Human Action Recognition from Inter-Temporal Dictionaries of Key-Sequences

Analí Alfaro, Domingo Mery, and Alvaro Soto

Department of Computer Science
Pontificia Universidad Católica de Chile
ajalfaro@uc.cl,
{dmery, asoto}@ing.puc.cl

Abstract. This paper addresses the human action recognition in video by proposing a method based on three main processing steps. First, we tackle problems related to intraclass variations and differences in video lengths. We achieve this by reducing an input video to a set of key-sequences that represent atomic meaningful acts of each action class. Second, we use sparse coding techniques to learn a representation for each key-sequence. We then join these representations still preserving information about temporal relationships. We believe that this is a key step of our approach because it provides not only a suitable shared representation to characterize atomic acts, but it also encodes global temporal consistency among these acts. Accordingly, we call this representation inter-temporal acts descriptor. Third, we use this representation and sparse coding techniques to classify new videos. Finally, we show that, our approach outperforms several state-of-the-art methods when is tested using common benchmarks.

Keywords: human action recognition, key-sequences, sparse coding, inter-temporal acts descriptor.

1 Introduction

Human action recognition is relevant to the development of potential applications such as surveillance systems, human-computer interaction and video annotation. However, there are many challenges that deserve careful attention. These include *i)* the fact that features should be reliable so that researchers can develop programs that achieve adequate representation of human actions; *ii)* the existence of high inner-class variations such as human poses, occlusions, viewpoints and dynamic backgrounds, which stand as obstacles to the task of classification; and *iii)* variations in the duration of an action, which can prevent quick recognition. Many studies have tried to address these problems over the past few years. Approaches to action representation use both global and local representations. Global representations consider holistic features such as silhouettes, motion and volumes. Bobick and Davis [3] and Efros et al. [6] proposed to describe the motion inside of a volume (stack of person-centered frames). Another methods

represent silhouettes as three-dimensional shapes in order to extract features [9]. Fathi and Mori [7] argue that low-level features are uninformative and propose the use of mid-level ones. In general, global features fail to fully address the noise introduced by clothes, lighting and body deformations. A local representation may be more robust than a global one in situations of high variation. The emergence of methods that can be useful for finding spatio-temporal interest points [5, 14] have allowed researchers to address the aforementioned problems. Many researchers have used dictionaries based on local features to develop robust action recognition systems. Several approaches can be used to produce a dictionary, including Bag-of-Words [16, 8, 15], Random Forest [25, 26], and more recently Sparse Dictionary Learning [10, 11, 23]. The sparse analysis establishes that a natural signal (eg. images) can be broken down into a linear combination of atoms which form a dictionary. Sparse coding techniques can be used to model an action based on the combination of training samples. The dictionary is adapted in order to capture the inherent structure of the data. For example, Guo et al. [11] built a class-specific dictionary from silhouette-based descriptors. The reconstruction error of each class is used for classification. Tanaya and Ward [10] proposed local motion pattern descriptors and evaluated several classification methods using sparse coding. Tran et al. [23] used a body-part descriptor to develop a dictionary for each part and class, and classified them based on the reconstruction error. These approaches have a common aspect: they use low-level features to find dictionaries. However, this information may be similar between different actions and it may incur into misclassification.

On the other hand, there are methods that use a single frame to achieve recognition [4]. Still others use a set of frames [8, 10, 11, 18, 21]. These approaches tend to manually select short sequences containing the atomic information of the action in order to address the time variation problem. Unlike of these methods we argue that the key-sequences selected to represent a video should be carefully selected. This is an important aspect to ensure an appropriate video representation. Additionally, we claim that it is necessary to rescue the inter-temporal relationship among which key-sequences.

The main focus of our work is to break down an action into acts such that the interaction among these parts summarizes the whole. The video of an action can thus be represented as a set of *key-sequences* containing the acts and their temporal relationships. By articulating both aspects, we can arrive at a complete description of the action. We propose breaking a video down into key-sequences using a non-supervised procedure. The key-sequences are then described using a descriptor obtained from Sparse Dictionary Learning. Our findings prove that our approach is effective and that it outperforms the state of the art on several datasets.

