This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:https://doi.org/10.1016/j.cviu.2015.09.003, Computer Vision and Image Understanding

Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Action Recognition by using kernels on aclets sequences

Luc Brun[b], Gennaro Percannella[a], Alessia Saggese[a,**], Mario Vento, IAPR Fellow[a]

[a]*Dept. of Information and Electrical Engineering and Applied Mathematics, University of Salerno, via Giovanni Paolo II, 84084 Fisciano (SA), Italy*
[b]*GREYC UMR CNRS 6072, ENSICAEN - Université de Caen Basse-Normandie, 14050 Caen, France*

ABSTRACT

In this paper we propose a method for human action recognition based on a string kernel framework. An action is represented as a string, where each symbol composing it is associated to an *aclet*, that is an atomic unit of the action encoding a feature vector extracted from raw data. In this way, measuring similarities between actions leads to design a similarity measure between strings. We propose to define this string's similarity using the global alignment kernel framework. In this context, the similarity between two aclets is computed by a novel soft evaluation method based on an enhanced gaussian kernel. The main advantage of the proposed approach lies in its ability to effectively deal with actions of different lengths or different temporal scales as well as with noise introduced during the features extraction step. The proposed method has been tested over three publicly available datasets, namely the MIVIA, the CAD and the MHAD, and the obtained results, compared with several state of the art approaches, confirm the effectiveness and the applicability of our system in real environments, where unexperienced operators can easily configure it.

1. Introduction

In the last years the analysis and the recognition of human motion from images and videos have interested several scientists working in pattern recognition and computer vision fields (Hu et al. (2004), Chaquet et al. (2013), Vishwakarma and Agrawal (2013), Turaga et al. (2008), Lee et al. (2014)). The large interest shown by many researchers toward action recognition is motivated by the virtually infinite applicative domains where it is possible to exploit detailed information regarding actions for achieving real time situation awareness, augmented reality, improved human-machine interaction, etc. Examples range from surveillance systems and ambient assisted living with the possibility of identifying some events of interest occurring into a video (an aggression, a robbery, a tumble of an elder person, etc), to video retrieval by allowing a human operator to easily retrieve specific types of actions performed by persons in a video footage, and to computer gaming for delivering more and more advanced, realistic, and engaging experience for the player.

Despite the difficulty to define a strict partition among the types of movements, three main classes are identified in Moeslund et al. (2006) and Aggarwal and Ryoo (2011) with different nomenclatures: gestures, actions and activities. Gestures (Mitra and Acharya (2007)) are elementary movements of a single body part, such as a hand, an arm, the face or the head. An action (Poppe (2010), Weinland et al. (2011)) describes a movement of the whole human body, such as walking or drinking. Finally, an activity (Aggarwal and Ryoo (2011)) is evaluated over a longer period of time, often analyzing object's trajectories, and may involve more persons and/or objects as two persons fighting, a person leaving a baggage, and so on. In this paper, we focus on the recognition of the second class of movement, namely actions.

Methods available in the literature formalize the action recognition as a traditional pattern recognition problem, where, as stated in Poppe (2010), a set of features is computed on the sequence of images and is used to feed a classifier. Several issues make this problem very challenging: indeed, a method for recognizing human actions has to be sufficiently general to deal with intra-class variations (actions performed by different persons, but typically in different ways), and at the same time should be able to distinguish among different actions even when performed by the same person. Furthermore, the same class of

**Corresponding author: Tel.: +39 089 963006
e-mail: asaggese@unisa.it (Alessia Saggese)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:https://doi.org/10.1016/j.cviu.2015.09.003, Computer Vision and Image Understanding

actions can be performed in different places, so implying different conditions (for instance related to the illumination of the environment or the angle under which the camera takes a person performing the action) which may change the appearance of the subject under test. Finally, algorithms should be able to recover actions having different durations and performed at different paces.

Within this framework, the contribution of the scientific community has been mainly devoted to the definition of novel feature vectors, aiming at describing the spatio-temporal movement of humans so as to be sufficiently general to deal with different persons and actions performed under different conditions, but at the same time sufficiently discriminant to discern among the different actions of interest. In particular, according to Poppe (2010), two different categories of descriptors can be defined, namely *local* and *global*. Local descriptors are based on a bottom-up approach: a set of spatio temporal interest points is detected by locally analyzing each pixel of the image and assigning a local patch to each of these points. These patches are then combined, for instance using a traditional bag of features approach (Dollar et al. (2005), Kovashka and Grauman (2010), Wang et al. (2009), Jhuang et al. (2007), Lee et al. (2014)). The main advantage of local descriptors is that they are less sensitive to partial occlusions and are typically computed on the whole image, so without the need to detect moving objects and to track them in successive frames. This is a useful property in all those cluttered and/or crowded scenes when reliable object detection and tracking can not be achieved. However, these advantages are counter balanced by the high computational effort required to extract these descriptors. Furthermore, the performance that can be achieved when using local descriptors strongly depends on the amount and on the reliability of the interest points detected by the system: in some cases, for instance due to camera movements, interest points extraction may be unreliable thus requiring specific pre-processing steps.

On the other hand, global descriptors are evaluated by using a top-down approach; the person of interest is first located in the scene and its *blob* is extracted using detection and/or tracking techniques. Starting from the blob, the action is globally described by features derived from silhouettes, edges or optical flow (Chen et al. (2011), Wang et al. (2007), Carletti et al. (2013), Ofli et al. (2012), Kellokumpu et al. (2011), Jiang et al. (2012), Jiang et al. (2013)). The main advantage of this family of approaches is that the feature vector extracted from the blob is solely based on the useful information strictly related to the movement of the person. Of course, it is evident that the higher is the accuracy in the extraction of the blob, the better will be the overall performance of such approaches. However, their main limitation is in general their sensitivity to noise, partial occlusions and variations in viewpoint. In order to overcome such limitations, Megavannan et al. (2012) propose a representation that is a trade off between the global and local approaches through a grid-based description: the minimum box enclosing the silhouette is partitioned into cells in order to also consider a kind of local information about parts of the objects.

Unfortunately the low level representation provided by both local and global descriptors is sensitive to errors introduced by either interest point or blob detection methods. In order to avoid this problem, a common solution recently adopted in the literature consists in introducing a high-level representation, able to improve the discriminative power of the overall recognition system. For instance, in Sung et al. (2012) the authors adopt a classification strategy based on Hidden Markov Model (HMM): they create a two-layers Maximum Entropy Markov strategy for modeling an activity as a set of sub-activities and exploit a dynamic programming approach for the inference. In general, the main drawback of HMM based approaches lies in the large amount of labeled data required during the training step, which is often not simple to obtain in real environments. Furthermore, the performance of HMM strongly depends on a good configuration of hidden states, which is difficult to achieve and in general requires an expertise in this field for a proper choice.

In Foggia et al. (2014) the high level representation is obtained directly from the data, by exploiting regularities in the training set: in particular, this is obtained through a Deep Belief Network trained by a Restricted Boltzmann Machine. Given a suitably large dataset, the deep representation typically outperforms a hand-crafted description scheme, without requiring an heavy effort in feature design. However, as in the case of the HMM architecture, the parameters used to train the network strongly influence the achieved performance, and usually require the intervention of an expert of the paradigm.

A high level representation based on bag of words has been exploited in (Foggia et al. (2013)): the action is represented by an histogram, which encodes the occurrence of feature vectors in a given temporal window (the so called visual words) according to a dictionary extracted during the learning phase. In this way, the similarity between two actions is expressed in terms of similarity between histograms rather than in terms of similarity between low level feature vectors, so making the system less sensitive to errors introduced during the features extraction step. One of the main advantages of this approach lies on the simplicity of its configuration (only few parameters need to be set up). However, although the promising performance obtained by Foggia et al. (2013), a larger experimentation has highlighted its main limitation: this method bases its decision only on the occurrence or on the absence of relevant visual words within a given temporal window, then the temporal information can not be fully taken into account. It is important to highlight that the temporal information is instead a very important feature, since human vision implicitly analyzes the sequence of the motion patterns composing the actions in order to distinguish them (Johansson (1973)).

