Video-Based Human Action Recognition Using Kernel Relevance Analysis

Jorge Fernández-Ramírez[1,2(✉)], Andrés Álvarez-Meza[1,2], and Álvaro Orozco-Gutiérrez[1,2]

[1] Automatics Research Group, Pereira, Colombia
[2] Universidad Tecnolgica de Pereira, Pereira, Colombia
jorgeferram17@utp.edu.co

Abstract. This paper presents a video-based Human Action Recognition using kernel relevance analysis. Our approach, termed HARK, comprises the conventional pipeline employed in action recognition, with a two-fold post-processing stage: (i) A descriptor relevance ranking based on the centered kernel alignment (CKA) algorithm to match trajectory-aligned descriptors with the output labels (action categories), and (ii) a feature embedding based on the same algorithm to project the video samples into the CKA space, where the class separability is preserved, and the number of dimensions is reduced. For concrete testing, the UCF50 human action dataset is employed to assess the HARK under a leave-one-group-out cross-validation scheme. Attained results show that the proposed approach correctly classifies the 90.97% of human actions samples using an average input data dimension of 105 in the classification stage, which outperforms state-of-the-art results concerning the trade-off between accuracy and dimensionality of the final video representation. Also, the relevance analysis allows to increase the video data interpretability, by ranking trajectory-aligned descriptors according to their importance to support action recognition.

Keywords: Human action recognition · Relevance analysis
Feature embedding · Kernel methods

1 Introduction

Human action recognition has become an important research area in the computer vision field due to its wide range of applications, including automatic video analysis, video indexing and retrieval, video surveillance, and virtual reality [5]. As a result of the increasing amount of video data available both on internet repositories and personal collections, there is a strong demand for understanding the content of complex real-world data. However, different challenges arise for action recognition in realistic video data [13]. First, there is large intra-class variation caused by factors such as the style and duration of the performed action, scale changes, dynamic viewpoint, and sudden motion. Second, background clutter, occlusions, and low-quality video data are known to affect robust recognition

G. Bebis et al. (Eds.): ISVC 2018, LNCS 11241, pp. 116–125, 2018.
https://doi.org/10.1007/978-3-030-03801-4_11

as well. Finally, for large-scale datasets, the data processing represents a crucial computational challenge to be addressed [3].

The most popular framework for action recognition is the Bag of visual Words (BOW) with its variations [11,12]. The BOW pipeline contains three main stages: feature estimation, feature encoding, and classification. Besides, there are several pre-processing and post-processing stages, such as relevance analysis and feature embedding to enhance data decorrelation, separability and interpretability [2]. Furthermore, different normalization techniques have been introduced for improving the performance of the recognition system. For the feature estimation step, the recent success of local space-time features like Dense Trajectories (DT) and Improved Dense Trajectories (iDT) has lead researchers to use them on a variety of datasets, obtaining excellent recognition performance [12,13]. Regarding the feature encoding step, super-vector based encoding methods such as Fisher Vector (FV) and Vector of Locally Aggregated Descriptors (VLAD) are presented as the state-of-the-art approaches for feature encoding in action recognition tasks [5,11]. Lastly, the classification stage has usually been performed by Support Vector Machines (SVM) in most recognition frameworks [8,10].

The feature encoding method that provides the final video representation is crucial for the performance of an action recognition system, as it influences directly the classifier ability to predict the class labels. However, video representations generated by methods such as FV or VLAD are known to provide high dimensional encoding vectors which increases the computational requirements in the classification stage [5,13]. On the other hand, the high dimensionality of the input data could affect the classifier accuracy adversely, by using redundant information and even noise, which do not enhance data separability. Therefore, the Dimensionality Reduction (DR), which consists of feature selection and feature embedding methods, is imperative to lighten the burden associated with the encoding stage, eliminate redundant information, and project samples into new spaces to increase separability [1]. Conventional methods, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) have been proposed to decorrelate the individual features of descriptors and reduce their length, in a pre-processing stage to the encoding [6]. Nevertheless, these methods are specially designated to work with real-valued vectors coming from flat Euclidean spaces. Thus, in modern computer vision due to real-world data and models, there has been growing interest to go beyond the extensively studied Euclidean spaces and analyse more realistic non-linear scenarios for better representation of the data [7].