The main intuitions and contributions of this paper can be summarized as follows:

- We demonstrate the relevance of exploring the subsequence feature space in order to identify key atomic acts, or key-sequences, that can be unique or

shared among action classes. In particular, we propose a method where each video is represented by a set of meaningful acts or key-sequences.
– We noted that human action classes can be characterized as a composition of unique and shared atomic acts. Furthermore, the temporal ordering of these atomic acts provides highly discriminative information to recognize different human actions. We then propose a new representation, called an *inter-temporal acts descriptor*, that is able to encode local information to characterize atomic acts and also to preserve a global temporal ordering among them.

The rest of the paper is organized as follows. Section 2 describes the proposed method for describing a video using the called an *inter-temporal acts descriptor* and the classification strategy. Section 3 presents the experiments and the discussion of our results. Finally, Section 4 contains concluding remarks.

2 Method

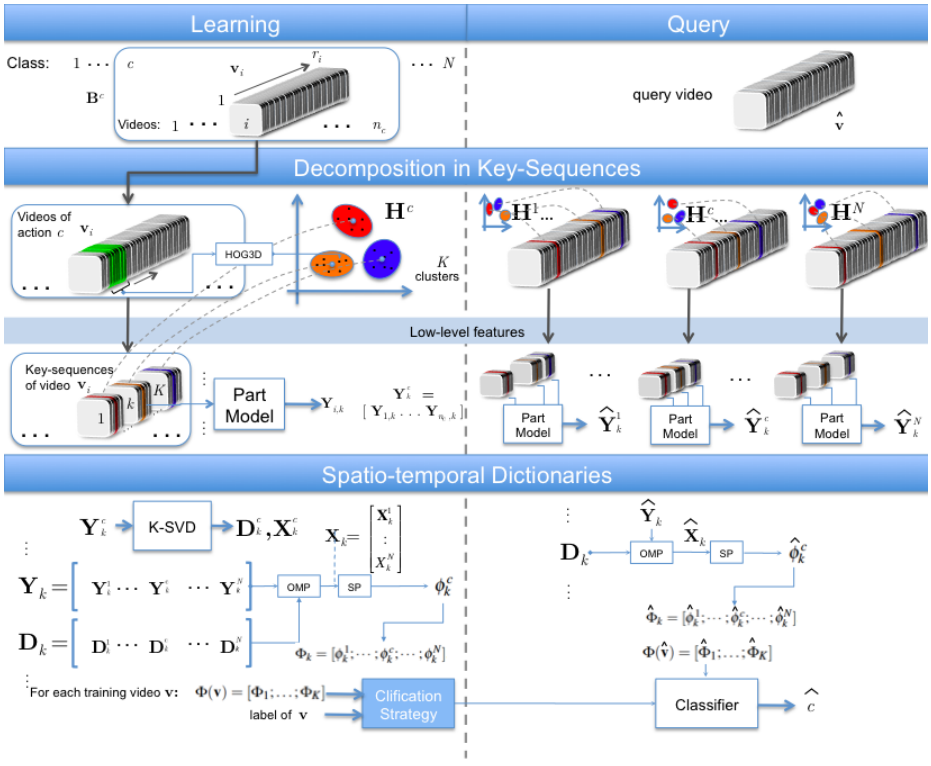

Fig. 1. Block diagram of proposed method.

In this section, we describe the proposed method following Fig. 1. The method consists of three main steps: *i)* Video decomposition in key-sequences, *ii)* Learning of spatio-temporal dictionaries of key-sequences, and *iii)* The classification strategy.

2.1 Video decomposition in Key-Sequences

In order to reduce the dimensionality of the recognition problem, we propose summarizing a video sequence using a few representative *key-sequences*. A key-sequence is a small number of consecutive frames composed of acts (or atomic motions) of an action that can be used to recognize it. For instance, the action 'diving' can be split into three acts: 'jump', 'twist' and 'entry'. These acts should contain enough information to address a successful classification of the action. The key-sequences should satisfy the following *consistency criteria*: *i)* They should strongly characterize the class of an action, which means that an action can be visually recognized by observing its key-sequences. *ii)* The key-sequences should be temporally sorted, thus, the k-th key-sequence of a video of an action should be similar to the k-th key-sequence of another video of the same action. Note that each person has his or her own way for doing an action, which means that the acts may appear to be different. Nevertheless, the acts should be consistent for all videos belonging to the same class.