In order to overcome this limitation, a *temporal* bag of words has been recently exploited (Shukla et al. (2013), Bettadapura et al. (2013), Hernandez-Garcia et al. (2014)): the common idea is to encode in the histogram information about the relative position of a frame with respect to the other ones. For instance, in Shukla et al. (2013) the video is partitioned into fixed length bins and a single histogram is extracted for each bin; the video is finally represented as a k-dimensional feature vector, where each dimension encodes a single bin. The partitioning of the video is finally iterated in a hierarchical way by considering

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:https://doi.org/10.1016/j.cviu.2015.09.003, Computer Vision and Image Understanding

different bins' lengths, so as to deal with actions performed at different rates. However, the bigger is the number of layers, the higher will be the performance but at the same time the higher will be the computational effort required by the system. In Bettadapura et al. (2013) a n-grams based approach is exploited: a set of grams, namely the possible sequences of l words, being $l = \{1, .., n\}$, are used to built the codeword so as to consider both the temporal and causal information; furthermore, randomly generated sequences are added to the codeword using regular expressions. Finally, the action is represented by computing the histogram of occurrences, as in the traditional bag of words approach. A similar strategy has been also proposed in Hernandez-Garcia et al. (2014), where the authors define the grams to be considered by a graph based approach. Although being very promising, the main limitation of such approaches is that only small values of n, typically lower than 3, can be used in real time applications, thus allowing to evaluate only local temporal information (three consecutive frames) instead than global one.

In order to explicitly exploit the temporal evolution of actions, several researchers have recently explored a string based representation. In Ballan et al. (2009), for instance, the string is built as a sequence of histograms, one for each frame. A different approach has been exploited in Gaur et al. (2011), where a set of spatio-temporal interest points (STIPs) is detected from a fixed length time window; in each window, the STIPs are used to form a feature graph where nodes and edges encode respectively STIPS and their spatio-temporal distance. The string is finally built by concatenating the so obtained feature graphs. Although being very promising, such approaches are very expensive from two points of view: on one side, for each frame they need to store a large amount of information (an histogram and a graph, respectively) for encoding a single symbol of the string; on the other side, the matching between such symbols is computationally intensive. In Brun et al. (2014a) the authors propose a very compact representation of the actions based on strings and a novel similarity measure based on the Dirac kernel, which allows to overcome limitations due to the computational effort of previous string based methods.

Starting from the seminal work in Brun et al. (2014a), in this paper we propose an efficient and robust method for recognizing human actions, providing the following original contributions:

- the temporal dimension of the action is exploited by means of a very compact representation based on strings, which allows to reduce the amount of memory required for storing the past information about the action without losing relevant information;
- a novel similarity measure based on a soft global alignment kernel is proposed for managing actions of different lengths and performed at different speeds;
- an extensive experimentation is carried out over three publicly available datasets and the achieved results demonstrate the effectiveness of the proposed approach both in terms of accuracy and time required for the elaboration; furthermore, a sensitivity analysis has confirmed its robustness with respect to the configuration parameters.

The rest of the paper is organized as follows: we introduce the proposed method first focusing on its rationale in Section 2 and then providing a detailed description in Section 3; in Section 4 we report and comment the results of the experimental analysis carried out on three public datasets and the comparison with state of the art action recognition methods; finally, in Section 5, we draw final conclusions and delineate future direction of our research efforts in this area.

2. Rationale of the method

In order to model motion patterns composing actions, we introduce the concept of *aclet*, considered as the atomic unit of an action. Each aclet is encoded through a symbol corresponding to an entry in an alphabet learnt during a preliminary learning step. This alphabet is obtained from quantization of the low level feature vectors computed on the silhouette at each time instant. The main advantage deriving from the introduction of aclets to represent feature vectors is that an aclet allows to encode information associated to a frame into a single value instead of in a high dimensional feature vector, so significantly decreasing the amount of memory required to maintain the history of the action. The sequence of aclets is then used to build a high level representation based on strings, able to explicitly take into account the temporal information about the sequence of aclets. This is a very important feature, since it allows to distinguish among similar actions whose main difference pertains the sequence of sub-actions (such as, for instance, *sit down and stand up* and *stand up and sit down*).

In order to deal with the uncertainty introduced during the extraction of aclets, we introduce a novel similarity measure based on the *global alignment kernel* framework, which allows to evaluate all possible alignments between two strings. This framework has been successfully used in other applications domains, ranging from gesture recognition (Pfister et al. (2014)) to emotional expression classification (Lörincz et al. (2013)) and trajectory analysis (Brun et al. (2014b)). The main improvement of the proposed approach with respect to the method in Brun et al. (2014a) regards the typology of assignment performed between a feature vector and an aclet. In fact, in Brun et al. (2014a) an hard assignment has been considered: each feature vector is assigned to a single aclet and the similarity between two aclets is evaluated by a traditional Dirac kernel. Unfortunately, such a strategy induces a strong sensitivity to the parameters required to set up the system. In order to mitigate this problem, in this paper we introduce a novel similarity measure evaluating the *closeness* of two aclets (soft assignment) and not only if two aclets are exactly the same (hard assignment); this is achieved by measuring the similarity between two aclets using an enhanced version of the gaussian kernel, which allows to implicitly assign a single feature vector to more than one aclet, so as to increase the overall robustness of the proposed approach with respect to errors introduced during features extraction step, as well as to make it less dependent on large variations of the configuration parameters. This is due to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:https://doi.org/10.1016/j.cviu.2015.09.003, Computer Vision and Image Understanding

the fact that soft assignment can be seen as an estimation of the posteriori probability that a feature vector belongs to an aclet (Liu et al. (2011)), thus it is able to strongly mitigate the quantization error introduced by the traditional hard assignment.

Another important contribution provided by this paper with respect to Brun et al. (2014a) is a more significant experimentation carried out on three public datasets, together with a sensitivity analysis that shows the strong improvements of the proposed approach with respect to the state of the art. Furthermore, as we will show, our method is able to outperform state of the art methodologies while remaining simple to configure: indeed, only a few parameters need to be selected during the configuration steps, implying that our approach is especially suited for working in real environments where unexperienced operators may configure the application.

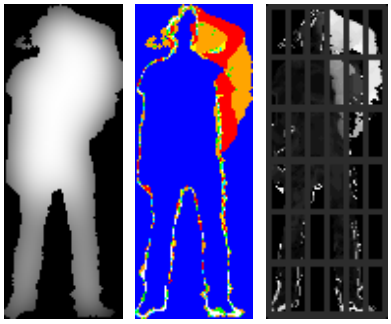
3. The proposed method

An overview of the proposed method is shown in Figure 1, where the two levels of representation used by our method are highlighted. In the first stage the low level representation is extracted by analyzing a sequence of depth images. Then, actions are modeled at the second level of the representation through strings whose symbols are named *aclets*. Each aclet is defined from feature vectors extracted in the first stage. Each string is finally used to feed a classifier, able to identify the action occurring in the video. In this section, details about each module are provided.

3.1. First-level representation

In the last years, the scientific community has deeply investigated the possibility to use depth images instead of traditional optical cameras. In fact, depth images allow to limit the loss of 3D information due to the use of traditional optical cameras: the silhouette of a person can be extracted in an easier and more reliable way since depth images are more robust to illumination changes or bad lighting conditions (Aggarwal and Xia (2014), Ye et al. (2013)). Starting from this consideration, we decided to extract the feature vector by processing depth images provided by a Kinect sensor. Here we adopt the low level description based on global features recently introduced in Carletti et al. (2013): the system first detects the human silhouette by a traditional background subtraction method (Conte et al. (2010)) applied over the depth image. Then, starting from the foreground image containing the silhouette, it computes three derived images: the *average depth image* (ADI), the *motion history image* (MHI) and the *depth difference image* (DDI). The main advantage provided by this representation is that the movement of a person in both spatial (x,y) and depth (z) dimensions can be represented in a very compact way, so significantly decreasing the amount of data to be processed without losing relevant information. For each derived image, an example is shown in Figure 2. Finally, for each frame, the feature vector is computed by analyzing ADI, MHI and DDI.

Given an observation window of N temporally adjacent images, the $ADI(x, y, t)$ (Megavannan et al. (2012)) is used to capture the motion information in the depth dimension. Let be



(a) (b) (c)

Fig. 2: An example of ADI (a), MHI (b) and DDI (c) for the action *drinking*.

(a) (b) (c) (d) (e) (f)

Fig. 3: An example of min-max depth variations features extracted with hierarchical grids for the action *walking*.

$FM(x, y, t)$ the value of the foreground mask of the pixel x, y at time t, being $FM(x, y, t) \in \{0, 1\}$, and $D(x, y, t)$ the homologous value in the depth image. $ADI(x, y, t)$ is the average depth at position (x, y) over the non zero values of the foreground mask at times $t - N + 1, \ldots, t$:

$$ADI(x, y, t) = \frac{\sum_{k=t-N+1}^{t} D(x, y, k) FM(x, y, k)}{\sum_{k=t-N+1}^{t} FM(x, y, k)}. \quad (1)$$

The MHI (Davis (2001)), is used to capture into a single static image the sequence of motions. The value of $MHI(x, y, t)$ is updated as follows:

$$MHI(x, y, t) = 255 \quad (2)$$

if point (x, y) passed from background to foreground at time t, and

$$MHI(x, y, t) = max\{MHI(x, y, t - 1) - \tau, 0\} \quad (3)$$

otherwise; τ is a constant value set to $(256/N) - 1$, corresponding to an observation over N frames.