In this work, we introduce a new human action recognition system using kernel relevance analysis. The system, based on a non-linear representation of the super-vector obtained by the FV encoding technique, seeks to reduce the input space dimensionality, as well as, enhance separability and interpretability of video data. Specifically, our approach includes a centered kernel alignment (CKA) technique to recognize relevant descriptors related to action recognition. Hence, we match trajectory-aligned descriptors with the output labels (action categories) through non-linear representations [2]. Also, the CKA-algorithm allows to compute a linear projection matrix, where the columns quantify the

required number of dimensions to preserve the 90% of the input data variability. Therefore, by projecting the video samples into the CKA generated space, the class separability is preserved, and the number of dimensions is reduced. Attained results on the UCF50 database demonstrate that our proposal favors the interpretability of the commonly employed descriptors in action recognition, and presents a system able to obtain competitive recognition accuracy using a drastically reduced input space dimensionality to the classification stage.

The rest of the paper is organized as follows: Sect. 2 presents the main theoretical background, Sect. 3 describes the experimental set-up, Sect. 4 introduces the results and discussions. Finally, Sect. 5 shows the conclusions.

2 Kernel-Based Descriptor Relevance Analysis and Feature Embedding

Let $\{\boldsymbol{Z}_n \in \mathbb{R}^{T \times D}, y_n \in \mathbb{N}\}_n^N$ be an input-output pair set holding N video samples, each of them represented by T trajectories generated while tracking a dense grid of pixels, whose local space is characterized by a descriptor of dimensionality D, as presented in [13]. Here, the samples are related to a set of human action videos meanwhile the descriptor in turn is one of the following trajectory-aligned measure: trajectory positions (Trajectory), Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF), Motion Boundary Histograms (MBHx and MBHy), yielding a total of $F = 5$ descriptors. Likewise, the output label y_n denotes the human action being performed in the corresponding video representation. From $\boldsymbol{Z}_n$, we aim to encode T described trajectories concerning a Gaussian Mixture Model (GMM), trained to be a generative model of the descriptor in turn. Therefore, the Fisher Vector (FV) feature encoding technique is employed, as follows [9]:

Let $\boldsymbol{Z}_n$ be a matrix holding T described trajectories $\boldsymbol{z}_t \in \mathbb{R}^D$, and υ^λ be a GMM with parameters $\lambda = \{w_i \in \mathbb{R}, \boldsymbol{\mu}_i \in \mathbb{R}^D, \sigma_i^2 \boldsymbol{I} \in \mathbb{R}^{D \times D}\}_{i=1}^K$, which are respectively the mixture weight, mean vector, and diagonal covariance matrix of K Gaussians. We assume that $\boldsymbol{z}_t$ is generated independently by υ^λ. Therefore, the gradient of the log-likelihood describes the contribution of the parameters to the generation process:

$$\boldsymbol{x}_n^\lambda = \frac{1}{T} \sum_{t=1}^{T} \nabla_\lambda \log \upsilon_\lambda(\boldsymbol{z}_t) \tag{1}$$

where ∇_λ is the gradient operator w.r.t λ. Mathematical derivations lead $\boldsymbol{x}_n^{\mu,i}$ and $\boldsymbol{x}_n^{\sigma,i}$ to be the D-dimensional gradient vectors w.r.t the mean and standard deviation of the Gaussian i, that is:

$$\boldsymbol{x}_n^{\mu,i} = \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \boldsymbol{\gamma}_t(i) \left(\frac{\boldsymbol{z}_t - \boldsymbol{\mu}_i}{\sigma_i} \right), \tag{2}$$

$$\boldsymbol{x}_n^{\sigma,i} = \frac{1}{T\sqrt{2w_i}} \sum_{t=1}^{T} \boldsymbol{\gamma}_t(i) \left[\frac{(\boldsymbol{z}_t - \boldsymbol{\mu}_i)^2}{\sigma_i^2} - 1 \right] \tag{3}$$

where $\boldsymbol{\gamma}_t(i)$ is the soft assignment of trajectory $\boldsymbol{z}_t$ to the Gaussian i, that is:

$$\boldsymbol{\gamma}_t(i) = \frac{w_i \upsilon_i(\boldsymbol{z}_t)}{\sum_{j=1}^{K} w_j \upsilon_j(\boldsymbol{z}_t)} \tag{4}$$

The final gradient vector $\boldsymbol{x}_n^{\lambda}$ is a concatenation of the $\boldsymbol{x}_n^{\mu,i}$ and $\boldsymbol{x}_n^{\sigma,i}$ vectors for $i = 1, \ldots, K$ and is $2KD$-dimensional.