The procedure for obtaining the key-sequences of a class is repeated for each class c, for $c = 1 \dots N$. Let $\mathbf{B}^c = \{\mathbf{v}_i\}_{i=1}^{p}$ be a set of p training videos of class c, where $\mathbf{v}_i = \{\mathbf{f}_{i,j}\}_{j=1}^{r_i}$ is a set of r_i frames. For each video $\mathbf{v}_i$, we built a set $\mathbf{S}_i$ with all of the possible subsequences $\mathbf{s}_{i,j} = \{\mathbf{f}_{i,k}\}_{k=j}^{j+t-1}$ of t consecutive frames: $\mathbf{S}_i = \{\mathbf{s}_{i,j}\}_{j=1}^{r_i-t+1}$. Thus, the set of subsequences of the class is $\mathbf{S}^c = \{\mathbf{S}_i\}_{i=1}^{p}$.

Each subsequence in $\mathbf{S}^c$ is described in appearance and motion using the well-known HOG3D descriptor [13]. It generates a feature space $\mathbf{H}^c = f_{\text{HOG3D}}(\mathbf{S}^c)$ with a large collection of high-dimensional spatio-temporal descriptors. Our goal is to find groups of descriptors with high levels of similarity that summarize the overall behavior from the class. We find these groups by applying a K-means clustering algorithm over $\mathbf{H}^c$, and define $\mathbf{Z}^c = \{\mathbf{z}_k\}_{k=1}^{K}$ as the set of the K estimated centroids for class c. We expect each centroid to represent an act which should be present in every video. Thus, we define the K key-sequences of video $\mathbf{v}_i$ as those K subsequences $\{\hat{\mathbf{s}}_{i,k}\}_{k=1}^{K} \in \mathbf{S}_i$, where $\hat{\mathbf{s}}_{i,k} = \mathbf{s}_{i,q(i,k)}$ is a subsequence, and where description $\mathbf{h}_{i,q} = f_{\text{HOG3D}}(\mathbf{s}_{i,q})$ is the most similar one to each centroid of $\mathbf{Z}^c$ and $q(i,1) < q(i,2) \dots < q(i,K)$ in order to ensure the temporally sorted subsequences. The key-sequences are estimated as follows: First, we compute the indices of the most similar subsequences as $w_k = argmin_{q'} \|\mathbf{z}_k - \mathbf{h}_{i,q'}\|$. Second, we sort the indices w as $\{q(i,k)\}_{k=1}^{K} = \text{sort}(\{w_k\}_{k=1}^{K})$ (see Fig. 2).

Low-Level Feature Extraction: A sequence of the video $\mathbf{v}_i$ could contain not only the actor, but also some noise from dynamic backgrounds, the objects intervening in the scene, clothes, etc. In order to overcome this problem, the actor is detected using a model of body part detection applied to the key-sequences of the video [2]. The goal is to extract relevant information from limbs where

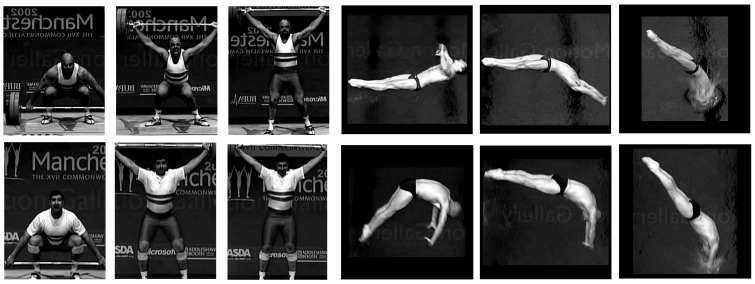

Fig. 2. Central frame from key-sequences of lifting and diving actions(UCF-Sports).