Finally, the DDI (Megavannan et al. (2012)) is used to evaluate motions changes in the depth dimension:

$$DDI(x, y, t) = D_{max}(x, y, t) - D_{min}(x, y, t), \quad (4)$$

where $D_{max}(x, y, t)$ and $D_{min}(x, y, t)$ are respectively the maximum and minimum depth for position (x, y) over the images at times $t - N + 1, \ldots, t$. Note that only non zero pixels values of the FM across N frames are evaluated. In our experiments N has been set to one second, as it is the minimum time required by a person to move a part of his body into a meaningful way.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:https://doi.org/10.1016/j.cviu.2015.09.003, Computer Vision and Image Understanding

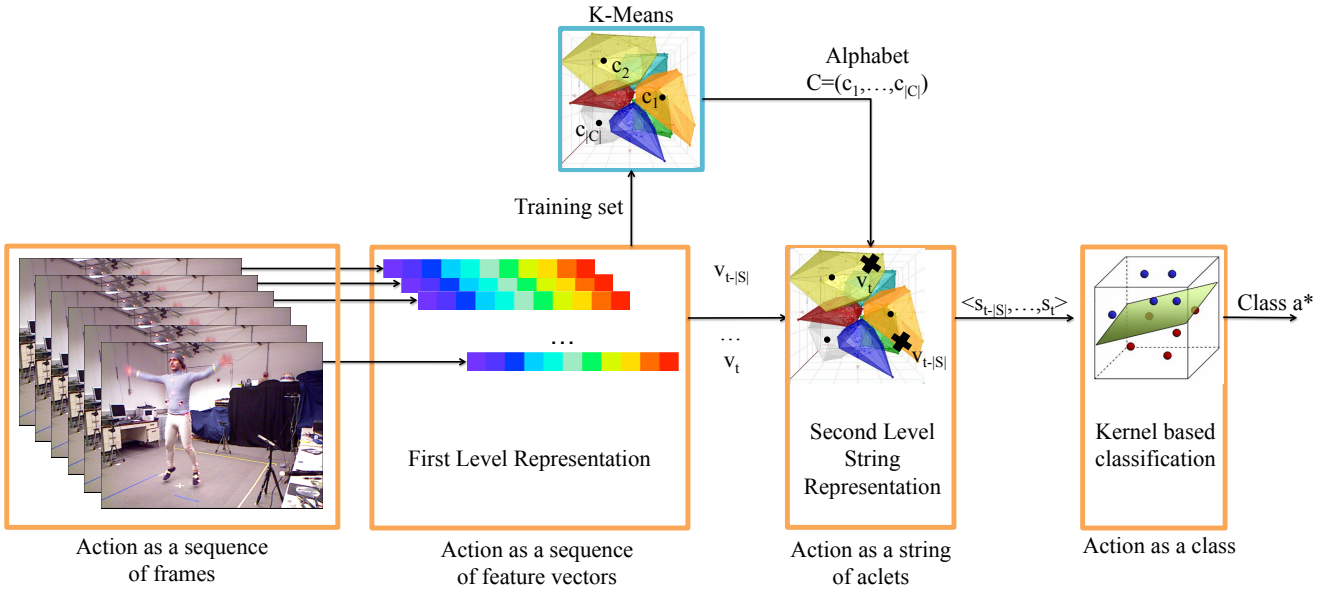


Fig. 1: An overview of the proposed method: once extracted the sequence of first level feature vectors $v_{t-|S|+1}, ..., v_t$, a high level representation $< s_{t-|S|}, ..., s_t >$ is obtained by mapping each vector into a symbol, according to the alphabet C computed during the training step. Finally, a kernel based classifier is employed in order to decide for the class a^*.

Once obtained the above derived images, we extract three different types of features. Both the *MHI* and the *ADI* are represented through the seven translation, scale and orientation invariant Hu moments (Bobick and Davis (2001)). Hu moments, computed over both *MHI* and *ADI*, result in a 14 sized feature vector.

On the other side, the *DDI* is represented through two different kinds of features: the $\Re$ transform (Tabbone et al. (2006)) and the *min-max depth Variations* (Megavannan et al. (2012)). The former is an extended Radon transform, which captures the geometric information of interest points. Its main advantage lies in the geometric invariance, to both scale and translation; furthermore, it is robust to the errors of the detection phase, such as disjoint silhouettes and holes in the shape. As experimentally demonstrated in Wang et al. (2007), $\Re$ transform outperforms methods using silhouette-based moment descriptors while requiring low computational costs.

Finally, in order to enrich the above global description, we also evaluate local information using a grid based approach: we first find the bounding box containing the silhouette in order to capture the complete motion of an action in it. Then, the box is hierarchically partitioned into cells of equal size (1x1, 2x1, 1x2, 2x2, 3x3, and 6x6) and the maximum and the minimum depth values in each cell are computed (Figure 3). The computation of this min-max depth variations results in a feature vector composed by 108 elements.

Note that the proposed description allows us to maintain all the advantages of the global representation, without suffering of the typical problems of this kind of descriptors, already discussed in the introduction: in fact, the silhouette, extracted from depth images instead of traditional intensity images, allows to significantly reduce the sensitivity to noise or partial occlusions. Furthermore, the $\Re$ transform captures local properties related to the alignment of image's subregions, while Hu moments and min-max depth variations provide a global representation of pixels' distribution.

3.2. *Second-level representation*

As outlined at the beginning of this section, an action is modeled as a string whose symbols, namely the aclets, belong to an alphabet C of finite size. The symbols in C are determined during the training phase of our method by performing a non linear quantization of the feature space by K-means clustering.

A finite alphabet of symbols, each one associated to an aclet, is thus generated by assimilating the i-th cluster to its centroid c_i:

$$C = (c_1, \ldots, c_{|C|}), \tag{5}$$

being $|C|$ the number of clusters.

It is important to highlight that, as we will show in the experimental section, the cardinality of the alphabet is not a critical parameter, as it can be chosen in a large range without significantly affecting the overall performance. Furthermore, the construction of C only requires the labeling of actions, thus limiting dramatically the human intervention in the ground truthing phase. The latter two points are noticeable as they strongly simplify the use of this method in real scenarios.

During the operating phase, the obtained alphabet C allows to determine for each low level feature vector v_i the closest cluster centroid c_j and then to associate the i-th symbol s_i:

$$s_i = \arg\min_j \|v_i - c_j\| \text{ for } j \in \{1, \ldots, |C|\}. \tag{6}$$

In order to model the fact that different actions may have different durations, for instance, the action *sit down* may takes a few seconds, while the action *boxing* may have a duration of tenth of seconds, in our approach we represent the k-th class of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:https://doi.org/10.1016/j.cviu.2015.09.003, Computer Vision and Image Understanding

action through a string of length l_k with $k \in \{1, ..., K\}$, being K the cardinality of the set of considered classes of action. Values l_k are automatically learnt during the training phase by computing the average string length associated to the actions belonging to that class.

At time t, the following strings are calculated

$$S_t^k =< s_{t-l_k}, ..., s_t >, \quad k = 1, ..., K \tag{7}$$

by concatenating the last l_k obtained symbols. The obtained strings are then used to feed a classifier and to evaluate the particular action a person is performing.

3.3. Similarity evaluation

The similarity between two actions, each represented as a string as introduced in the previous subsection, is evaluated within a kernel framework. The main advantage in the use of a kernel based approach is that kernel functions encode a similarity measure corresponding to a scalar product in some Hilbert space defined by the kernel. This last property allows to combine the rich description provided by strings with efficient machine learning methods such as SVM using the kernel trick (Aizerman et al. (1964)). This trick allows to replace scalar products by calls to our kernel in any method which can be expressed solely through scalar products.

In this paper, we propose a novel similarity metric based on the general framework of the *fast global alignment kernel* (FGAK) (Cuturi (2011)). This kernel computes a soft-minimum of all alignment scores. Hence this kernel, as argued by Cuturi (2011), takes into account a richer statistic on both strings being compared than their minimal alignment score.

