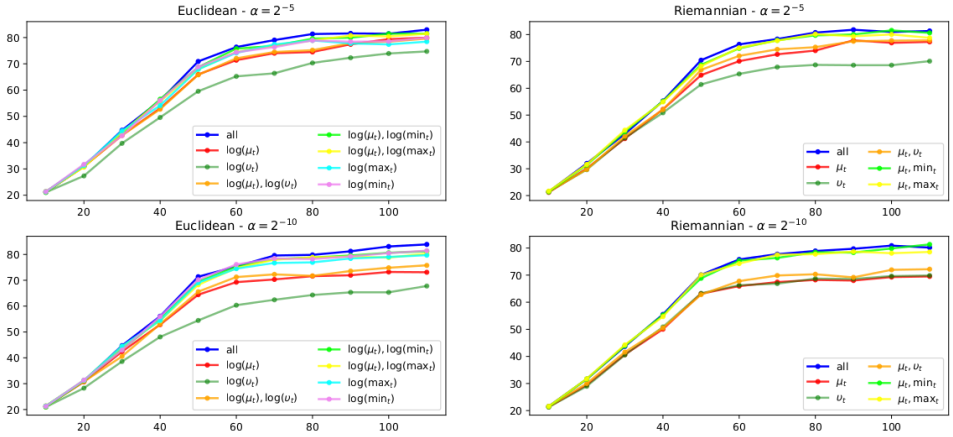

Fig. 5 KTH action online recognition. Video descriptors that code different statistic combinations were evaluated. As expected, a growing accuracy is achieved while the statistics are updated along time. In almost all cases, after 60% of the sequence is achieved stable and coherent results.

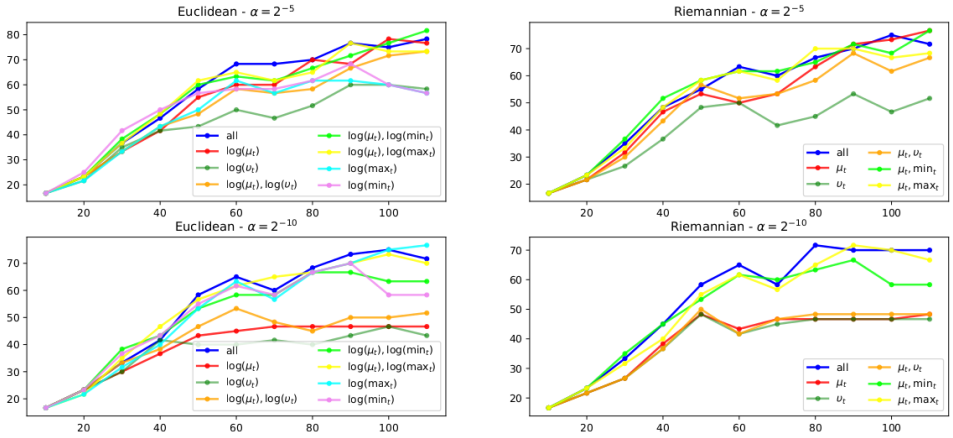

Fig. 6 UT1 online action recognition performance. The best performance is observed over statistics computed on Euclidean space by projecting incoming covariance to identity tangent plane. In almost all cases, a limitation on action representation is reported when the variance statistic is included on action descriptor. Projecting from on manifold results interesting as it improves the stability after 65% of the video sequence.

achieve a good trade-off between accuracy and speed of computation. The support trajectories contain by themselves relevant kinematic information that result robust to appearance changes, allow a natural fusion of different features that can compactly represent activities. These KTFs are spatially pooled into covariance matrices that lie within a video Riemannian manifold.

This action covariance-based descriptor represents each particular sequence as Riemannian manifold, whose statistics approximate topology and geometry of actions in such space. In this work we analyze intrinsic statistics to operate on the video Riemannian manifold, which are treated as an optimization problem by projecting tangent planes and operating in such corresponding Euclidean space. From such statistics we achieve competitive action classification results, by coding complete video sequences with video descriptors of only 78 scalar values. In many cases, the intrinsic Riemannian mean results sufficient to model an individual activity recorded in a video sequence. For more complex cases, an extension of such statistics was herein proposed to build video descriptors on the fly, using a recurrent scheme, where each statistic is updated and mapped to a machine learning algorithm to obtain a frame level prediction. In all cases, the computed recursive statistics result compact and relevant to represent actions over the evaluated datasets.

For KTH, where the actions are periodic and with a relatively controlled background, the video descriptor only requires 40% of video sequence to achieve stable prediction results. Regarding UT interaction, it was obtained a progressive accuracy on partial representations with stable results at 70% of the sequences. In almost all experiments, the mean and the recursive mean result the most descriptive measures to code activities. Such measures generally regularize covariances and filter out abrupt changes during the development of the activity. On the other hand, the variance often showed limited representation power for activities.

