3D Cylindrical Trace Transform based feature extraction for effective human action classification

Georgios Goudelis*, Georgios Tsatiris*, Kostas Karpouzis*, Stefanos Kollias*†
*School of Electrical and Computer Engineering
National Technical University of Athens
9, Heroon Politechniou str., 15780, Athens, Greece
{ggoudelis, gtsatiris}@image.ntua.gr, kkarpou@cs.ntua.gr
†School of Computer Science
University of Lincoln
Brayford Pool, Lincoln, Lincolnshire, UK
skollias@lincoln.ac.uk

***Abstract*—Human action recognition is currently one of the hottest areas in pattern recognition and machine intelligence. Its applications vary from console and exertion gaming and human-computer interaction to automated surveillance and assistive environments. In this paper, we present a novel feature extraction method for action recognition, extending the capabilities of the Trace transform to the 3D domain. We define the notion of a 3D form of the Trace transform on discrete volumes extracted from spatio-temporal image sequences. On a second level, we propose the combination of the novel transform, named *3D Cylindrical Trace Transform*, with *Selective Spatio-Temporal Interest Points*, in a feature extraction scheme called *Volumetric Triple Features*, which manages to capture the valuable geometrical distribution of interest points in spatio-temporal sequences and to give prominence to their action-discriminant geometrical correlations. The technique provides noise robust, distortion invariant and temporally sensitive features for the classification of human actions. Experiments on different challenging action recognition datasets provided impressive results indicating the efficiency of the proposed transform and of the overall proposed scheme for the specific task.**

I. Introduction

Human action recognition has become one of the very important topics on the field of pattern recognition, especially due to its continually growing use in modern applications in everyday life. In the field of digital games, this relates to (mostly) console games which utilize cameras like Kinect, PlayStation Play or the Eyetoy, and exertion games [13] [25].

The problem of human action recognition is the automatic detection and classification of human activities from information acquired from cameras or other sensing modalities, such as accelerometers [26]. Although the idea is simple, the specific task is notably challenging as any relevant system has to overcome a large number of restrictive parameters: illumination variations, camera view angle, complicated backgrounds, and occlusions are only a fraction of the existing set of problems. In addition to the above mentioned, individuality [27] is another and very important factor that cannot be neglected, as every person performs the same set of movements (action) in a unique and different to every other person's way.

A. Related work

During the last decade, a large number of relevant algorithms have been proposed, as 3D information has started to play a leading role on emerging technologies. The first approaches on human action recognition based on 3D data appeared in the early 1980s; that research mostly utilized data received by visible-light cameras and monocular sensors, which demonstrate considerable loss of information. As a result, the release of low-cost depth sensors boosted further the growth of research on 3D data. A recent review in [2] summarizes the major techniques based in 3D imagery and separates them into four categories: *3D from stereo*, *3D from motion capture* and *3D from depth sensors*.

In terms of feature extraction strategies, a comprehensive analysis from Sun et al. [1] classifies them into into four categories: *motion based*, *appearance based*, *space-time volume based*, and *space-time interest points or local features based*. Regarding representation of action sequences, the same study detects again four classes, namely: *human silhouettes*, *space-time shapes*, *dense trajectories* and *local 3D patches*. In another recent survey [3], the authors split action recognition techniques into the *single-layered approaches*, which regard the action as a single entity, and the *hierarchical approaches*, which focus on the structural primitives that comprise an action (subactivities).

In a typical example of a pose-based technique that does not rely on temporal information, authors in [6] presented a technique that focuses on extracting *key poses* from action sequences. In essence, it makes a selection among the most distinctive poses from a specific set, in an attempt to avoid using complex action representations. More recent pose-based approaches range from crafting pyramidal features to represent human poses [7], to physics-inspired representations that leverage environment information [9]. In a more applied framework, Deboeverie et al. [5] employ a Random Forest based classification on skeletal poses, in order to detect dynamic gestures in physiotherapy scenarios.