Furthermore, one of the main advantages with respect to the other kernel based approaches is that it avoids the diagonal dominance of the Gram matrix, which is an undesirable property. As a matter of fact, a Gram matrix with a large diagonal dominance corresponds to a set of objects mainly similar to themselves according to kernel values. Consequently, a kernel with such kind of Gram matrix does not allow efficient generalization from training set. Furthermore, the FGAK allows to properly deal with strings having different lengths and different temporal scales.

Formally, given two strings, $X =< x_1, ..., x_n >$ and $Y =< y_1, ..., y_m >$ of length n and m respectively, an alignment between X and Y is a pair of increasing integral vectors (π_1, π_2) of length $p < n + m$, such that $1 = \pi_1(1) \leq ... \leq \pi_1(p) = n$ and $1 = \pi_2(1) \leq ... \leq \pi_2(p) = m$, with unary increments and no simultaneous repetitions. Let $A(n, m)$ be the set of all possible alignments between X and Y. The global alignment kernel k_{GA} is defined as:

$$k_{GA}(X, Y) = \sum_{\pi \in A(n,m)} \prod_{i=1}^{|\pi|} k(x_{\pi_1(i)}, y_{\pi_2(i)}, \pi_1(i), \pi_2(i)). \tag{8}$$

The kernel k that we consider in our approach is given by a combination of the triangular kernel k_t and a *soft* weighted kernel k_w:

$$k(x_i, y_j, i, j) = \frac{k_t(i, j) \cdot k_w(x_i, y_j)}{1 - k_t(i, j) \cdot k_w(x_i, y_j)}, \tag{9}$$

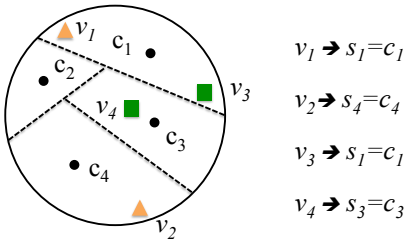


Fig. 4: Black circles $c_1, ..., c_4$ represent the cluster centroids encoding the alphabet. Triangular and squared markers $v_1, ..., v_4$ represent the feature vectors. The limitation of the traditional Dirac kernel is evident if we consider that the similarity between v_1 and v_2 is 0, equal to the one between v_3 and v_4.

where x_i and y_j encode two symbols and i and j represent the position of these symbols inside strings X and Y, respectively.

In particular, k_t has been introduced in order to speed up the computation of the kernel by considering only a small but feasible subset of all the possible alignments induced by the global alignment kernel; it allows to make the kernel computation faster, so as to obtain a *Fast* GAK: if the indices of two symbols differ by more than T, their kernel value is equal to 0. k_t is defined as follows:

$$k_t(i, j) = \left(1 - \frac{|i - j|}{T}\right)_+, \tag{10}$$

where T is the order of the kernel and + refers to the fact that $k_t(i, j) = 0$ if $|i - j| \geq T$.

The weighted kernel k_w on the other side, implicitly induces the soft assignment and allows to avoid the problems related to the *hard* evaluation of the traditional Dirac kernel $\delta(x_i, y_i)$, used in Brun et al. (2014a), and defined as:

$$\delta(x_i, y_i) = \begin{cases} 0 & \text{if } x_i \neq y_i \\ 1 & \text{if } x_i = y_i \end{cases} \tag{11}$$

As we can see, the traditional Dirac kernel does not take into account how close two clusters are. In order to better understand this problem, we can consider the following example. Let us consider the alphabet C shown in Figure 4. The similarity between symbols s_1 and s_4, associated to feature vectors v_1 and v_2 respectively, is equal to 0, since $\delta(s_1, s_4) = 0$. Similarly, we can note that $\delta(s_1, s_3) = 0$, being s_1 and s_3 the symbols associated to the feature vectors v_3 and v_4 respectively. However, although belonging to two different clusters, v_3 and v_4 are located in two close clusters and their similarity shouldn't be null.

Starting from the above considerations, in this paper we introduce a novel kernel with a double purpose: in fact, on one side, we should avoid the crisp behavior induced by the hard evaluation of the Dirac kernel; on the other side, we can not make its computation too heavy so as to still allow a computation in real time. For these reasons, we propose a novel similarity metric based on k_w, that is an enhanced version of the traditional gaussian kernel (Cuturi (2011)). Let be $d(x_i, y_i)$ the distance between the cluster centroids c_{x_i} and c_{y_i}, associated to the symbols x_i and y_i, respectively:

$$d(x_i, y_i) = \|c_{x_i} - c_{y_i}\|^2. \tag{12}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:https://doi.org/10.1016/j.cviu.2015.09.003, Computer Vision and Image Understanding

The kernel $k_w(x_i, y_i)$ can be computed as follows:

$$k_w(x_i, y_i) = e^{-\phi_\sigma(x_i, y_i)}, \tag{13}$$

where:

$$\phi_\sigma(x_i, y_i) = \frac{1}{2\sigma^2} d(x_i, y_i) + \log\left(2 - e^{-\frac{d(x_i, y_i)}{2\sigma^2}}\right). \tag{14}$$

This kernel performs a soft evaluation, since it is based on the distance between cluster centroids. Such distances can be computed only once, during the setup of the system, so significantly decreasing the computational effort required by the use of the kernel at run time.

It is important to highlight that the proposed kernel is a Mercer kernel. In fact, as demonstrated in Cuturi et al. (2006), k_{GA} is positive definite if both k and $k/(1 + k)$ are positive definite. Using Equation 9 our kernel k may be written as:

$$k(x, y, i, j) = \sum_{n=1}^{+\infty} (k_t(i, j) k_w(x, y))^n$$

and is thus definite positive as the limit of a sum of definite positive kernels. Moreover:

$$\frac{k(x, y, i, j)}{1 + k(x, y, i, j)} = k_t(i, j) k_w(x, y)$$

is trivially definite positive as a tensor product of definite positive kernels. Hence (Cuturi (2011), Remark 1), our global kernel k_{GA} is definite positive.This is a an important property: indeed, as described in Lei and Sun (2007), positive definite kernels usually strongly outperform not positive definite ones in kernel machines (for instance in combination with support vector machines).

The kernel defined by Equation 8 is finally normalized so as to obtain a similarity value k^N_{GA} in the interval $[0, 1]$:

$$k^N_{GA}(X, Y) = \frac{k_{GA}(X, Y)}{\sqrt{k_{GA}(X, X) * k_{GA}(Y, Y)}}. \tag{15}$$

Computational Cost: it can be shown that the cost for computing k_{GA} is $O(T \min(m, n))$ where T is the order of kernel k_t (Equation 10 and Cuturi (2011)). This cost consitutes only a marginal overhead with respect to the computation of the low level processing module, which instead depends on the original image resolution, which is significantly higher.

3.4. The classifier

The aim of the proposed method is to identify a particular action a^* occurring in a video stream, among the $|A|$ different classes a_i the system has been trained on, being $i = 1, ..., |A|$.

In order to confirm the generality of the proposed approach with respect to different classification paradigms, we considered two different well-known and widely adopted classifiers, namely a multi-class support vector machine (SVM) and a k nearest neighbors (k-NN). As for the SVM, we design $|A|$ different classifiers, one for each class, using a one versus all approach. The i-th classifier is trained on the whole training data set in order to classify the members of i-th class against the rest. It means that the training set is relabeled: the samples belonging to the i-th class are labeled as positive examples, while samples belonging to other classes are labeled as negative ones. During the operating phase, a new string is assigned to the class whose distance to the margin is maximum.

As for the k-NN, during the training step the different strings encoding the prototypes are extracted, one for each action. During the operating phase, the strings S^k_t representing each action at the generic time t are compared with the prototypes using the proposed kernel and the decision is finally taken by a majority vote of their K nearest neighbors. The number of nearest neighbors has been experimentally set to 5.

4. Experimental results

In this section we describe the results of the performance assessment of the proposed method. In particular, we first introduce the datasets adopted for validating the action recognition method: each dataset is analyzed in terms of the classes of actions proposed for the tests and the kind of difficulties arising from their use. Successively, we focus on the adopted experimental protocol; we define the indices that are used for measuring the performance and evaluating the robustness of the method with respect to the set up parameters. We also compare the proposed method with several state of the art methods on the same datasets.We conclude this section by a discussion on the advantages and limitations of our method.

4.1. Datasets description

The proposed method has been tested over three public datasets: the Berkeley multimodal human action detection dataset (hereinafter MHAD), the MIVIA dataset and the Cornel activity dataset (hereinafter CAD). The first two datasets provide depth images and, for each sequence, the corresponding background, while the third one does not provide the backgrounds. Thus, in this case, silhouettes are extracted by filtering depth images using an evaluation of the distance between persons and the camera. In all the datasets the subjects were asked to perform a given action, without detailed instructions about how to do it, as well as about the time and the execution rate. Thus, different styles in performing the actions have been applied, so making the datasets very challenging.