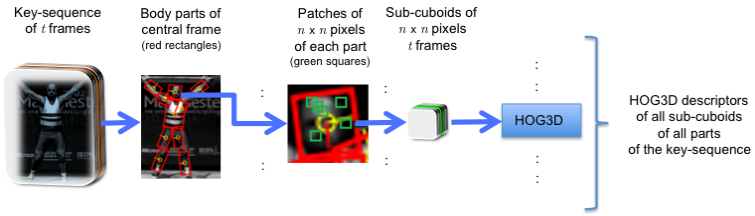

Fig. 3. Part model: low-level features of a key-sequence.

motions are performed. Thus, given a key-sequence of a video $\hat{\mathbf{s}}_{i,k}$ with t frames, we apply the body part model only to the central frame $(t-1)/2$, for an odd t, because it is enough to produce an estimate of the location of the parts and to avoid the cost of making this in each frame. The model delivers the bounding box of ten parts of the body: torso, forearms, arms, thighs, legs and head (see Fig. 3).

We then extract local spatio-temporal features from spatio-temporal *cuboids* defined by the bounding box of the body parts propagated across the frames of the key-sequence. For each cuboid, we randomly generate spatial patches of $n \times n$ pixels and extract *sub-cuboids*. A sub-cuboid is formed by lining up the 2D patches from each frame of the key-sequence. The size of each spatio-temporal sub-cuboid is $n \times n \times t$. These sub-cuboids are described using HOG3D, which yields a collection of descriptors $\mathbf{Y}_{i,k}$, for video $\mathbf{v}_i$ and key-sequence k. We call this function $\mathbf{Y}_{i,k} = f_{\text{low-level}}(\mathbf{v}_i, k, c)$. Hereafter, the descriptors of the sub-cuboids will be denominated *low-level features*. The description of the k-th key-sequence of all n_c videos of class c are arranged in $\mathbf{Y}_k^c = [\mathbf{Y}_{1,k} \dots \mathbf{Y}_{ik} \dots \mathbf{Y}_{n_c,k}]$, and it will be called the low-level features from class c and temporal order k.

2.2 Learning of spatio-temporal dictionaries of key-sequences

The actions can be similar to each other or they can share features. This is true for the actions 'run' and 'long jump,' for example. When these actions are broken

down into their key-sequences we note the presence of the act of running in both actions. The local information extracted from this act is clearly similar in both actions. Therefore, we believe that it is possible to represent a video as a combination of characteristics from different actions. A high value of this combination in a given class indicates a high probability that a video belongs to that class. It follows that each key-sequence from a video can be encoded at a high level so that the temporal relationship between these descriptors can be identified. The temporal relationship is important for articulating the relationship between sequences. Our method computes a *inter-temporal acts descriptor* from temporal dictionaries. These features provide information about the contribution of all classes to the performance of an act in a key-sequence.

We use two levels of *sparse dictionaries* [1] to calculate the temporal dictionaries. At the lower level, we identify one dictionary for each class from low-level features. At the high level, we identify temporal dictionaries and their sparse representations. These sparse representations are used to build a feature that describes a video as a combination of the contributions of each class that considers the temporal relationship between their key-sequences.

Class-based Temporal Sparse Dictionary: In this level, a sparse dictionary of each class is created using the low-level features obtained in Section 2.1. The descriptors are organized into groups according to the ordering imposed by the key-sequences. Hence, we calculate a dictionary $\mathbf{D}_k^c$ for each class and each temporal order $\mathbf{Y}_k^c$ using K-SVD [1]:

$$\min_{\mathbf{D}_k^c, \mathbf{X}_k^c} ||\mathbf{Y}_k^c - \mathbf{D}_k^c \mathbf{X}_k^c||_F^2 \text{ subject to } ||x_l||_0 \leq \lambda_1, \tag{1}$$

where the last term indicates that the number of nonzero entries of each column of $\mathbf{X}_k^c$ must not exceed λ_1.

