Human activity recognition in the semantic simplex of elementary actions

Beaudry Cyrille
cyrille.beaudry@univ-lr.fr
Péteri Renaud
renaud.peteri@univ-lr.fr
Mascarilla Laurent
laurent.mascarilla@univ-lr.fr

MIA
University of La Rochelle
La Rochelle, France

Abstract

This paper presents an original approach for recognizing human activities in video sequences. A human activity is seen as a temporal sequence of elementary action probabilities. Actions are first generically learned using a robust action recognition method based on optical flow estimation and a cross-dataset training process. Activities are then projected as trajectories on the semantic simplex in order to be characterized and discriminated. A new trajectory attribute based on the total curvature Fourier descriptor is introduced. This attribute takes into account the induced geometry of the simplex manifold. Experiments on labelled datasets of human activities prove the efficiency of the proposed method for discriminating complex actions.

1 Introduction

1.1 Context

Analyzing and recognizing actions has received considerable attention for many years in the computer vision community. Works on this topic are motivated by several potential applications (video monitoring, automatic video indexing, crowd analysis, human-machine interaction, etc). The wide variability of human actions in videos makes difficult to design generic methods (datasets of sport activities, daily activities, different contexts of action, etc). Two kinds of approaches for tackling human action recognition can be outlined. The first approaches tend to consider an action as a set of low-level features extracted from a group of frames (for instance histograms of spatio-temporal points). This constitutes what we call an elementary action, such as walking or jumping. The second approaches represent an order set of semantic attributes, and is called an activity. This is the framework of the proposed method: human activities are complex actions which are constituted of an ordered set of different elementary actions. For instance high jumping, can be decomposed of different elementary actions over time: walking, running and jumping.
1.2 Human activities: a brief state of the art

Different approaches have been developed to address human action recognition. Most of them are based on discriminative supervised models. The goal is to discriminate different actions performed by one or several subjects using algorithmic methods trained on already labeled video sequences. To detect and describe relevant features in videos, several discriminative approaches are using a temporal extension of 2D interest point detector. Laptev et al. [3] were the first to propose the Spatio-Temporal Interest Point detector (STIP) which is a temporal extension of the Harris-Laplace 2D detector [1]. It is efficient on constrained video datasets such as KTH Dataset [16]. Dollar et al. [3] provide the cuboïd detector and descriptor adapted for periodic movements in video, or for facial expression recognition. In [23] Willem et al. extend the 2D detector SURF [3] in the temporal domain to detect saliency using the determinant of 3D Hessian matrix. Wang et al. in [21] are adding temporal information by estimating point trajectories, using a dense sampling strategy at regular time intervals. Trajectories allow to better capture temporal information of motion of interest. Raptis et al. also use gradient and optical flow information to encode salient point trajectories. Vrigkas et al. [20] represent actions using a Gaussian mixture model by clustering motion curves computed from optical flow estimation.

Other studies are focused on generative probabilistic models for human activities or complex actions. Unlike elementary actions, activities require a much longer temporal observation. They commonly represent human daily behavior, sports action, humans interaction and most of them can be decomposed into different short elementary actions. Activities have a higher semantic level compared to elementary actions. Most generative models are based on Latent Dirichlet Allocation algorithm (LDA) [3] originating from document retrieval. This algorithm brings out underlying document topics. This framework allows the characterization of any type of data as a proportion of topics which compose it. A complex action is then defined as a collection of topics. A SVM classifier is generally applied on the collection of topics to discriminate between activities. Niebles et al. explore topic generation for human activities in [12] as a non-supervised action learning using Bag of visual Words. The BoW is built using features such as the cuboïd descriptor [1]. Tavernard et al. represent videos as sequential occurrences of visual words obtained from STIP detector. Hierarchical LDA is thereafter used to take into account the chronological structure of visual words. Y. Wang et al. have introduced semi-supervised LDA to constrain correspondance between generated topics and already known action classes. Nevertheless, generative models such as LDA fail to match already known actions occuring in videos with topics generated in a non-supervised way. Moreover, it is difficult to semantically analyze discovered topics. In fact, in the original version of the LDA, there is no possibility to bring an a priori information on already known actions and to ensure a correspondance between generated topics and present actions in the video. Methods using a priori information are less efficient than discriminative methods on classification of elementary actions. Moreover, most of them are using global descriptors which have shown their limitation for action recognition.