The MHAD dataset, by Ofli et al. (2013), is based on actions with movements in both upper and lower extremities (such as *jumping in place*), actions with high dynamics in upper extremities (such as *waving hands*) and finally actions with high dynamics in lower extremities (such as *sit down*). It contains 11 actions performed by 7 male and 5 female subjects. All the subjects performed 5 repetitions of each action, yielding about 660 action sequences which correspond to approximatively 82 minutes of total recording time. The main difficulty of this dataset is that several movements share common sub-movements. For instance, actions *sit down* and *stand up* are a part of the action *sit down then stand up*; the jumping movement is also a part of both actions *jumping in place* and *jumping jacks*. It implies that discriminating such actions may be a very difficult task and thus

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:https://doi.org/10.1016/j.cviu.2015.09.003, Computer Vision and Image Understanding

(a) Jumping in place (b) Jumping jacks (c) Waving - one hand (right) (d) Clapping

(e) Opening a jar (f) Drinking (g) Free Movements (h) Interacting with a table

(i) Brushing teeth (j) Cooking (chopping) (k) Wearing Contact lenses (l) Working on computer

Fig. 5: Some examples from the MHAD (a-d), from the MIVIA (e-h) and from the CAD (i-l) datasets.

may strongly influence the overall performance of the methods for action recognition.

The MIVIA dataset, by Foggia et al. (2013), is composed of 7 actions performed by 14 subjects, namely 7 males and 7 females. All subjects performed 2 repetitions of each action. This dataset mostly contains actions with movement in upper extremities (such as *drinking* or *interacting with the table*). It is important to underline that, although being composed by a lower number of classes (7 vs 11), the MIVIA dataset also includes *free movements*: in this case, persons have been asked to freely move inside the scene. It is evident that this kind of action is difficult to recognize, especially by a string kernel based approach, which implicitly considers the similarity between actions by evaluating the sequence of aclets.

Finally, the CAD dataset, by Sung et al. (2012), is composed by 12 actions performed by 4 subjects (two males, two females, one of them left-handed). Differently from the other considered datasets, it has not been acquired in a laboratory but instead in five different environments: office, kitchen, bedroom, bathroom and living room. For each action, around 45 seconds of data for each person have been acquired, so resulting in more than 30 minutes of video recording.

Few examples from the datasets are shown in Figure 5, while more information about actions are reported in Tables 1 and 2.

4.2. Performance assessment

The tests over the three datasets have been carried out adopting a leave-one-out cross validation strategy: the training set is composed by all the repetitions of $n-1$ persons and the repetitions of the remaining person have been used to test the system. Finally, the average performance has been reported. In this way, the bias due to the particular way in which a single person performs an action is avoided.

During the configuration phase, the user needs to set up the following parameters: the number of clusters $|C|$ and the parameters of the kernel, namely T and σ.

As suggested in Cuturi (2011), the value of σ is related to both the number of alignments, which are of the order of median length $\|X\|$, and the values of ϕ_σ, which are of the order of median distances between cluster centroids $\|C_X - C_Y\|$. Thus, it can be set up as follows:

$$\sigma \in \alpha_\sigma \cdot \widetilde{\|C_X - C_Y\|} \sqrt{\widetilde{\|\tilde{X}\|}}, \tag{16}$$

being α_σ set to 2 in our experiments as suggested in Cuturi (2011). Starting from this consideration, we fixed σ and performed a sensitivity analysis by evaluating the robustness of the proposed approach with respect to large variations of $|C|$ and T. Results obtained on the MHAD, MIVIA and CAD datasets are shown in Figure 6. The figure is organized as follows: in the first row (a,b,c) results obtained by the traditional Dirac kernel (Brun et al. (2014a)) are reported; the second and the third rows show the proposed weighted kernel with a k-NN classifier (d,e,f) and with a SVM classifier (g,h,i), respectively. All figures show the $|C|$ values on the x-axis and the true positive rate on the y-axis; different curves in the same figure represent performance achieved using different values of T. For each dataset and each method, mean and variance values of all curves are finally summarized in (j,k,l).

Results in Figure 6 highlight the great robustness of the pro-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:https://doi.org/10.1016/j.cviu.2015.09.003, Computer Vision and Image Understanding

posed method with respect to variations of $|C|$ and T over all the datasets. This is a very important feature that allows to largely simplify the configuration procedure. We can also note that the performance obtained using the proposed approach does not depend on the employed classification architecture as very similar performance are achieved using the k-NN and the SVM classifier, thus confirming the good discriminative power of the proposed similarity measure. This observation appears evident by considering results reported in Figures 6j, 6k and 6l, which report a summary of the sensitivity analysis carried out on the method. In particular, the values of mean and variance of the true positive rates are shown. The best average results are achieved on the three datasets using the proposed kernel combined with a k-NN classifier (91.26%, 93.52% and 73.59% for MHAD, MIVIA and CAD datasets, respectively). Moreover, the SVM method provides good performances, since it maintains a very low variance, still comparable with the one of the k-NN, and a high average true positive rate(89.70%, 92.56% and 73.28% for MHAD, MIVIA and CAD datasets, respectively). It is also important to highlight that our kernel, with both classification schemas and over all the datasets, significantly reduces the variance introduced by varying the parameters (one order of magnitude, from 10^{-4} to 10^{-5} for MHAD and MIVIA datasets and from 10^{-3} to 10^{-4} for CAD dataset) with respect to Brun et al. (2014a).

4.3. Comparison with state of the art

In order to further confirm the effectiveness of our approach with respect to the state of the art, the accuracy for each class is reported in Tables 1 and 2. As for MHAD and MIVIA dataset, the performance refers to the following configuration: $|C| = 512$ and $|T| = 65$, while as for the CAD dataset, we consider $|C| = 32$ and $|T| = 25$. The large difference between the values of C over the first two datasets and the CAD dataset ($|C| = 512$ and $|C| = 32$, respectively) is motivated by the fact that in the CAD dataset the actions are performed by each single actor sequentially, so reducing the intra class variability per actor, while on the contrary on the remaining two datasets the repetitions are performed in different moments, so increasing the intra class variability. Consequently, in the last case a higher number of clusters is needed to achieve a good generalization, while the same number of clusters would provide a higher degree of specialization on the CAD dataset.

We can note that in all the datasets our method is able to correctly recognize most actions. However, it is interesting to analyze how our approach deals with the action *free movement* belonging to the MIVIA dataset, where the obtained accuracy is 85.5%; this result is due to the fact that, as anticipated before, *free movement* is a very challenging action since the evaluation of the temporal information can not increase the reliability in the recognition process.

It is also interesting to note that actions achieving lowest performance in the MHAD dataset are *waving one or two hands* and *clapping hands*. It is mainly due to the fact that in these cases only a very localized part of the body (one or two hands) is moving; it implies that at a low level it is difficult to find out a good feature set, able to better discriminate these two typologies of actions. This consideration is also confirmed by the fact that among the state of the art methodologies reported in Table 1, the best accuracy over such actions has been achieved by Foggia et al. (2014), which uses a method based on a deep architecture to automatically learn the representation.