Research on spatio-temporal feature extraction for actions

includes seminal works by Laptev [11] and Chakrabotry et al. [12], who delved further into the concept of exploiting spatio-temporal features in a Bag-Of-Video Words (BoVW) pipeline. This technique will be explained further in a following section. Recent motion based techniques include the one by Kumar and John [4], which analyzes optical flow. The one presented in [10] proposes skeleton fitting and motion trajectory extraction.

B. The proposed work: A preample

The work introduced in this paper is inspired by the study in [15], where we examined the potential of the original Trace Transform for human action recognition and proposed two feature extraction methods for the particular task. The proposed techniques manage to produce noise robust features that proved to be sufficient for successful recognition of human activity when tested on two popular datasets. However, both of the these techniques were based on modeling actions in a per-frame fashion, not taking into account any temporal interlinking between prominent features in the action sequence. Although they show resilience to occlusion, this may reduce their applicability on highly occluded environments, where spatial information can be distorted. In the current work, we propose a new form of the Trace Transform, extending its capabilities to the 3D space, and we present a novel feature extraction pipeline, suitable for activity recognition in videos.

The rest of the paper is organized as follows. The fundamental theory behind the Trace transform and the 3D Radon transform, which are the source of inspiration for the proposed 3D Cylindrical Trace transform, is presented in Section II. The presentation and the notation for the proposed transform are also found in the same section. The overview of the proposed feature extraction scheme is described in Section III. The experimental procedure and a discussion on the results are provided in Section IV, followed by a short conclusion in Section V.

II. The 3D Cylindrical Trace and its relevant transforms

The Trace transform can be considered as a generalization of the Radon [16] transform, in the sense that Radon is a subcase of Trace. While the Radon transform of an image is a 2D representation of the image in coordinates ϕ and p (with the value of the integral of the image computed along the corresponding line, placed at cell $[\phi, p]$), Trace calculates functional T along the tracing line. This functional may not necessarily be the integral. The final transform is created by tracing an image with straight lines and calculating certain functionals of the image values along these lines. This way, a variety of transforms, baring different properties can be extracted from the same image. The transform produced is a 2-dimensional function of the parameters (ϕ, p) of the tracing lines. Definition of these parameters is given in Figure 1. Examples of Radon and Trace transforms for different action snapshots are given in Figure 2. A detailed overview of the fundamental theory behind the Trace transform can be found in [17] and [15].

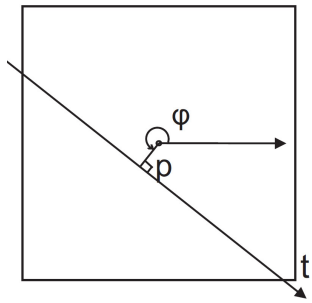

Fig. 1. Definition of the parameters of an image tracing line.

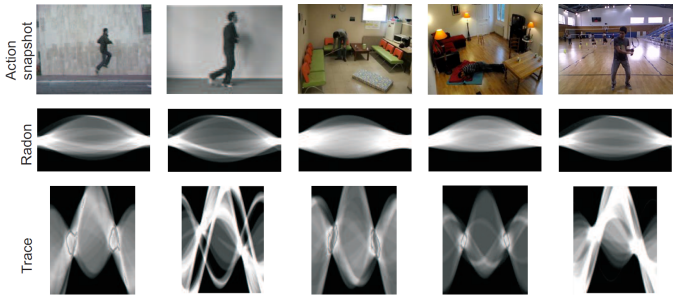

Fig. 2. Examples of Radon and Trace transforms created from the silhouettes of different action snapshots taken from various datasets.

A. The 3D Generalized Radon Transform

The definition of the transform proposed in this paper is inspired by the 3D Radon transform [18]. The latter is defined using 1D projections of a 3D object $f(x, y, z)$ where these projections are obtained by integrating $f(x, y, z)$ on a plane, whose orientation can be described by a unit vector $\vec{a}$. Geometrically, the continuous 3D Radon transform maps a function in $\mathbb{R}^3$ into the set of its plane integrals in $\mathbb{R}^3$. To make the comprehension of the proposed transform easier, we provide a brief overview of the 3D Radon transform, as it is formulated in [19]. The definitions and the basic properties of the continuous form of the transform proven in [18], are also valid for the discrete form presented bellow.