Concatenated Temporal Sparse Dictionary: In this level, all sparse dictionaries $\mathbf{D}_k^c$ $(c = 1 \dots N)$ with the same k-th order are concatenated to form a temporal dictionary . Let $\mathbf{Y}_k = [\mathbf{Y}_k^1 \parallel \dots \parallel \mathbf{Y}_k^c \parallel \dots \parallel \mathbf{Y}_k^N]$ be the matrix of the k-th order containing the concatenated low-level features from the N classes with the same temporal order position k. Let $\mathbf{D}_k = [\mathbf{D}_k^1 \parallel \dots \parallel \mathbf{D}_k^c \parallel \dots \parallel \mathbf{D}_k^N]$ be the concatenated temporal dictionary. The goal is to find $\mathbf{X}_k$, the sparse representation of $\mathbf{Y}_k$ by solving:

$$\min_{\mathbf{X}_k} ||\mathbf{Y}_k - \mathbf{D}_k \mathbf{X}_k||_F^2 \text{ subject to } ||x_l||_0 \leq \lambda_2, \tag{2}$$

using, for example, a matching pursuit algorithm like OMP [19]. This sparse representation is useful because it provides information about the contribution of the classes for each sample. Thus, $\mathbf{X}_k$ can be seen as $[\mathbf{X}_k^1 ; \dots \mathbf{X}_k^c ; \dots \mathbf{X}_k^N]$, where $\mathbf{X}_k^c$ represents the sparse matrix of class c in temporal order k. Therefore, we can take advantage of this information for computing the contribution of a certain class for a video.

Describing inter-temporal acts: Our purpose is to learn how to produce a representation which characterizes the videos as a combination of the classes and their temporal relationships. Let a video $\mathbf{v}$ represent a set of low-level features

$\{\mathbf{Y}_k^c\}_{c=1}^N$ for all classes for temporal order k, where $\mathbf{Y}_k^c = f_{\text{low-level}}(\mathbf{v}, k, c)$. The contribution of a given class and a given order to the video $\mathbf{v}$ is computed by applying a sum pooling (SP) in the atoms from its class. We define the sum pooling operator SP as:

$$\phi_k^c = \sum_{l=1}^{P} x_k^c(l), \tag{3}$$

where the $\mathbf{X}_k^c = [x_k^c(1); \ldots; x_k^c(P)]$ indicates the P atoms of the sparse representation of $\mathbf{Y}_k^c$ according to dictionary $\mathbf{D}_k$. Each low-level feature belongs to a given video $\mathbf{v}$ of class c and temporal order k. The $\sum(\cdot)$ term is the sum pooling (SP) generated by projecting on the P-atoms of the class. Finally, the sum pooling application produces a value ϕ_k^c that is interpreted as the contribution that is delivered by the class c to the video $\mathbf{v}$ in the temporal order k. We can thus represent a video $\mathbf{v}$ as the contribution of all classes to the composition of an *act descriptor* $\Phi_k = [\phi_k^1; \cdots; \phi_k^c; \cdots; \phi_k^N]$. This feature is important because allows us to express shared features between classes. Finally, in order to retrieve the temporal relationship between key-sequences of the video, we define a function to create an *inter-temporal acts descriptor*. The *inter-temporal acts descriptor* of a video $\mathbf{v}$ is defined by concatenating each *act descriptor* from each ordering as $\Phi(\mathbf{v}) = [\Phi_1; \ldots; \Phi_K]$.

2.3 Classification Strategy

The high-level features from the previous step can be used in any supervised classification approach. Again, we chose to use a classification based on sparse coding. Let β^c be a dictionary learned from the *inter-temporal acts descriptors* of the class c in the train stage. Let $\beta = [\beta^1 \ldots \beta^c \ldots \beta^N]$ be a concatenated overcomplete dictionary. Let $\Phi(\hat{\mathbf{v}})$ be the matrix containing the *inter-temporal acts descriptor* of the query video $\hat{\mathbf{v}}$. The classification consists of identifying a sparse representation called α for the *inter-temporal acts descriptor* of the test sample. This sparse representation is found using OMP [19]:

$$\min_{\alpha} ||\Phi(\hat{\mathbf{v}}) - \beta\alpha||_F^2 \text{ subject to } ||\alpha_l||_0 \leq \lambda_3 \tag{4}$$

Again the sparse representation α can be seen as a block matrix of the classes. Thus, it is possible to calculate the contribution of each class by applying the sum pooling operator SP. Finally, we choose $\hat{c}$, the class with the highest contribution.