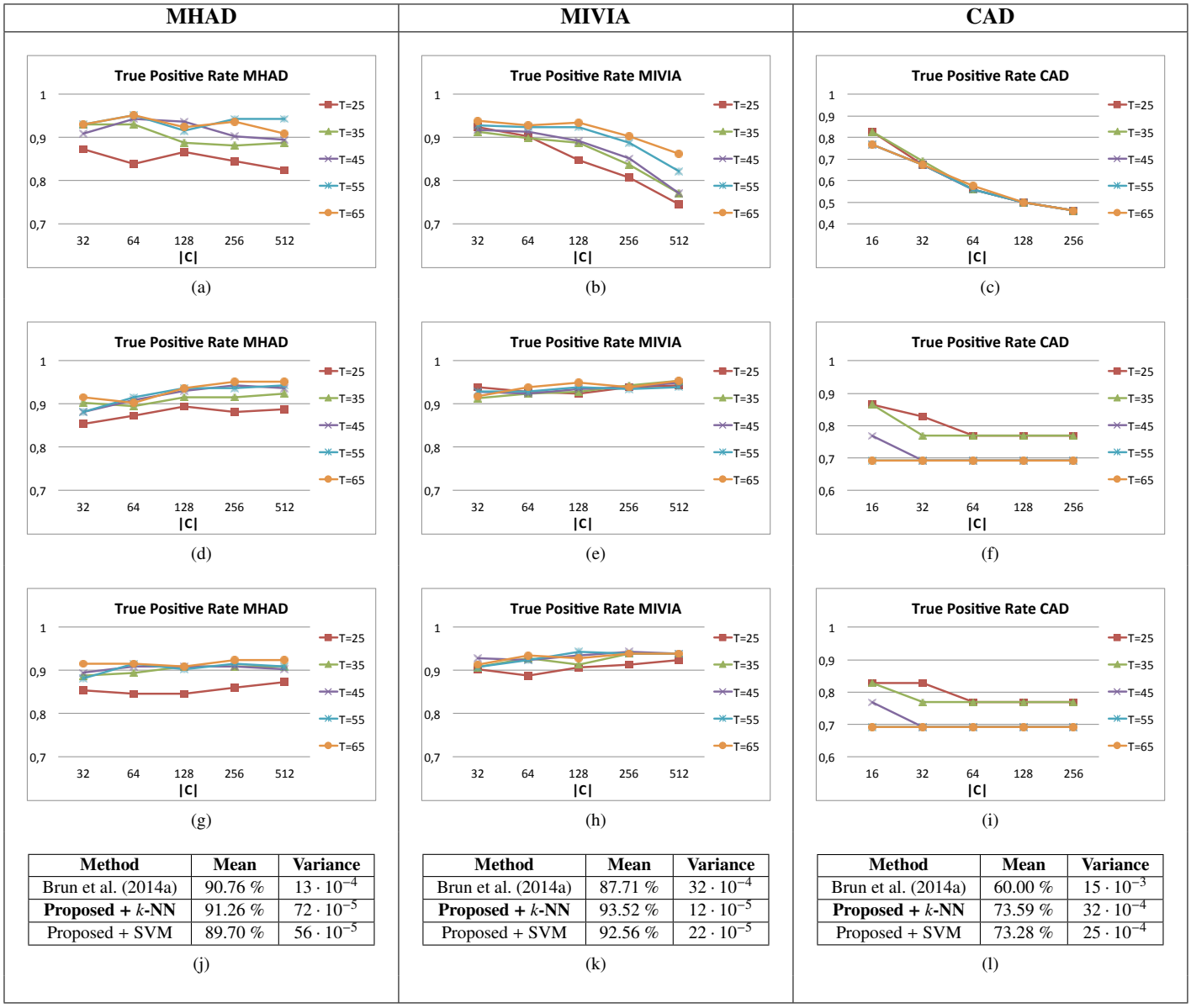
As for the CAD dataset, we can note that most of the actions (8 over 13) are perfectly recognized; as for the remaining ones, we can note that they mainly involve a very limited part of the body, thus making difficult, as in the case of MHAD dataset, the definition of a feature set able to perfectly discriminate such actions.

Table 1 also reports results achieved by other state of the art methodologies. The considered methods are based both over local (Dollar et al. (2005)) and global (Foggia et al. (2014), Foggia et al. (2013), Carletti et al. (2013)) descriptors and some of them exploit a high level representation of the data (Foggia et al. (2014), Foggia et al. (2013)). Our approach outperforms all the considered methodologies in 8 over 11 classes for the MHAD dataset and in 6 over 7 for the MIVIA dataset, while obtaining the higher average accuracy over the latter two datasets (95.1% and 95.4% for the MHAD and MIVIA datasets, respectively). Such results confirm the high effectiveness of the proposed approach, able to overcome several methodologies, based on different typologies of features as well as different high level representations.

A more compact comparison is reported in Table 3; note that only methods based on *reject option*, *bag of words*, *cuboids*, *BM1*, *edit distance*, *deep learning*,*dynamic time warping* and *K-mer kernel* listed in Table 3 are based on visual information, while the remaining approaches are based on the skeleton acquired by a Mocap system (the optical motion capture system Impulse, able to capture 3D position of active LED markers): it is clear that the introduction of this kind of marker, even if improving the performance of action recognition algorithms, strongly limits its applicability to real environments. Although we extract the first level feature vector from depth images instead than from Mocap system, we can note that our method outperforms all the other considered approaches, both based on traditional intensity images, depth images and Mocap systems. This is a very encouraging result, since it confirms the effectiveness of the proposed approach also compared with several state of the art methodologies.

Data in Table 3 highlight another important result related to the accuracy that can be achieved using different string based approaches as the one proposed in this paper and those using dynamic time warping or the K-mer string kernel, the latter widely adopted for the classification of protein sequence data (Leslie and Kuang (2004)). In order to fairly compare the above string based methods for the tests we considered the same low level and high level representations, and also the same classifier (the k-NN), while we change only the metric used to evaluate the similarity between strings. The improvement achieved by our method using the proposed kernel with respect to the same method using the dynamic time warping or the K-mer kernel (95.1% vs 79.9% and 62.3% over the MHAD, 95.4% vs 79.8% and 87.5% over the MIVIA, 86.5% vs 57.7% and 54.2 over the CAD) confirms the effectiveness of the approach. We deem that the explanation of these results is that our similarity measure

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:https://doi.org/10.1016/j.cviu.2015.09.003, Computer Vision and Image Understanding



Method	Mean	Variance
Brun et al. (2014a)	90.76 %	$13 \cdot 10^{-4}$
Proposed + *k*-NN	91.26 %	$72 \cdot 10^{-5}$
Proposed + SVM	89.70 %	$56 \cdot 10^{-5}$

(j)

Method	Mean	Variance
Brun et al. (2014a)	87.71 %	$32 \cdot 10^{-4}$
Proposed + *k*-NN	93.52 %	$12 \cdot 10^{-5}$
Proposed + SVM	92.56 %	$22 \cdot 10^{-5}$

(k)

Method	Mean	Variance
Brun et al. (2014a)	60.00 %	$15 \cdot 10^{-3}$
Proposed + *k*-NN	73.59 %	$32 \cdot 10^{-4}$
Proposed + SVM	73.28 %	$25 \cdot 10^{-4}$

(l)

Fig. 6: True positive rate computed over the MHAD (a,d,g,j), MIVIA (b,e,h,k) and CAD (c,f,i,l) datasets by varying the $|C|$ parameter by using (a,b,c) Dirac kernel (Brun et al. (2014a)), (d,e,f) weighted kernel with a k-NN classifier and (g,h,i) with a SVM classifier. In (j,k,l) mean and variance values are reported for each considered method.

uses a soft evaluation, revealing to be more robust with respect to the noise introduced during the features extraction step as well as during the assignment of a symbol to a feature vector.

4.4. Discussion

One of the strengths of the proposed approach lies in its generality and its robustness, confirmed by the fact that it is able to achieve very high performance in several conditions, both by varying the configuration parameters (T and $|C|$) and the classifier (SVM and k-NN). This consideration is furthermore confirmed if we note that the results are similar for all the datasets containing different typologies of actions and acquired in different environments. In fact, we obtain on average an accuracy of 91.26%, 93.52% and 86.50% over the MHAD, the MIVIA and the CAD datasets, respectively. Note that the proposed approach is optimized for working with human silhouettes, thus being provided with the background of the scene. Even if this is not available for the CAD dataset and thus the extraction of the silhouettes as well as of the low level features from such images are affected by noise, the performances of the proposed approach are very promising, especially if compared with state of the art methodologies.

It is also important to highlight that all these advantages are not paid in terms of computational cost. In fact, our method is able to work in real time: over a MacBook Pro equipped with Intel Core 2 Duo running at 2.4 GHz, the system processes during the operating phase 1 CIF video streams at 10 frame per seconds.

Finally, the comparison of the proposed approach with respect to state of the art methodologies confirms the accuracy of our method, since it achieves the highest performances over

three publicly available datasets with respect to methods relying on different information sources(depth image or Mocap system), different low level feature vectors (local and global descriptors) and different high level representations (bag of words and deep representation).

CAD	
Action	**Proposed + k-NN**
Talking on the phone	100.0
Drinking water	100.0
Talking on couch	75.0
Relaxing on couch	100.0
Opening pill container	100.0
Working on computer	100.0
Cooking (stirring)	100.0
Cooking (chopping)	75.0
Writing on whiteboard	50.0
Brushing teeth	75.0
Wearing contact lenses	100.0
Rinsing mouth with water	50.0
Still	100.0
Average Accuracy	**86.50**

Table 2: Recognition accuracy for each class of actions achieved by the proposed method over the CAD dataset. All recordings of the dataset have a duration of 45 secs.

5. Conclusion

In this paper we have presented a new method for recognizing human actions. Differently from most methods in the literature which pay their high accuracy in terms of computational cost or difficulties in the parameters set up, our method runs in real time and is at the same time robust with respect to large variations of configuration parameters, making it particularly suited for working in real applications.