Assuming that the same procedure is performed for each descriptor, the concatenation of the resulting vectors generates the set $\{\boldsymbol{x}_n \in \mathbb{R}^{2K(D_1+\cdots+D_F)}, y_n \in \mathbb{N}\}_n^N$. Afterwards, a *Centered Kernel Alignment* (CKA) approach is performed to compute a linear projection matrix, and to determine the relevance weight from each trajectory-aligned descriptor individual feature, as follows [2]:

Let $\kappa_X : \mathbb{R}^S \times \mathbb{R}^S \to \mathbb{R}$, where $S = 2K(D_1 + \cdots + D_F)$, be a positive definite kernel function, which reflects an implicit mapping $\phi : \mathbb{R}^S \to \mathcal{H}_X$, associating an element $\boldsymbol{x}_n \in \mathbb{R}^S$ with the element $\phi(\boldsymbol{x}_n) \in \mathcal{H}_X$, that belongs to the Reproducing Kernel Hilbert Space (RKHS), $\mathcal{H}_X$. In particular, the Gaussian kernel is preferred since it seeks an RKHS with universal approximation capability, as follows [4,14]:

$$\kappa_X(\boldsymbol{x}_n, \boldsymbol{x}_{n'}; \sigma) = \exp\left(-\upsilon^2(\boldsymbol{x}_n, \boldsymbol{x}_{n'})/2\sigma^2\right);\ n, n' \in \{1, 2, \ldots, N\}, \tag{5}$$

where $\upsilon(\cdot,\cdot) : \mathbb{R}^S \times \mathbb{R}^S \to \mathbb{R}$ is a distance function in the input space, and $\sigma \in \mathbb{R}^+$ is the kernel bandwidth that rules the observation window within the assessed similarity metric. Likewise, for the output labels space $\mathcal{L} \in \mathbb{N}$, we also set a positive definite kernel $\kappa_L : \mathcal{L} \times \mathcal{L} \to \mathcal{H}_L$. In this case, the pairwise similarity distance between samples is defined as $\kappa_L(y_n, y_{n'}) = \delta(y_n - y_{n'})$, being $\delta(\cdot)$ the Dirac delta function. Each of the above defined kernels reflects a different notion of similarity and represents the elements of the matrices $\mathbf{K}_X, \mathbf{K}_L \in \mathbb{R}^{N \times N}$, respectively. In turn, to evaluate how well the kernel matrix $\mathbf{K}_X$ matches the target $\mathbf{K}_L$, we use the statistical alignment between those two kernel matrices as [2]:

$$\hat{\rho}(\mathbf{K}_X, \mathbf{K}_L) = \frac{\langle \bar{\mathbf{K}}_X, \bar{\mathbf{K}}_L \rangle_{\mathrm{F}}}{\sqrt{\langle \bar{\mathbf{K}}_X \bar{\mathbf{K}}_X \rangle_{\mathrm{F}} \langle \bar{\mathbf{K}}_L \bar{\mathbf{K}}_L \rangle_{\mathrm{F}}}}, \tag{6}$$

where the notation $\bar{\mathbf{K}}$ stands for the centered kernel matrix calculated as $\bar{\mathbf{K}} = \tilde{\boldsymbol{I}}\mathbf{K}\tilde{\boldsymbol{I}}$, being $\tilde{\boldsymbol{I}} = \boldsymbol{I} - \mathbf{1}^\top \mathbf{1}/N$ the empirical centering matrix, $\boldsymbol{I} \in \mathbb{R}^{N \times N}$ is the identity matrix, and $\mathbf{1} \in \mathbb{R}^N$ is the ones vector. The notation $\langle \cdot, \cdot \rangle_{\mathrm{F}}$ represents the matrix-based Frobenius norm. Hence, Eq. (6) is a data driven estimator that allows to quantify the similarity between the input feature space and the output label space [2]. In particular, for the Gaussian kernel κ_X, the Mahalanobis distance is selected to perform the pairwise comparison between samples:

$$\upsilon_{\boldsymbol{A}}^2(\boldsymbol{x}_n, \boldsymbol{x}_{n'}) = (\boldsymbol{x}_n - \boldsymbol{x}_{n'})\boldsymbol{A}\boldsymbol{A}^\top(\boldsymbol{x}_n - \boldsymbol{x}_{n'})^\top,\ n, n' \in \{1, 2, \ldots, N\}, \tag{7}$$