In this paper, we present an original approach for human activities recognition in videos. It relies on a semantic representation of videos rather than a Bag of visual features approach, allowing better generalization. We characterize activities as temporal sequences of elementary actions by estimating their probabilities over time. Elementary actions are not discovered as in generative probabilistic models but learned via a robust action recognition method based on a discriminative model. These activities are then projected as trajectories on the semantic simplex of elementary actions. These action trajectories are processed and characterized...
using a metric respecting the geometry of the simplex manifold.

## 2 Recognition of elementary actions

### 2.1 A method based on optical flow estimation

To recognize elementary actions, a discriminative approach based on the optical flow estimation provided by Beaudry *et al.* \cite{beaudry2016} is used. Video sequences are characterized by critical points of optical flow and by their temporal trajectories. These features are computed at different spatio-temporal scales, using a dyadic subdivision of the sequence. Features are extracted and correspond to different frequency scales (fast and slow movements, respectively at high and small scales - illustrated on Figure 1).

![Figure 1: Example of movements captured at different frequency scales. Red trajectories correspond to high frequency movements (fast). Blue trajectories correspond to low frequency movements (slow). Green trajectories correspond to an intermediate frequency scale.](image)

Critical points are locally described by spatial gradients and motion orientation descriptors \cite{image} (namely the HOG and HOF descriptors). Multi-scale trajectories are thereafter frequency described using Fourier transform coefficients, which ensures robust invariance to geometric transformations. These three characteristics (shape, orientation of movement and frequency) have proven to be complementary and relevant for elementary action recognition.

### 2.2 Cross-dataset learning

In order to characterize an activity as an ordered sequence of elementary actions, a robust and generic representation of the elementary action has to be extracted. Here, we have selected the KTH dataset \cite{image}, the Weizmann dataset \cite{image}, the UCF-11 dataset \cite{image} and the UCF-50 dataset \cite{image} to represent different aspects of human actions. KTH and Weizmann are datasets containing videos with acquisition constraints. Subjects in those videos are performing elementary actions (*jumping, waving, walking, running, boxing, etc*) in a canonical way. In UCF-11 and UCF-50, elementary actions are captured in real situations and contexts. These latter bring visual variabilities and movements which are not specific to the action of interest. Generic and constrained datasets are complementary for representing elementary actions at different frequencies of movement. We combine videos from both
types of datasets to perform a cross-dataset learning in order to provide a robust and generic
description of elementary actions.

Four elementary actions from common activities are thereafter considered: Jump, Run, Walk, and Handwave. In the cross-dataset learning process, we have made the choice of using 1/3 of generic videos (UCF-11, UCF-50) and 2/3 of constrained videos (KTH and Weizmann datasets).

Table 1 shows results obtained after a Leave-One-Out cross-validation test when the classifier is trained on this mixture dataset. The global recognition rate is 96.87%. Confusion appears between semantically related classes. We have used an Adaboost late fusion scheme to combine each feature descriptor.

<table>
<thead>
<tr>
<th>Actions</th>
<th>jump</th>
<th>walk</th>
<th>run</th>
<th>wave</th>
</tr>
</thead>
<tbody>
<tr>
<td>jump</td>
<td>90.62</td>
<td>0</td>
<td>9.37</td>
<td>0</td>
</tr>
<tr>
<td>walk</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>run</td>
<td>0</td>
<td>3.12</td>
<td>96.87</td>
<td>0</td>
</tr>
<tr>
<td>wave</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 1: Confusion matrix of the mixture dataset learning

The recognition rate per descriptor and Adaboost weight is shown in Table 2. The HOG descriptor computed weight is the lowest among these three descriptors, as expected with a hybrid dataset. Efros et al. have indeed shown that most of common datasets have an important visual bias [18].