These advantages have been achieved thanks to the following main contributions: (i) starting from our preliminary work (Brun et al. (2014a)), we use a string based representation which allows to explicitly evaluate the temporal evolution of aclets composing the actions; (ii) the similarity between two actions is evaluated within a kernel based framework, which makes the system robust with respect to actions having different temporal lengths and different rates; (iii) the similarity between two aclets is based on an enhanced version of the gaussian kernel, which implicitly introduces a kind of soft evaluation, since it allows to analyze how two aclets are close each other.

Acknowledgment

This research has been partially supported by A.I.Tech s.r.l. (http://www.aitech.vision) and by POR Campania FSE 2007/2013, *Bando Sviluppo di Reti di eccellenza* within the project Embedded Systems in Critical Domains.

MHAD			
Method	**Source**	**Year**	**Accuracy**
Proposed	-	-	**95.1**
Dynamic time warping	-	-	79.9
K-mer kernel	-	-	62.3
Bag of words	Foggia et al. (2013)	2013	72.9
Edit sistance	Brun et al. (2015)	2015	87.1
BM1	Cheema et al. (2013)	2013	77.7
Deep learning	Foggia et al. (2014)	2014	85.8
Cuboids	Dollar et al. (2005)	2005	66.2
Reject option	Carletti et al. (2013)	2013	57.4
SMIJ	Ofli et al. (2012)	2012	94.2
HMIJ	Ofli et al. (2012)	2012	82.9
HMW	Ofli et al. (2012)	2012	81.1
LDSP	Ofli et al. (2012)	2012	82.2
BM2	Cheema et al. (2013)	2013	87.8

(a)

MIVIA			
Method	**Source**	**Year**	**Accuracy**
Proposed	-	-	**95.4**
Dynamic time warping	-	-	79.8
K-mer kernel	-	-	87.5
Cuboids	Dollar et al. (2005)	2005	74.6
Reject option	Carletti et al. (2013)	2013	82.5
Bag of words	Foggia et al. (2013)	2013	83.0
HAcK	Brun et al. (2014a)	2014	93.9
Deep learning	Foggia et al. (2014)	2014	84.7
Edit distance	Brun et al. (2015)	2015	85.2

(b)

CAD			
Method	**Source**	**Year**	**Accuracy**
Proposed	-	-	**86.5**
Dynamic time warping	-	-	57.7
K-mer kernel	-	-	54.2
HAcK	Brun et al. (2014a)	2014	82.7
Edit distance	Brun et al. (2015)	2015	63.5
Sparse coding	Ni et al. (2012)	2012	65.3
MKL	Wang et al. (2014)	2014	74.7
MEMM	Sung et al. (2012)	2012	51.3

(c)

Table 3: Comparison with state of the art approaches over the MHAD dataset (a), over the MIVIA dataset (b) and over the CAD dataset (c). The first three rows of each table report the performance of the our method using the proposed kernel, dynamic time warping and the K-mer kernel, respectively.

References

Aggarwal, J., Ryoo, M., 2011. Human activity analysis: A review. ACM Comput. Surv. 43, 16:1–16:43. URL: http://doi.acm.org/10.1145/1922649.1922653, doi:10.1145/1922649.1922653.

Aggarwal, J., Xia, L., 2014. Human activity recognition from 3d data: A review. Pattern Recogn Lett , –doi:http://dx.doi.org/10.1016/j.patrec.2014.04.011.

Aizerman, M.A., Braverman, E.A., Rozonoer, L., 1964. Theoretical foundations of the potential function method in pattern recognition learning., in: Automation and Remote Control, pp. 821–837.

Ballan, L., Bertini, M., Del Bimbo, A., Serra, G., 2009. Action categorization in soccer videos using string kernels, in: Content-Based Multimedia Indexing, 2009. CBMI '09. Seventh International Workshop on, pp. 13–18. doi:10.1109/CBMI.2009.10.

Bettadapura, V., Schindler, G., Ploetz, T., Essa, I., 2013. Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition, in: IEEE CVPR, pp. 2619–2626. doi:10.1109/CVPR.2013.338.

Bobick, A.F., Davis, J.W., 2001. The recognition of human movement us-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI:https://doi.org/10.1016/j.cviu.2015.09.003, Computer Vision and Image Understanding

MHAD						
Action	**Length (secs)**	**Proposed**	**Foggia et al. (2014)**	**Foggia et al. (2013)**	**Carletti et al. (2013)**	**Dollar et al. (2005)**
Jumping in place	5	**100.0**	78.0	75.0	74.9	44.9
Jumping jacks	7	**100.0**	92.8	96.4	58.9	63.7
Bending - hands up all the way down	12	**100.0**	90.8	60.7	47.5	73.9
Punching (boxing)	10	**100.0**	83.2	92.9	67.9	66.9
Waving - two hands	7	80.6	**94.9**	58.9	70.2	72.2
Waving - one hand (right)	7	83.7	**88.5**	75	35.5	75.0
Clapping hands	5	78.7	**84.6**	60.7	59.9	63.3
Throwing a ball	3	**100.0**	47.4	75.0	48.9	32.5
Sit down then stand up	15	**100.0**	95.5	75.0	57.3	77.0
Sit down	2	**100.0**	32.7	76.8	79.5	10.8
Stand up	2	**100.0**	28.7	82.1	64.9	14.0
Average Accuracy		**95.1**	85.8	72.9	57.4	66.2

(a)

MIVIA						
Action	**Length (secs)**	**Proposed**	**Foggia et al. (2014)**	**Foggia et al. (2013)**	**Carletti et al. (2013)**	**Dollar et al. (2005)**
Opening a jar	2	**100.0**	73.3	67.9	73.3	49.6
Drinking	3	**100.0**	70.4	53.6	73.9	53.4
Sleeping	3	**100.0**	76.8	98.2	65.7	84.6
Free Movements	11	85.5	96.8	75.0	**99.0**	79.6
Stopping	7	**100.0**	87.3	**100.0**	62.9	60.0
Interacting with a table	3	**100.0**	85.3	**100.0**	94.0	84.6
Sitting	3	**100.0**	72.7	82.1	80.9	89.9
Average Accuracy		**95.4**	84.7	83.0	82.5	74.6

(b)

Table 1: Average Length per recording (in seconds) and accuracy rate of the single actions tested over both the MHAD (a) and the MIVIA (b) datasets. The obtained results are compared with Foggia et al. (2014) Foggia et al. (2013), Carletti et al. (2013) and Dollar et al. (2005).

ing temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. 23, 257–267. URL: http://dx.doi.org/10.1109/34.910878, doi:10.1109/34.910878.

Brun, L., Foggia, P., Saggese, A., Vento, M., 2015. Recognition of human actions using edit distance on aclets strings, in: VISAPP 2015.

Brun, L., Percannella, G., Saggese, A., Vento, M., 2014a. Hack: Recognition of human actions by kernels of visual strings, in: IEEE AVSS 2014.

Brun, L., Saggese, A., Vento, M., 2014b. Dynamic scene understanding for behavior analysis based on string kernels. IEEE Trans. Circuits Syst. Video Technol. doi:10.1109/TCSVT.2014.2302521.

Carletti, V., Foggia, P., Percannella, G., Saggese, A., Vento, M., 2013. Recognition of human actions from rgb-d videos using a reject option, in: ICIAP 2013. volume 8158, pp. 436–445. URL: http://dx.doi.org/10.1007/978-3-642-41190-8_47, doi:10.1007/978-3-642-41190-8_47.

Chaquet, J.M., Carmona, E.J., Fernndez-Caballero, A., 2013. A survey of video datasets for human action and activity recognition. Comput Vis Image Und 117, 633 – 659. doi:http://dx.doi.org/10.1016/j.cviu.2013.01.013.

Cheema, M.S., Eweiwi, A., Bauckhage, C., 2013. Human activity recognition by separating style and content. Pattern Recogn Lett , –URL: http://www.sciencedirect.com/science/article/pii/S0167865513003607, doi:http://dx.doi.org/10.1016/j.patrec.2013.09.024.

Chen, Y., Wu, Q., He, X., 2011. Human action recognition based on radon transform, in: Multimedia Analysis, Processing and Communications, Springer Berlin Heidelberg. pp. 369–389. doi:10.1007/978-3-642-19551-8_13.

Conte, D., Foggia, P., Percannella, G., Tufano, F., Vento, M., 2010. An experimental evaluation of foreground detection algorithms in real scenes. EURASIP Journal on Advances in Signal Processing 2010, 373941. URL: http://asp.eurasipjournals.com/content/2010/1/373941, doi:10.1155/2010/373941.