3 Experiments and Results

In order to validate the method outlined in Section 2, in this section, we describe the experimental settings and report the results obtained in three well-known datasets of actions: KTH, UCF-Sports, and Olympic.

3.1 Experimental Settings

We configure the training and testing parameters for each dataset as follows:
Decomposition in Key-Sequences: We believe that a key-sequence should contain a short number of frames for two reasons. First, short sequences can capture atomic motion. Second, short sequences can be processed quickly. Then, we take a fixed group of $t = 7$ frames to form a sequence. In short videos, choosing many key-sequences may imply a high overlap. Specifically, the key-sequences of periodic actions can capture similar information because these actions are highly repetitive. For example, the action 'run' is a monotone action (arms and legs have a cyclic motion), and hence their key-sequences may be similar. A reasonable strategy is thus to select a few key-sequences in order to avoid providing the same information. We select a number of $K = 3$ key-sequences to represent a video. For each key-sequence and for each body-part, we select 5 random body part positions to select 5 patches of 20×20 pixels. Around each of these patches we select another 8 patches of the same size to complete a total of 45 patches to cover each of the 10 body-parts. Thus, each key-sequence produces 450 spatio-temporal sub-cuboids with size $20 \times 20 \times 7$. Each sub-cuboid is described using HOG3D (300 dimensions). In order to deal with uninformative sub-cuboids, we apply a threshold to the magnitude of the descriptor to filter them out. The values of the threshold are 1.8, 1 and 2 to UCF-Sports, KTH and Olympic dataset, respectively. Afterwards, the remaining descriptors are normalized.
Temporal Dictionaries: We create the class-based dictionaries using $P = 600$ atoms (*i.e.* with redundancy $\mu = 2$[1]). For each concatenated temporal dictionary, we set $\mu = 2$ and $P = \mu \times N \times K$. In particular, the number of atoms for KTH, Olympic and UCF-Sports are $36, 96$ and 54, respectively. The parameter of sparsity is set to the 10% of the number of atoms in both levels ($\lambda_1 = \lambda_2$) in accordance with [10]. Finally, the dimension of our *inter-temporal acts descriptor* is $N \times K$. Therefore, the dimension of the final descriptors in KTH, Olympic dataset and UCF-Sports are 36, 48 and 27, respectively.
Classification Strategy: We create a concatenated temporal dictionary for classification containing $P = \mu \times N$ atoms ($\mu = 2$) and set λ_3 to the 10% of the number of atoms. Thus, the number of atoms in KTH, Olympic dataset and UCF-Sports are 12, 32 and 18.

3.2 Results

• *KTH Dataset:* The dataset contains 2391 sequences from six types of human actions. Each action is performed by 25 actors in four different scenarios with view point changes and variations in scale, lighting and appearance (clothes). We use the original setup [22], *i.e.* splitting the data in training and testing. In order to engage in a fair comparison, we selected methods that use the same testing protocol. Table 1 shows a summary of the comparison. We obtained 95.7 % accuracy in the recognition compared to other state-of-the-art methods. Regarding

[1] The redundancy indicates the folds of basis vectors that are to be identified with respect to the dimension of the descriptor.

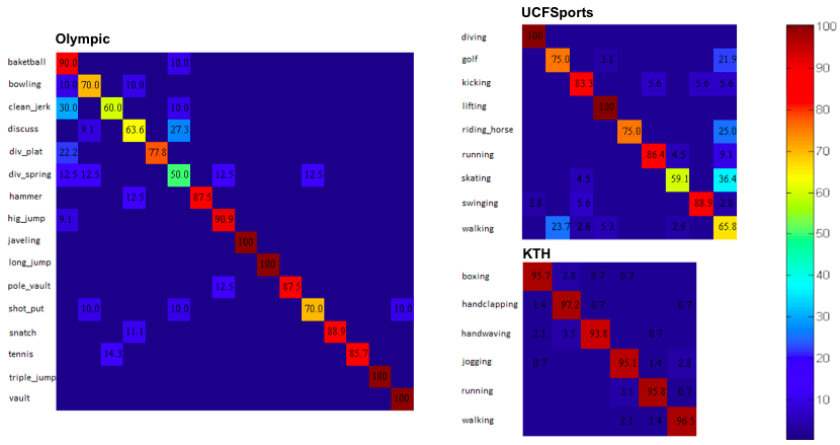

Fig. 4. Confusion matrix on the tested datasets using proposed method.