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>FCD</th>
<th>HOF</th>
<th>HOG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rec. rate</td>
<td>94.53%</td>
<td>95.31%</td>
<td>53.12%</td>
</tr>
<tr>
<td>Adaboost weight</td>
<td>2.91</td>
<td>1.82</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Table 2: Recognition per descriptor and weight obtained by Adaboost late fusion.

Mixing videos from different datasets increases the visual variability in the scene and weaken gradient information whereas information related to motion remains quite stable.

3 Representation of complex actions by a sequence of elementary actions

3.1 Probability of elementary actions

In our method, a complex action is viewed as a sequence of elementary action proportions. To evaluate these proportions evolution in time, the decision boundaries of the classifier are transformed into probabilities [5].

Elementary action probabilities are then computed over time using a sliding window along the video sequence. The goal is to characterize a frame \( t \) by its elementary action probabilities in a \([t - N; t + N]\) window. It is assumed that elementary actions are commonly performed on a short time period. K.Schindler et al. [15] have indeed shown that few images are needed to achieve good recognition rates on elementary action datasets. Figure 2 illustrates the decomposition on a video from the Weizmann dataset. The Jack action is composed of Jump and Handwave elementary actions. The graph represents the evolution of elementary action probabilities over time. Red curve is for Handwave action, blue curve for Jump action. The periodicity and alternation between the two elementary actions is well
noticeable on the graph. The other figure represents a Basket-ball action, where the green bar indicates the "handwaving" instant of the shoot, which corresponds to a high proportion of the Handwave elementary action. The blue bar emphasizes the "jump" instant of the shoot, where Jump elementary action probability is high. These examples illustrate the ability of the action recognition method to provide a meaningful representation of generic actions using a mixture dataset.

### 3.2 Characterizing the absence of action

When elementary action probabilities are estimated, values depend on the most relevant actions among those that have been learned. When there is no movement in the sequence, it is necessary to adapt the classifier. To do so, the amount of potential movements present in the sliding window is estimated using the optical flow. Each frame of the sequence is subdivided in vertical and horizontal blocks and the mean power of the optical flow is computed. Finally, each frame $t$ of the sequence is characterized by a coefficient $\text{coef}_{\text{standing}}(t) \in [0, 1]$,
reflecting this absence of motion. Probabilities estimated from the classifier are then normalized. An artificially generated action class is then introduced, named the "Standing class". This class allows to inject in the classifier the presence or absence of movement in the sequence at time \( t \).

The new \textit{a posteriori} probability vector is then:

\[
\text{Prob}_{\text{estimates}}(t) = [\text{coef}_{\text{standing}}(t) \ast (\lambda_1(t), \ldots, \lambda_k(t), \ldots, \lambda_L(t)), 1 - \text{coef}_{\text{standing}}(t)]
\]

with \( \lambda_k(t) \) the probability of the elementary action at time \( t \).

Figure 3 shows an example of improvement obtained using the \textit{Standing} class (a related video is also available online in the supplementary materials). When no action is present in the sequence, probabilities are close to 0, except for the new standing class. The characterization of the absence of movement provides a richer description and a more relevant representation of the elementary actions over time.

4 Action semantic trajectories

4.1 Trajectory in the semantic space

Once a frame is characterized by its elementary action probabilities, its feature vector lies in a simplex \( \mathcal{P}_L \) defined such as:

\[
\mathcal{P}_L = \{ \pi \in \mathbb{R}^{L+1} \mid \sum_{i=1}^{L+1} \pi_i = 1, \pi > 0 \}, \quad \mathcal{P}_L \text{ being a submanifold of } \mathbb{R}^{L+1}.
\]

Figure 4 shows the global scheme for projecting activities in \( \mathcal{P}_L \).