Cuturi, M., 2011. Fast global alignment kernels, in: Getoor, L., Scheffer, T. (Eds.), ICML, ACM. pp. 929–936.

Cuturi, M., Vert, J.P., Birkenes, O., Matsui, T., 2006. A kernel for time series based on global alignments. CoRR abs/cs/0610033, 413–416.

Davis, J., 2001. Hierarchical motion history images for recognizing human motion, in: Detection and Recognition of Events in Video, 2001. Proceedings. IEEE Workshop on, pp. 39 –46. doi:10.1109/EVENT.2001.938864.

Dollar, P., Rabaud, V., Cottrell, G., Belongie, S., 2005. Behavior recognition via sparse spatio-temporal features, in: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72. doi:10.1109/VSPETS.2005.1570899.

Foggia, P., Percannella, G., Saggese, A., Vento, M., 2013. Recognizing human actions by a bag of visual words, in: IEEE SMC 2013.

Foggia, P., Saggese, A., Strisciuglio, N., Vento, M., 2014. Exploiting the deep learning paradigm for recognizing human actions, in: IEEE AVSS 2014.

Gaur, U., Zhu, Y., Song, B., Roy-Chowdhury, A., 2011. A "string of feature graphs" model for recognition of complex activities in natural videos, in: ICCV, IEEE Computer Society, Washington, DC, USA. pp. 2595–2602. doi:10.1109/ICCV.2011.6126548.

Hernandez-Garcia, R., Garcia-Reyes, E., Ramos-Cozar, J., Guil, N., 2014. Human action classification using n-grams visual vocabulary, in: Bayro-Corrochano, E., Hancock, E. (Eds.), Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer International Publishing. volume 8827 of *Lecture Notes in Computer Science*, pp. 319–326. URL: http://dx.doi.org/10.1007/978-3-319-12568-8_39, doi:10.1007/978-3-319-12568-8_39.

Hu, W., Tan, T., Wang, L., Maybank, S., 2004. A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst., Man, Cybern. C 34, 334–352. doi:10.1109/TSMCC.2004.829274.

Jhuang, H., Serre, T., Wolf, L., Poggio, T., 2007. A biologically inspired system for action recognition, in: IEEE ICCV, pp. 1–8.

Jiang, Z., Lin, Z., Davis, L., 2012. Recognizing human actions by learning and matching shape-motion prototype trees. IEEE Trans. Pattern Anal. Mach. Intell. 34, 533–547. doi:10.1109/TPAMI.2011.147.

Jiang, Z., Lin, Z., Davis, L.S., 2013. A unified tree-based framework for joint action localization, recognition and segmentation. Comput Vis Image Und 117, 1345 – 1355. URL: http://www.sciencedirect.com/science/article/pii/S1077314212001749, doi:http://dx.doi.org/10.1016/j.cviu.2012.09.008.

Johansson, G., 1973. Visual perception of biological motion and a model for its analysis. Perception and Psychophysics 14, 201–211. URL: http://dx.doi.org/10.3758/BF03212378.

Kellokumpu, V., Zhao, G., PietikŁinen, M., 2011. Recognition of human actions using texture descriptors. Mach. Vis. Appl. 22, 767–780.

Kovashka, A., Grauman, K., 2010. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition, in: IEEE CVPR, pp. 2046–2053. doi:10.1109/CVPR.2010.5539881.

Lee, H., Morariu, V., Davis, L., 2014. Robust pose features for action recognition, in: IEEE CVPRW, pp. 365–372. doi:10.1109/CVPRW.2014.60.

Lei, H., Sun, B., 2007. A study on the dynamic time warping in kernel machines, in: IEEE SITIS, pp. 839–845. doi:10.1109/SITIS.2007.112.

Leslie, C., Kuang, R., 2004. Fast string kernels using inexact matching for protein sequences. J. Mach. Learn. Res. 5, 1435–1455. URL: http://dl.acm.org/citation.cfm?id=1005332.1044708.

Liu, L., Wang, L., Liu, X., 2011. In defense of soft-assignment coding, in: Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 2486–2493. doi:10.1109/ICCV.2011.6126534.

Lörincz, A., Jeni, L.A., Szabó, Z., Cohn, J.F., Kanade, T., 2013. Emotional expression classification using time-series kernels. CoRR abs/1306.1913. URL: http://arxiv.org/abs/1306.1913.

Megavannan, V., Agarwal, B., Babu, R., 2012. Human action recognition using depth maps, in: SPCOM 2012, pp. 1 –5. doi:10.1109/SPCOM.2012.6290032.

Mitra, S., Acharya, T., 2007. Gesture recognition: A survey. IEEE Trans. Syst., Man, Cybern. C 37, 311–324. doi:10.1109/TSMCC.2007.893280.

Moeslund, T.B., Hilton, A., Krüger, V., 2006. A survey of advances in vision-based human motion capture and analysis. Comput Vis Image Und 104, 90–126. URL: http://linkinghub.elsevier.com/retrieve/pii/S1077314206001263, doi:10.1016/j.cviu.2006.08.002.

Ni, B., Moulin, P., Yan, S., 2012. Order-preserving sparse coding for sequence classification, in: ECCV, pp. 173–187. doi:10.1007/978-3-642-33709-3_13.

Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R., 2012. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition, in: IEEE CVPRW, pp. 8–13. doi:10.1109/CVPRW.2012.6239231.

Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R., 2013. Berkeley mhad: A comprehensive multimodal human action database, in: WACV.

Pfister, T., Charles, J., Zisserman, A., 2014. Domain-adaptive discriminative one-shot learning of gestures, in: ECCV 2014. Springer International Publishing. volume 8694 of *LNCS*, pp. 814–829. URL: http://dx.doi.org/10.1007/978-3-319-10599-4_52, doi:10.1007/978-3-319-10599-4_52.

Poppe, R., 2010. A survey on vision-based human action recognition. Image Vision Comput. 28, 976–990. URL: http://dx.doi.org/10.1016/j.imavis.2009.11.014, doi:10.1016/j.imavis.2009.11.014.

Shukla, P., Biswas, K.K., Kalra, P.K., 2013. Action recognition using temporal bag-of-words from depth maps, in: IAPR MVA 2013, pp. 41–44. URL: http://www.mva-org.jp/Proceedings/2013USB/papers/04-02.pdf.

Sung, J., Ponce, C., Selman, B., Saxena, A., 2012. Unstructured human activity detection from rgbd images., in: ICRA, IEEE. pp. 842–849.

Tabbone, S., Wendling, L., Salmon, J.P., 2006. A new shape descriptor defined on the radon transform. Comput. Vis. Image Underst. 102, 42–51. URL: http://dx.doi.org/10.1016/j.cviu.2005.06.005, doi:10.1016/j.cviu.2005.06.005.

Turaga, P., Chellappa, R., Subrahmanian, V., Udrea, O., 2008. Machine recognition of human activities: A survey. IEEE Trans. Circuits Syst. Video Technol. 18, 1473–1488. doi:10.1109/TCSVT.2008.2005594.

Vishwakarma, S., Agrawal, A., 2013. A survey on activity recognition and behavior understanding in video surveillance. The Visual Computer 29, 983–1009. URL: http://dx.doi.org/10.1007/s00371-012-0752-6, doi:10.1007/s00371-012-0752-6.

Wang, H., Ullah, M.M., Kliser, A., Laptev, I., Schmid, C., 2009. Evaluation of Local Spatio-temporal Features for Action Recognition, in: British Machine Vision Conference.

Wang, J., Liu, Z., Wu, Y., Yuan, J., 2014. Learning actionlet ensemble for 3d human action recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on 36, 914–927. doi:10.1109/TPAMI.2013.198.

Wang, Y., Huang, K., Tan, T., 2007. Human activity recognition based on r transform, in: CVPR 2007, pp. 1 –8. doi:10.1109/CVPR.2007.383505.

Weinland, D., Ronfard, R., Boyer, E., 2011. A survey of vision-based methods for action representation, segmentation and recognition. Comput Vis Image Und 115, 224 – 241. URL: http://www.sciencedirect.com/science/article/pii/S1077314210002171, doi:http://dx.doi.org/10.1016/j.cviu.2010.10.002.

Ye, M., Zhang, Q., Wang, L., Zhu, J., Yang, R., Gall, J., 2013. A survey on human motion analysis from depth data, in: Grzegorzek, M., Theobalt, C., Koch, R., Kolb, A. (Eds.), Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications. volume 8200 of *LNCS*, pp. 149–187. URL: http://dx.doi.org/10.1007/978-3-642-44964-2_8, doi:10.1007/978-3-642-44964-2_8.