the confusion matrix (see Fig. 4), actions such as boxing, clapping and waving are confused. We anticipated these findings because these actions are characterized by similar hand motions. The same situation is observable with actions jogging, running and walking, due to the leg-centered motion. Nevertheless, our method obtains high accuracy in all classes.

Method	Year	Acc.(%)
Laptev et al. [15]	2008	91.8
Niebles et al. [18]	2010	91.3
Zang et al. [27]	2012	94.0
Our method		95.7

Table 1. Comparison of accuracy using KTH dataset

• *Olympic Dataset:* The Olympic Sports Dataset contains 16 actions and about 783 videos of athletes practicing different sports from Youtube [18]. This dataset is challenging because it contains camera motion and several views. Unlike the KTH and UCF-Sports datasets, the bounding boxes are not available. Due to the complexity of the scenes, applying the body-part model may produce errors in the actor detection. Therefore, we use the raw frames from each key-sequence to extract 450 random sub-cuboids. These sub-cuboids are used to obtain the temporal dictionaries. Table 2 shows the results where we obtain an overall performance of 81.3%.

We noted from the confusion matrix (see Fig. 4) that the actions associated with javeling, long jump, triple jump and vault are perfectly classified. These

Method	Year	Acc.(%)
Niebles et al. [18]	2010	72.1
Liu et al. [17]	2011	74.4
Jiang et al. [12]	2012	80.6
Our method		81.3

Table 2. Comparison of accuracy using Olympic dataset.

actions are characterized by a similar start and a different ending. In this case, it seems that key-sequences are important for breaking the actions down into atomic acts, which implies a good classification. However, actions that involve moderate motion and short duration such as clean and jerk, discus throwing, diving springboard and shot put have a lower recognition rate.

• *UCF-Sports Dataset:* This data set consists of 150 videos of several sports obtained from several sources. We decided to present results in 9 actions: diving, golf, kicking, lifting, riding horse, running, skateboarding, swinging and walking. This dataset present many challenges, including dynamic backgrounds, camera motion, scale changes, and variations in lighting and appearance. We increased the amount of data samples by adding a vertically flipped version of each video. We trained and tested all sequences (original and flipped version) and obtain an overall performance of 80.4 %. Table 3 shows the comparison our results using a leave-one-out setup.

Method	Year	Acc.(%)
Rodríguez et al. [20]	2008	69.2
Wang et al. [24]	2009	85.6
Guha et al. [10]	2012	83.8
Our method		80.4

Table 3. Comparison of accuracy using UCF-Sports dataset

Diving, lifting, kicking, running and swinging are correctly classified because they were broken down into differentiable acts. In addition, our method notably differentiates between running and kicking despite the fact that they involve similar acts. Other actions such as golf, riding horse and skate boarding, which involve objects and moderate body motions, present issues of misclassification. However, we believe that this is due to the fact that the key-sequence may contain similar information (due to the short duration of videos) and the lack of context information (see Fig. 4).

4 Conclusions

We proposed a non-supervised method to extract key-sequences that potentially contain the acts to summarize a video. Also, our method describes the video key-sequences as a mixture of different classes and different times called *inter-temporal acts descriptor*. Our approach achieved a robust representation to face the inner-class variations and several video lengths. The experiments have shown that our approach is effective and it obtains a good performance compared to other methods in the state-of-the-art in standard and challenging datasets. In a future work, we plan to integrated the overall process of our method using sparse dictionary learning.

Acknowledge

The work of Analí Alfaro has been supported by Mecesup Program for Doctoral Scholarship.