where the matrix $\boldsymbol{A} \in \mathbb{R}^{S \times P}$ holds the linear projection in the form $w_n = \boldsymbol{x}_n \boldsymbol{A}$, with $w_n \in \mathbb{R}^P$, being P the required number of dimensions to preserve the 90% of

the input data variability, and $\boldsymbol{A}\boldsymbol{A}^\top$ the corresponding inverse covariance matrix in Eq. (7), assuming $P \leq S$. Therefore, intending to compute the projection matrix $\boldsymbol{A}$, the formulation of a CKA-based optimizing function can be integrated into the following kernel-based learning algorithm:

$$\hat{\boldsymbol{A}} = \arg\max_{\boldsymbol{A}} \log\left(\hat{\rho}(\mathbf{K}_X(\boldsymbol{A};\sigma), \mathbf{K}_L)\right), \tag{8}$$

where the logarithm function is employed for mathematical convenience. The optimization problem from Eq. (8) is solved using a recursive solution based on the well-known gradient descent approach. After the estimation of the projection matrix $\hat{\boldsymbol{A}}$, we assess the relevance of the S input features. To this end, the most contributing features are assumed to have the higher values of similarity relationship with the provided output labels. Specifically, the CKA-based relevance analysis calculates the relevance vector index $\varrho \in \mathbb{R}^S$, holding elements $\varrho_s \in \mathbb{R}^+$ that allows to measure the contribution from each of the s-th input features in building the projection matrix $\hat{\boldsymbol{A}}$. Hence, to calculate those elements, a stochastic measure of variability is utilized as follows: $\varrho_s = \mathbb{E}_P\{|a_{s,p}|\}$; where $p \in \{1, 2, \dots P\}$, $s \in \{1, \dots, S\}$, and $a_{s,p} \in \hat{\boldsymbol{A}}$.

3 Experimental Set-Up

Database. To test our *video-based human action recognition using kernel relevance analysis* (HARK), we employ the UCF50 database [10]. This database contains realistic videos taken from Youtube, with large variations in camera motion, object appearance and pose, illumination conditions, scale, etc. For concrete testing, we use $N = 5967$ videos concerning the 46 human action categories in which the human bounding box file was available [13]. The video frames size is 320×240 pixels, and the length varies from around 70–200 frames. The dataset is divided into 25 predefined groups. Following the standard procedure, we perform a leave-one-group-out cross-validation scheme and report the average classification accuracy overall 25 folds.

HARK Training. Initially, for each video sample in the dataset we employ the Improved Dense Trajectory feature estimation technique (iDT), with the code provided by the authors in [13], keeping the default parameter settings to extract $F = 5$ different descriptors: Trajectory (x, y normalized positions along 15 frames), HOG, HOF, MBHx, MBHy. The iDT technique is an improved version of the previously realized Dense Trajectory technique from the same author, which removes the trajectories generated by the camera motion and the inconsistent matches due to humans. Thus, the human detection is a challenging requirement in this technique, as people in action datasets appear in many different poses, and could only be partially visible due to occlusion or by being partially out-of-scene. These five descriptors are extracted along all valid trajectories and the resulting dimensionality D_f is 30 for the trajectory, 96 for HOG, MBHx and MBHy, and 108 for HOF.

We then randomly select a subsample of $5000 \times K$ trajectories from the training set to estimate a GMM codebook with $K = 256$ Gaussians, and the FV

encoding is performed as explained in Sect. 2. Afterwards, we apply to the resulting vector a Power Normalization (PN) followed by the L2-Normalization ($||\text{sign}(x)|x|^{\alpha}||$, where $0 \leq a \leq 1$ is the normalization parameter). The above procedure is performed per descriptor, fixing $\alpha = 0.1$. Next, all five normalized FV representations are concatenated together, yielding $S = 218112$ encoding dimension. The linear projection matrix $\hat{\boldsymbol{A}} \in \mathbb{R}^{S \times P}$ and the relevance vector index $\boldsymbol{\varrho} \in \mathbb{R}^{S}$ are computed as explained in section Sect. 2; where P=104.8, is the average required number of dimensions, through 25 leave-one-out iterations, to preserve the 90% of the input data variability.