![Figure 4: Global scheme for characterizing activities in the semantic simplex.](image)

Activities are then represented as trajectories on the semantic simplex \( \mathcal{P}_L \).

The transformation:

\[
F : \left\{ \mathcal{P}_L \rightarrow S_L^+ \right\} \pi = (\pi_1, \ldots, \pi_{L+1}) \rightarrow \theta = (2 \sqrt{\pi_1}, \ldots, 2 \sqrt{\pi_{L+1}})
\]

with:

\[
S_L^+ = \{ \theta \in \mathbb{R}^{L+1} \mid \sum_{i=1}^{L+1} \theta_i^2 = 2, \theta > 0 \},
\]

is a diffeomorphism of \( P_L \) into \( S_L^+ \). The \( L \)-simplex \( P_L \) is endowed with the Fisher information metric and the positive \( L \)-hypersphere \( S_L^+ \) is endowed with the standard Euclidean metric on its surface [8].

Since \( F \) is an isometry, it emplies that geodesic distances in \( P_L \) can be computed as shortest curves on \( S_L^+ \).
The geodesic distance between two points \((\pi_{k_1}, \pi_{k_2})\) of \(P_L\) is the great circle arc linking \((F(\pi_{k_1}), F(\pi_{k_2}))\) on \(S^+_L\) such as:

\[
d_{S^+_L}(F(\pi_{k_1}), F(\pi_{k_2})) = d_{S^+_L}(\theta_{k_1}, \theta_{k_2}) = 2 \cos^{-1}(\theta_{k_1} \theta_{k_2}^\top / 4)
\]

4.2 Characterization of semantic trajectories through total curvature Fourier descriptor

Trajectories are characterized by their shapes on the manifold. By using the diffeomorphism \(F\), trajectories lie on the hypersphere \(S^+_L\), and cartesian coordinates in \(R^{L+1}\), are converted into spherical coordinates. They are defined by one radial coordinate \(r\) (in our case \(r = 2\)) and \(L\) angular coordinates \(\phi_1, \phi_2, ..., \phi_L \in [0, 2\pi]\). The goal is to describe in the frequency domain the angular evolution over time of the shape on the half-positive hypersphere.

Fourier coefficients permit to obtain a robust shape descriptor. In the frequency domain, most of the shape information are included in the first low frequencies. One obtains a robust and global representation of trajectories by discarding high frequencies which correspond to less relevant information or noise. Considering only angular variations of the spherical coordinates ensure that any processing in the frequency domain will keep the resulting trajectory on \(S^+_L\) (thus the sum of associated probabilities equals to 1).

Activity trajectories are open shapes in \(S^+_L\). Reconstruction of shapes from low-frequency Fourier coefficients do not necessarily coincide with the end-points of the original shape. When removing high frequencies, it has a tendency to become a close shape and to oscillate near end-points. To avoid this problem, the method of U.Yoshinori et al. \[19\] for open curves is adapted. It corresponds to the Fourier transform performed on a cumulative angular curvature function. It permits to preserve end-point positions of the original shape when it is reconstructed from only low frequencies of its Fourier descriptor (see Figure 5).

Figure 5: Trajectories having the same shape but different positions in the simplex. First trajectory (top row) goes from Wave action to Run action. Second trajectory (bottom row) goes from Wave action to Jump action. Because of the concatenation of angular coordinates \(\phi\), the two resulting descriptors are different.

To characterize these trajectories, we concatenate the Fourier transform coefficients of the cumulative angular curvature of each trajectory angular coordinates. Because of the geometry of the hypersphere, the resulting descriptor is not invariant to translations, scales and rotations (invariance would not be here a desirable property). In fact, the position on
the simplex depends on the actions performed during the activity. Two different activities which have trajectories with the same shape but do not share necessarily the same elementary actions. They will be on two different positions on $S^+_L$ (see Figure 6).