References

1. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. In IEEE Trans. on Signal Processing, pp.4311-4322, vol. 54 (2006).
2. Andriluka, M., Roth, S., Schiele, B.: Pictorial Structures Revisited: People Detection and Articulated Pose Estimation. In Proc. in Computer Vision and Pattern Recognition (2009)
3. Bobick, A., Davis, J.: The Recognition of Human Movement Using Temporal Templates. In IEEE Trans. on Pattern Analysis and Machine Intelligence, pp 257-267, vol. 23 (2001).
4. Carlsson, S., Sullivan, J.: Action Recognition by Shape Matching to Key Frames. In Workshop on Models versus Exemplars in Computer Vision (2001).
5. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.:Behavior Recognition via Sparse Spatio-Temporal Features. In VS-PETS (2005).
6. Efros, A., Berg, A., Berg ,Er., Mori, G., Malik, J.:Recognizing Action at a Distance. In Proc. in International Conference on Computer Vision (2003).
7. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In Proc. in Computer Vision and Pattern Recognition (2008).
8. Gaidon, A., Harchaoui, Z., Schmid, C.:Actom sequence models for efficient action detection. In Proc. in Computer Vision and Pattern Recognition (2011).
9. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.:Actions as space-time shapes. In Proc. in International Conference on Computer Vision(2005).
10. Guha, T., Ward, R.: Learning Sparse Representations for Human Action Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, pp 1576-1588, vol. 34 (2012).
11. Guo, K., Ishwar, P., Konrad, J.:Action Recognition in Video by Sparse Representation on Covariance Manifolds of Silhouette Tunnels. In Proc. in International Conference on Pattern Recognition (2010).

12. Jiang, Y., Dai, Q., Xue, X., Liu, W., Ngo, Ch.: Trajectory-Based Modeling of Human Actions with Motion Reference Points. In Proc. in European Conference on Computer Vision (2012).
13. Kläser, A., Marszalek, M., Schmid, C.:A spatio-temporal descriptor based on 3d-gradients. In Proc. in British Machine Vision Conference (2008).
14. Laptev, I., Lindeberg, T.:Space-time interest points. In Proc. in International Conference on Computer Vision (2003).
15. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.:Learning realistic human actions from movies. In Proc. in Computer Vision and Pattern Recognition (2008).
16. Liu, J., Ali, S., Shah, M.: Recognizing human actions using multiple features. In Proc. in Computer Vision and Pattern Recognition (2008).
17. Liu, J., Kuipers, B., Savarese, S.:Recognizing human actions by attributes. In Proc. in Computer Vision and Pattern Recognition (2011).
18. Niebles, J., Chen, Ch., Fei-Fei,L.:Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. In Proc. European Conference on Computer Vision (2010).
19. Pati, Y., Rezaiifar R., Krishnaprasad, P.:Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition. Proceedings of the 27 th Annual Asilomar Conference on Signals, Systems,and Computers (1993).
20. Rodriguez, M., Ahmed, J., Shah, M.: Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. In Proc. in Computer Vision and Pattern Recognition (2008).
21. Schindler, K., Van Gool, L.: Action Snippets: How many frames does human action recognition require?. In Proc. in Computer Vision and Pattern Recognition (2008).
22. Schüldt, C., Laptev, I., Caputo, B.:Recognizing human actions: A local SVM approach. In Proc. in International Conference on Pattern Recognition (2004).
23. Tran, K., Kakadiaris, I., Shah, S.:Modeling Motion of Body Parts for Action Recognition. In Proc. in British Machine Vision Conference (2011).
24. Wang, H., Ullah, M., Kläser, A., Laptev, I., Schmid, C.:Evaluation of local spatio-temporal features for action recognition. In Proc. in British Machine Vision Conference (2009).
25. Yao, A., Gall, U., Van Gool, L.:A Hough transform-based voting framework for action recognition. In Proc. in Computer Vision and Pattern Recognition (2010).
26. Yu, T., Kim, T., Cipolla, R.:Real-time Action Recognition by Spatiotemporal Semantic and Structural Forest. In Proc. in British Machine Vision Conference (2010).
27. Zhang, Y., Liu, X., Chang, M., Ge, W., Chen, T.: Spatio-Temporal Phrases for Activity Recognition. In Proc. in European Conference on Computer Vision (2012).