Let M be a 3D model and $f(x)$ the volumetric binary function of M, which is defined as:

$$f(x) = \begin{cases} 1, & \text{when } x \text{ lies within the 3D model's volume} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Let also, $\boldsymbol{\eta}$ be the unit vector in $\mathbb{R}^3$. The 3D discrete Radon transform of the 3D model $f(x)$ is given by:

$$T_f(\boldsymbol{\eta}, \rho) = \sum_{n=1}^{N} f(x_n)\delta(x_n^T \boldsymbol{\eta} - \rho) \tag{2}$$

where $\boldsymbol{\eta}$ is a unit vector in 3D space, denoting the transformation of point x_n in spherical coordinates, ρ is a real number, and $\delta(\cdot)$ is the Dirac delta function. The unit vector $\boldsymbol{\eta}$ can be written in spherical coordinates as:

$\boldsymbol{\eta} = [\cos\phi\sin\theta, \sin\phi\sin\theta, \cos\theta]$. Thus, equation (2) is rewritten as:

$$T_f(\rho,\theta,\phi) = \sum_{n=1}^{N} f(x_n,y_n,t_n)\cdot\delta(x_n\cos\phi\sin\theta + y_n\sin\phi\sin\theta + t_n\cos\theta - \rho) \quad (3)$$

This specific transform is easy to calculate. It is not invariant to scaling, translation and rotation, though.

B. The 3D Cylindrical Trace Transform

The proposed 3D Cylindrical Trace Transform (CTT) is an extension of the Trace Transform [17] to the 3D space. The Cylindrical Trace Transform CTT_f of the binary function $f(x)$, of a 3D model M, associates a functional T to each tracing line placed at cell (p,φ,θ), with distance p and angle φ characterizing uniquely a line and the plane that forms angle θ with the origin. A depiction of this can be seen in Figure 3.

In essence, Trace transforms are continuously calculated on planes rotating in the direction of polar axis A, cutting the 3D mesh M. The virtual cylinder created by the continuous rotation of planes in the polar direction is of radius ρ and of length l, with origin $O:(0,0,0)$. The radius ρ is defined as the distance of the outermost point $x(\rho_{max})$ of M from the longitudinal axis L of the cylinder. The length l is defined as the parallel to the longitudinal axis distance between the two most distant antipodal points of the 3D model. Every Trace transform $\breve{g}$ is calculated with respect to the center of the 3D model which coincides with the point K of the longitudinal axis of the virtual cylinder with cylindrical coordinates $(0,0,l/2)$. Thus, after 180 degrees of rotation, the resulting $\breve{g}$ is equal to the transform calculated on the plane with $\theta = 0$.

Considering each cutting plane as a 2D function $\xi(x,y)$ formed by the projection of the mesh on that plane, its Trace transform $\breve{g}(p,\varphi)$ can be given by evaluating a functional T along all lines (p,φ) tracing ξ:

$$\breve{g}(p,\varphi) = T(\xi(x,y)\delta(p - x\cos\varphi - y\sin\varphi)) \quad (4)$$

The final representation of the 3D model is the proposed 3D CTT. This transform is also a 2D function of parameters (p,φ) and is given by the sum of the individual calculations of Trace transforms on planes rotated by angle θ relative to the origin. in the direction of the angular coordinates defined by θ:

$$CTT_f(p,\varphi) = \sum_{n=1}^{N} \breve{g}_n(p,\varphi). \quad (5)$$

where $\breve{g}_n$ is the n^{th} Trace transform, i.e. the transform calculated on the 2D planar projection of M on the plane that forms angle θ_n with the origin. Also: $N \geq 2$, $0 < \theta_n \leq \theta_{max}$ and $\theta_{max} = 180^o$. An illustration of the 3D Cylindrical Trace transform is given in Figure 3.

To make the transform scale invariant, the maximum distance d_{max} between the center of the mass and the most distant voxel of the 3D volume is calculated. In following, $f(x)$ is scaled so that $d_{max} = 1$. CTT is always calculated with respect to the center of the mass, which coincides with the point K. Thus, CTT is translation invariant. The same properties hold for the continuous and the discrete form as well.