For the classification step, we use a one-vs-all Linear SVM with regularization parameter equal to 1 and a Gaussian kernel SVM, varying the kernel bandwidth between the range $[0.1\sigma_o, \sigma_o]$, being $\sigma_o \in \mathbb{R}^{+}$ the median of the input space Euclidean distances; and searching the regularization parameter within the set $\{0.1, 1, 100, 500, 1000\}$, by nested cross-validation with the same leave-one-group-out scheme. Figure 1 summarizes the HARK training pipeline. It is worth noting that all experiments were performed using the Matlab software on a Debian server with 230 GB of RAM and 40 cores. The FV code is part of the open-source library VLFeat, the implementation is publicly available[1]. On the other hand, the CKA code was developed by Alvarez-Meza *et al.* in [2] and is also publicly available[2].

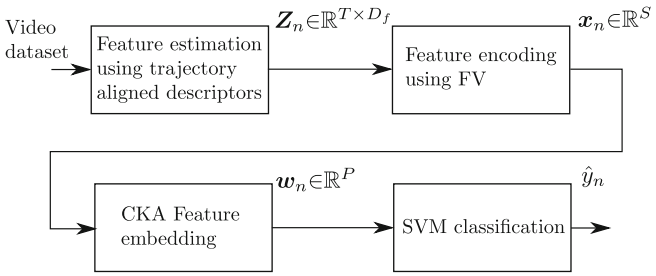


Fig. 1. Sketch of the proposed HARK-based action recognition system.

4 Results and Discussions

Figure 2, shows a visual example of feature estimation and encoding using trajectory-aligned descriptors and BOW. From the color points, where different colors represent the assignment of a given trajectory to one of the prototype vectors generated by the k-means algorithm, we can appreciate the hard assignment of trajectory descriptors in the BOW encoding. Also, different sizes of the points represent the scale in which the trajectory is generated. In contrast, this paper uses the soft assignment of the GMM-based FV encoding, which is not

[1] http://www.vlfeat.org/overview/encodings.html.
[2] https://github.com/andresmarino07utp/EKRA-ES.

as straightforward to express in a figure. It is worth noting that due to the human segmentation performed before the trajectory-based feature estimation, the encoding points are mainly grouped in the player whereabouts, which constrains the zone of interest to only characterize the player information. This strategy helps to reduce the uncertainty from the video representation, as the influence of the background is decreased.

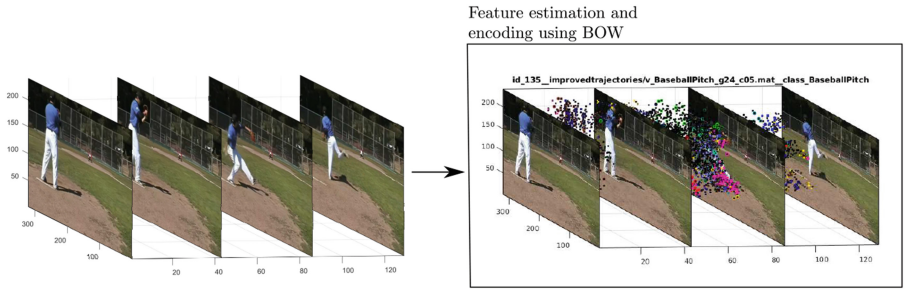
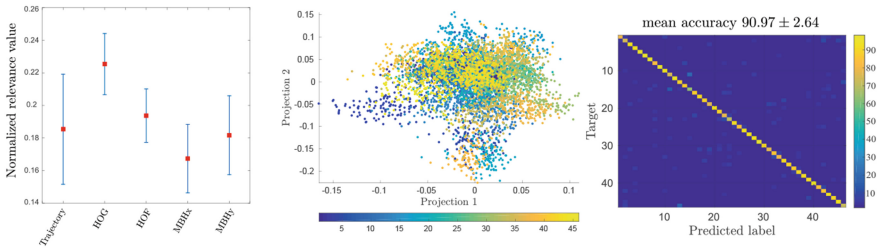


Fig. 2. Feature estimation and encoding using trajectory-aligned descriptors and BOW.