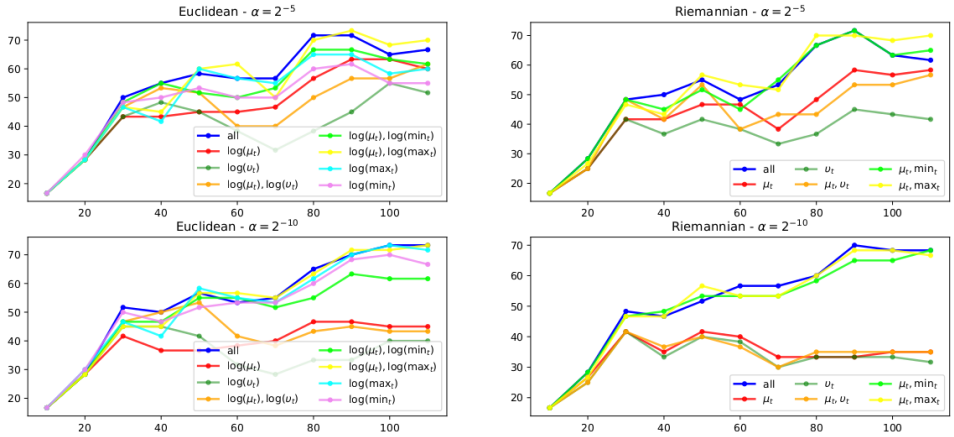

Fig. 7 UT2 online action recognition performance. In both projections it is illustrated a similar maximum performance. The variation of α parameter result important to obtain a smooth or stepped incremental accuracy. In almost whole case, the integration with variance limit the description of activities. Such fact could be related with abrupt motion of cameras and background.

Nevertheless, this statistic is helpful in the case of quick changes and strong variations on the dynamic of the activities. Representing videos by covariance Riemannian manifolds proves a powerful tool to compactly describe actions. In both cases, action classification and recognition, competitive results were achieved from very compact representations. Such representation could be complemented with additional image features or included in more sophisticated learning representations. Future works include evaluation of the proposed descriptor in more complex scenarios. Likewise, this descriptor will be extended to recognition of interactive actions such as human group activities.

Acknowledgements This research is funded by the RTRA Digiteo project MAPOCA.

References

1. Cao, X.t.: Action recognition using 3d daisy descriptor. Machine vision and applications 25(1), 159–171 (2014)
2. Fernando, B.t.: Modeling video evolution for action recognition. In: CVPR, pp. 5378–5387 (2015)
3. Fletcher, P.T., Joshi, S.: Principal geodesic analysis on symmetric spaces: Statistics of diffusion tensors. In: Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis, pp. 87–98 (2004)
4. Fletcher, P.T., Joshi, S.: Riemannian geometry for the statistical analysis of diffusion tensor data. Signal Processing 87(2), 250–262 (2007)
5. Gaidon, A., Harchaoui, Z., Schmid, C.: Actom sequence models for efficient action detection. In: CVPR 2011, pp.3201–3208. IEEE (2011)
6. Guo, K., Ishwar, P., Konrad, J.: Action recognition from video using feature covariance matrices. IEEE Transactions on Image Processing 22(6), 2479–2494 (2013)
7. Jain, M., Jegou, H., Bouthemy, P.: Better exploiting motion for better action recognition. In: CVPR, pp. 2555–2562 (2013)
8. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence 35(1),221–231 (2012)
9. Ji, X., Wang, C., Zuo, X., Wang, Y.: Multiple feature voting based human interaction recognition. International Journal of Signal Processing, Image Processing and Pattern Recognition 9(1), 323–334 (2016)
10. Jiang, Y.G., Dai, Q., Liu, W., Xue, X., Ngo, C.W.: Human action recognition in unconstrained videos by explicit motion modeling. IEEE Transactions on Image Processing 24(11), 3781–3795 (2015)
11. Laptev, I.: On space-time interest points. International journal of computer vision 64(2-3), 107–123 (2005)
12. Laptev, I., Marsza lek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR (2008)
13. Pennec, X.: Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. Journal of Mathematical Imaging and Vision 25(1), 127 (2006)
14. Pennec, X., Fillard, P., Ayache, N.: A riemannian framework for tensor computing. International Journal of computer vision 66(1), 41–66 (2006)
15. Ryoo, M.t.: An overview of contest on semantic description of human activities (sdha) 2010. In: ICPR, pp. 270–285. Springer (2010)
16. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: ICPR 2004., vol. 3, pp.32–36 (2004)
17. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems, pp. 568–576(2014)
18. Nour el houda Slimani, K., Benezeth, Y., Souami, F.: Human interaction recognition based on the co-occurrence of visual words. In: CVPR, pp. 455–460 (2014)
19. Tran, D.t.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR, pp. 6450–6459 (2018)
20. Varol, G., Laptev, I., Schmid, C.: Long-term temporal convolutions for action recognition. IEEE transactions on pattern analysis and machine intelligence 40(6), 1510–1517 (2018)
21. Veeriah, V., Zhuang, N., Qi, G.J.: Differential recurrent neural networks for action recognition. In: Proceedings of the IEEE international conference on computer vision, pp. 4041–4049 (2015)
22. Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behavior understanding in video surveillance. The Visual Computer 29(10), 983–1009 (2013)
23. Vrigkas, M., Nikou, C., Kakadiaris, I.A.: A review of human activity recognition methods. Frontiers in Robotics and AI 2, 28 (2015)
24. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: CVPR, pp. 3551–3558 (2013)
25. Wang, H.t.: Action recognition by dense trajectories. In: CVPR (2011)
26. Wang, L., Qiao, Y., Tang, X.: Action recognition with trajectory-pooled deep-convolutional descriptors. In: CVPR, pp. 4305–4314 (2015)
27. Wu, S., Oreifej, O., Shah, M.: Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories. In: CVPR, pp. 1419–1426. IEEE (2011)
28. Yu, G., Yuan, J., Liu, Z.: Propagative hough voting for human activity recognition. In: ECCV 2012, vol. 7574, pp. 693–706 (2012)