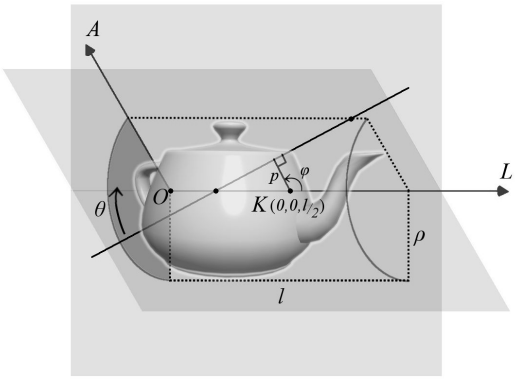

Fig. 3. 3D Cylindrical Trace transform illustration.

III. Overview of the Proposed System

In [15], two different ways based on the Trace transform have been proposed, for the extraction of features from human action videos. Both methods (*History Trace Templates* (HTTs) and *History Triple Features* (HTFs)) were able to create representations of low dimensionality from an action sequence. Both proved to be robust in noise and illumination variations. However, lack of time sensitivity and occlusion issues could not be effectively handled. This inspired the newly formulated transform and the methodology presented in this study.

The proposed scheme has been designed for the scenario of human action recognition in video sequences. It combines the proposed 3D CTT with a state of the art algorithm for spatio-temporal interest point acquisition, the so-called Selective Spatio-Temporal Interest Points (SSTIPs) [12]. A three-dimensional spatio-temporal volume is crafted based on the SSTIPs mesh and various 3D Cylindrical Trace Transforms, using different functionals, are calculated by it. Finally, the results are used in a triple-feature extraction scheme that produces feature vectors of very small length. The concept of this particular methodology is to take advantage of the most valuable attributes each one of these techniques has to provide and combine them in a final and straightforward pipeline.

Spatio-temporal feature acquisition methods, such as STIPs and Bag of Visual Words (BOVW), are very hot lately in the field of action recognition. However, this kind of representations ignore potential valuable information that refers to the global spatio-temporal distribution of interest points [19]. By introducing the Cylindrical Trace Transform, the methodology presented in this paper manages to capture detailed information about the geometrical distribution of interest points, while at the same time it provides the versatility of creating a large number of potential features for a variety of

capturing conditions, environments and applications. The use and the combination of different and suitable functionals for the calculation of different features can provide very robust representations of an action video sequence, in the form of feature vectors. More details on the individual techniques and the proposed scheme are provided in the following subsections.

A. Selective Spatio-Temporal Interest Points

As mentioned above, the proposed scheme incorporates the use of a novel approach to the STIPs acquisition problem, presented in [12], the so-called Selective Spatio-Temporal Interest Points technique. In this study, the authors proposed a Spatio-Temporal Interest Points (STIPs) extraction methodology which focuses on global motion instead of local spatio-temporal information, thus preventing the erroneous detection of interest points due to cluttered backgrounds and camera motion. Furthermore, they show that their method performs well in producing stable, repeatable STIPs, robust to the local properties of the detector throughout the motion sequence.

One could summarize the selective STIPs pipeline as a procedure that: 1. detects spatial interest points, 2. suppresses unwanted background points and 3. imposes local and temporal constraints on the result. The first step is essentially conducted using a Harris corner detector. The underlying idea behind the second step is the observation that corner points detected in the background follow some particular geometric pattern, while those on humans do not bear this property. Finally the spatial and temporal constraints are imposed, based on the notion that for an interest point to be considered an accurate and repeatable STIP, it should show a positional change through the motion sequence. An example of extracted SSTIPs from a sample of the THETIS dataset is given in Figure 4.