Figure 3(a) shows the normalized relevance value of the provided Trajectory, HOG, HOF, MBHx, and MBHy descriptors, this figure is generated by averaging the components of $\boldsymbol{\varrho} \in \mathbb{R}^S$ which corresponds to each descriptor. Therefore, the mean and standard deviation is presented to represent the descriptor relevance vector. As seen, the HOG descriptor exhibit the highest relevance value regarding our HARK criteria, this descriptor quantify the local appearance and shape within the trajectory-aligned space window through the distribution of intensity gradients. Notably, all the others descriptors mainly quantifies the human local motion (Trajectory normalized positions, HOF, MBHx, MBHy), are very close regarding their relevance value. Hence, the trajectory-aligned descriptors match similarly the human actions labels concerning the CKA-based analysis presented in Sect. 2, as they are all local measures of appearance, shape, and motion equally important to support action recognition. Remarkable, the relevance value in Fig. 3(a) mainly depends upon the discrimination capability of the Gaussian kernel in Eq. 5, and the local measure being performed by the descriptor. Now, as seen in Fig. 3(b), the CKA embedding in its first two projections provides an insight into the data overlapping. The studied classes overlapping (human actions) can be attributed to similar intra-class variations in several categories, as videos with realistic scenarios have inherent attributes such as background clutter, scale changes, dynamic viewpoint and sudden motion, that may be affecting adversely the class separability.

Furthermore, as it is evidenced by the confusion matrix of the test set in Fig. 3(c), an RBF SVM over the CKA feature embedding can obtain $90.97 \pm 2.64\%$ of accuracy in classifying human actions on the employed dataset. From this matrix, the classes 22 and 23 are generating classification problems because

the human movements performed in both are similar, these classes correspond to Nunchucks and Pizza tossing respectively. As expected, the RBF SVM can achieve more reliable recognition than a Linear SVM, as the data problem in Fig. 3(b) is non-linear, see the results presented for this paper in Table 1. Notable, our approach requires only 104.8 dimensions on average through 25 leave-one-out iterations to classify 46 actions of the UCF50 dataset, with competitive accuracy, which is very useful, because more elaborated classifiers (once discarded due to the data dimension) can be employed to increase the recognition rate further.



(a) (b) (c)

Fig. 3. Human action recognition on the UCF50 database. (a) Feature relevance values. (b) 2D input data projection from 46 action categories using CKA.(c) Confusion matrix for the test set under a nested leave-one-group-out validation scheme using an RBF SVM classifier.

In turn, Table 1 presents a comparative study of the results achieved by our HARK and other similar approaches from the state-of-the-art for human action recognition on the UCF50 database. To build this comparative analysis, approaches with similar experimental set-up are employed. Specifically, those approaches using iDT representation and similar descriptors. Primarily, the compared results exhibit a trade-off between data dimension and accuracy, more elaborate procedures such as the one presented in [5], uses Time Convolutional Networks (TCN) and Spatial Convolutional Networks (SCN) descriptors along with iDT descriptors, and Spatio-temporal VLAD (ST-VLAD) encoding to enhance the class separability. Thus, the mentioned approach obtain very high mean accuracy 97.7%. However, the data dimensionality is considerably high, which limits the usage of many classifiers. On the other hand, the approach presented in [13], enhances the spatial resolution of the iDT descriptors by using a strategy called spatiotemporal pyramids (STP) along with Spatial Fisher Vector encoding (SFV). Obtained results regarding the accuracy of [13] are comparable to ours. Nonetheless, the data dimension is drastically higher.

Table 1. Comparison with similar approaches in the state-of-the-art on the UCF50 dataset.

Reference	Representation	Descriptors	Feature encoding	Data dimension	Classification method	Accuracy [%]
Uijlings *et al* [11]	–	HOG+HOF+MBHx+MBHy	FV	36864	Linear SVM	81.8
Wang *et al* [13]	iDT	HOG+HOF+MBHx+MBHy	SFV + STP	611328	Linear SVM	91.7
Duta *et al* [5]	iDT+2St	HOG+HOF+MBHx+MBHy+SCN+TCN	ST-VLAD	258816	Linear SVM	97.7
HARK	iDT	Traj+ HOG+HOF+MBHx+MBHy	FV + CKA	**104.8**	Linear SVM	**87.9**
HARK	iDT	Traj+ HOG+HOF+MBHx+MBHy	FV + CKA	**104.8**	RBF SVM	**90.9**