![Figure 6: Trajectory smoothing on $S^+_L$. Left: original trajectory. Center: simplified reconstructed trajectory using standard Fourier descriptor. Right: simplified reconstructed trajectory using total curvature Fourier descriptor. Start-point and end-points keep the same position when using the total curvature function.](image)

5 Experiments

In order to test the discriminative performance of the proposed method, three complex actions from UCF datasets and Olympic Sport dataset are considered: High-Jump, Basket-ball and Base-ball. Four elementary actions are used: Wave, Jump, Run and Walk (setting $L = 5$ for taking into account the Standing class). We set the temporal window size to $N = 6$, and 10 videos for each class are considered.

The first step consists in up-sampling each trajectory to obtain the same number of points for all of them. The Fourier cumulative curvature descriptor is then used to characterize each trajectory. Only 50% of the Fourier coefficients is kept in the final the descriptor. Figure 7 illustrates trajectories of videos from each activity class (3 related videos are also available online in supplementary materials).

![Figure 7: Examples of activity trajectories: High-Jump activity, Base-ball activity, and Basket-ball activity. Please see related videos in online supplementary materials.](image)
5.1 Classification results

We perform a Leave-One-Out cross-validation test using a SVM with a RBF kernel. The goal is to evaluate the performance of the method and to assess the interest of characterizing trajectories in the simplex with the Fourier-shape descriptor. A recognition rate of 96.6% is obtained. Table 3 exposes results of the cross-validation test for each activity class. It illustrates the fact that Fourier shape descriptor permits a good characterization of activity trajectories on the semantic hypersphere. Trajectories of the same activities commonly share the same shape and the same sequence of elementary actions. Describing trajectories with Fourier coefficients permits a robust characterization of the shape in the frequency domain. Moreover, the fact that we use spherical coordinates permit to encode in the descriptor information about elementary actions variation over time, which is a crucial point here.

<table>
<thead>
<tr>
<th>Activities</th>
<th>High jump</th>
<th>Basketball</th>
<th>BaseBall</th>
</tr>
</thead>
<tbody>
<tr>
<td>High jump</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Basketball</td>
<td>0</td>
<td>100%</td>
<td>0%</td>
</tr>
<tr>
<td>Baseball</td>
<td>10%</td>
<td>0</td>
<td>90%</td>
</tr>
</tbody>
</table>

Table 3: Confusion matrix when a SVM classifier is trained on activity trajectory features from our method. (High-Jump, Basketball and Baseball). Recognition rate is 96.6%.

The recognition rate obtained with a Leave-one-out cross-validation test emphasizes the discriminative power of this representation. Considering trajectories on the simplex allows to take into account the temporal order of elementary actions for each activity class. In comparison, we have applied the STIP method provided by [10] on videos used in our experiments. The confusion matrix is presented in Table 4. The global recognition rate for the STIP on this set is 86.6%, to be compared with the 96.6% reached by the proposed method. The semantic aspect of our method also allows a better generalization of human activities.

<table>
<thead>
<tr>
<th>Activities</th>
<th>High jump</th>
<th>Basketball</th>
<th>BaseBall</th>
</tr>
</thead>
<tbody>
<tr>
<td>High jump</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Basketball</td>
<td>0</td>
<td>80%</td>
<td>20%</td>
</tr>
<tr>
<td>Baseball</td>
<td>0%</td>
<td>20%</td>
<td>80%</td>
</tr>
</tbody>
</table>

Table 4: Confusion matrix when a SVM classifier is trained with the STIP method from Laptev [10]. Recognition rate is 86.6%.

6 Conclusion

We develop in this paper an original approach for human activity recognition. The method characterizes human activities as a sequence over time of elementary actions probabilities. These sequences are then projected as trajectories on a semantic simplex to be characterized using the total curvature Fourier descriptor. The frequency domain allows to encode shape information on trajectories and permits to discriminate between different human activity classes. Unlike generative probabilistic model, the elementary actions which compose human activities are statistically learned with a robust action recognition method trained on a cross-dataset.

Considering human actions as trajectories on the semantic manifold opens the way to different applications, such as video summary (computation of a mean shape on the manifold).
References