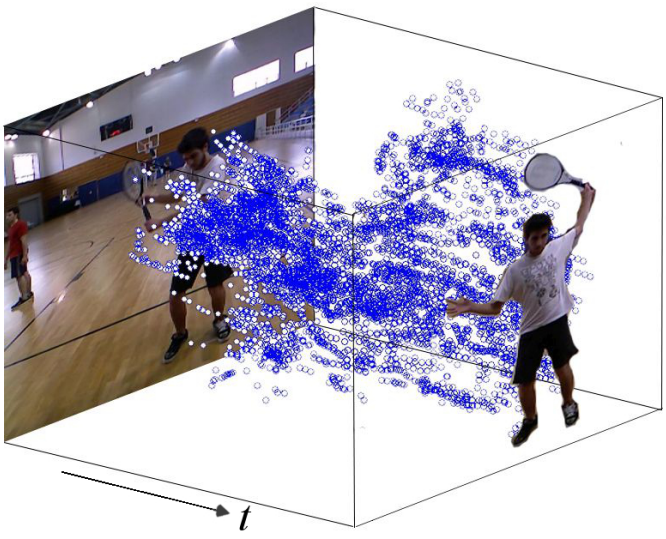

Fig. 4. Selective STIPs extracted from a backhand shot video sequence from the THETIS dataset. t denotes the direction of time.

B. 3D CTT on Selective Spatio Temporal Interest Points

Kadyrov and Petrou [17], through the formulation of the Trace transform, have shown that not only integrals (Radon transform) but also other selectively calculated functionals, such as the median or the mean, along straight lines tracing a two dimensional function, have the ability to reconstruct this function fully. As already explained above, by using a variety of functionals, we have the ability to calculate different Trace transforms of that function. Every transform is itself a 2D-function of the parameters (p, φ) of each tracing line.

Accordingly, 3D CTT is produced by tracing consecutive planes, rotated by angle θ, belonging to the same minimal window (cylinder) of a 3D model and having the same origin *K*. The 3D CTT is the sum of individual planar transformations which results in a new 2D function of parameters (p, φ). In the same way as in Trace transform, one can produce a variety of 3D CTTs providing different properties by calculating different functionals.

The ultimate goal is to capture the dynamic information and structure of an action to the best possible extent. At this point, Trace transform has already shown its suitability for this task [15]. The 3D CTT proposed in this paper is an evolution of the Trace transform that provides the benefits of the regular form and extends its capabilities into the 3D space, which, in the context of action recognition, is the spatio-temporal domain. Consequently, this formulation is more suitable for the representation of actions, which are inherently spatio-temporal volumes.

Let M be the 3D model formed by the SSTIPs mesh created from a human action video. Let ρ and l be the radius and the length respectively, of the smallest cylinder bounding the mesh. Then, let z be a random plane with size 2ρ x l where ρ is the radius and l is the length of the minimum cylinder bounding the 3D mesh. The Trace transform $\check{g}_z(p, \varphi)$, is a function defined on the space of straight lines that trace z. p and φ are the parameters that define the position of the line on the relative 2D coordinate system of z, which forms angle θ with the origin. So, as documented in equation 4, z can be expressed as a 2D projection $\xi(x, y)$ of the nearby (within a certain tolerance) and coincident points on z and $\check{g}_z(p, \varphi)$ is the transform calculated by evaluating a functional T over the line $p = x\cos\varphi + y\sin\varphi$ which traces z. The reference point is defined by the center of the SSTIPs volume. The sum of all the planar transforms gives the final 3D CTT of mesh M, as explained in equation 5.

By applying different functionals to the SSTIPs mesh M, a set of $CTT_{f_i}(p, \varphi)$ transforms is produced, where $i = 1...I$ and I is the number of transforms one chooses to calculate. The final set of 3D CTTs will become subject to the triple-feature extraction scheme provided in the next subsection.

This scheme provides the benefit that silhouette extraction or temporal alignment of the sequences is not required. Spatio-temporal interest points are extracted based on the spatial and temporal evaluation of movement and intensity changes. By calculating the CTT on such a mesh, the geometric distribution

of the points is embedded in the final transform, which will also capture the distinct and persistent spatio-temporal information that characterizes each action. Thus, temporal alignment is not required. An illustration of how 3D CTT is calculated across SSTIPs is given in Figure 5.