5 Conclusions

In this paper, we introduced a video-based human action recognition system using kernel relevance analysis (HARK). Our approach highlights the primary descriptors to predict the output labels of human action videos using trajectory representation. Therefore, HARK quantifies the relevance of $F = 5$ trajectory-aligned descriptors towards a CKA-based algorithm, that matches the input space with the output labels, to enhance the descriptor interpretability, as it allows to determine the importance of local measures (appearance, shape, and motion) to support action recognition. Also, the CKA-algorithm allows to compute a linear projection matrix, through a non-linear representation, where the columns quantify the required number of dimensions to preserve the 90% of the input data variability. Hence, by projecting the video samples into the generated CKA space, the class separability is preserved, and the number of dimensions is reduced. Attained results on the UCF50 database show that our proposal correctly classified the 90.97% of human actions samples using an average input data dimension of 104.8 in the classification stage, through 25 folds under a leave-one-group-out cross-validation scheme. In particular, according to the performed relevance analysis, the most relevant descriptor is the HOG which quantifies the local appearance and shape through the distribution of intensity gradients. Remarkable, HARK outperforms state-of-art results concerning the trade-off between the accuracy achieved and the required data dimension (Table 1). As future work, authors plan to employ other descriptors such as the deep features presented in [5]. Also, a HARK improvement based on the enhancement of spatial and temporal resolution, as the one presented in [13], could be an exciting research line.

Acknowledgments. Under grants provided by the project 1110-744-55958 funded by COLCIENCIAS. Also, J. Fernández is partially founded by the COLCIENCIAS project "ATTENDO" - code: FP44842-424-2017, and by the Maestría en Ingeniería Eléctrica from the Universidad Tecnológica de Pereira.

References

1. Ai, S., Lu, T., Xiong, Y.: Improved dense trajectories for action recognition based on random projection and fisher vectors. In: MIPPR 2017: Pattern Recognition and Computer Vision, International Society for Optics and Photonics, vol. 10609, p. 1060915 (2018)
2. Alvarez-Meza, A.M., Orozco-Gutierrez, A., Castellanos-Dominguez, G.: Kernel-based relevance analysis with enhanced interpretability for detection of brain activity patterns. Front. Neurosci. **11**, 550 (2017)
3. Álvarez-Meza, A.M., Molina-Giraldo, S., Castellanos-Dominguez, G.: Background modeling using object-based selective updating and correntropy adaptation. Image Vis. Comput. **45**, 22–36 (2016)
4. Brockmeier, A.J., et al.: Information-theoretic metric learning: 2-D linear projections of neural data for visualization. In: EMBC, pp. 5586–5589. IEEE (2013)
5. Duta, I.C., Ionescu, B., Aizawa, K., Sebe, N.: Spatio-temporal VLAD encoding for human action recognition in videos. In: Amsaleg, L., Guðmundsson, G.Þ., Gurrin, C., Jónsson, B.Þ., Satoh, S. (eds.) MMM 2017. LNCS, vol. 10132, pp. 365–378. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-51811-4_30
6. Guo, K., Ishwar, P., Konrad, J.: Action recognition from video using feature covariance matrices. IEEE Trans. Image Process. **22**(6), 2479–2494 (2013)
7. Harandi, M., Salzmann, M., Hartley, R.: Dimensionality reduction on spd manifolds: the emergence of geometry-aware methods. IEEE Trans. Pattern Anal. Mach. Intell. (2017)
8. Li, Q., Cheng, H., Zhou, Y., Huo, G.: Human action recognition using improved salient dense trajectories. Comput. Intell. Neurosci. (2016)
9. Perronnin, F., Snchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6314, LNCS (PART 4), pp. 143–156 (2010)
10. Reddy, K.K., Shah, M.: Recognizing 50 human action categories of web videos. Mach. Vis. Appl. **24**(5), 971–981 (2013)
11. Uijlings, J., Duta, I.C., Sangineto, E., Sebe, N.: Video classification with densely extracted hog/hof/mbh features: an evaluation of the accuracy/computational efficiency trade-off. Int. J. Multimedia Inf. Retrieval **4**(1), 33–44 (2015)
12. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vis. **103**(1), 60–79 (2013)
13. Wang, H., Oneata, D., Verbeek, J., Schmid, C.: A robust and efficient video representation for action recognition. Int. J. Comput. Vis. **119**(3), 219–238 (2016)
14. Wang, Y., et al.: Tracking neural modulation depth by dual sequential monte carlo estimation on point processes for brain-machine interfaces. IEEE Trans. Biomed. Eng. **63**(8), 1728–1741 (2016)