Another improvement compared with previous techniques is the ability of this pipeline to encode variations in the length of action sequences. In other words, if the speed at which an action is performed plays a significant role in the classification of that action, this pipeline has the ability to incorporate it in the extracted feature vector. We call this property *time-sensitivity*. Previous techniques based on the extraction of features in a per-frame fashion lack this property.

Finally, it should be noted that, as 3D CTT has been designed having in mind the extraction of features from spatio-temporal sequences, it also differentiates itself from 3D Radon in another way. Since the 3D models that CTT is applied on are in fact spatio-temporal volumes, the transform processing is always performed with the direction of time being perpendicular to the rotational axis. This way, it is assumed that the pose of a spatio-temporal mesh is constant and aligned with the time axis. As this is always the case of application of the proposed transform and the time dimension will never be rotated into another direction within the 3D space, CTT can take the form of a cylinder and not necessarily this of a sphere. This way, the proposed transform manages to be more time efficient.

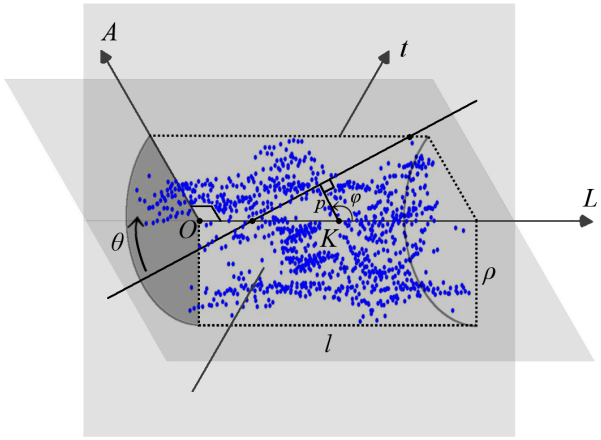

Fig. 5. 3D CTT calculation on Selective STIPs extracted from an action video sequence. t denotes the direction of time.

C. The Volumetric Triple Feature (VTF) extraction scheme

The fundamental work of Kadyrov and Petrou [17] proposes the formulation of a specific type of features, called *triple features*, as a very simple but rich representation of the result of the Trace transform of an image. Initially formulated as a solution for the classification problem of images of fish closely resembling each other, a triple feature is constructed in the following manner:

1) The Trace transform of a 2D function is produced, using a *Trace functional* T.
2) By calculating a *diametric functional* P along the columns of the 2D function's Trace transform, a circus function is obtained.
3) The final triple feature is ultimately produced by applying a *circus functional* Φ along the resulting vector of numbers from step 2.

The theory behind triple features is further elaborated in [17].

In [15] it is shown that *ratios of pairs* of different triple features, constructed by using different functionals T, P and Φ on the frames of an action sequence (video), can create feature vectors that effectively represent the action sequence. This feature vector extraction technique, called *History Triple Features* (HTFs), offered robust features, sufficient for the successful recognition of human activity. This technique, however, came with shortcomings.

The methodology presented in this paper aims to tackle these shortcomings, as seen in the previous section. Now, using the newly formulated 3D Cylindrical Trace Transform on spatio-temporal volumes produced by SSTIPs of action sequences, we propose the formulation of triple features on volumes, called *Volumetric Triple Features* (VTFs). The extraction scheme of VTFs follows the aforementioned pattern for triple feature extraction. For every $CTT_{f_i}(p, \varphi)$, calculated using functional T_i, we apply a diametric functional P_i along the columns of the transform. Then, a circus functional Φ_i is evaluated along the resulting string. This way, a set of Π triple features is computed. The procedure is illustrated in Figure 6.

All Π features are then divided by each other, to produce a new set of independent features. So, the given action sequence is finally depicted by a vector $\mathbf{v}$, the so-called VTF vector, which is essentially the set of all calculated triple feature ratios, based on the set of the different CTT_{f_i} applied on the SSTIPs of the action sequence:

$$\mathbf{v} = (\Pi_{rat_1}, \Pi_{rat_2}, ..., \Pi_{rat_{h-1}}, \Pi_{rat_h}) \tag{6}$$

where each Π_{rat} is a ratio of two triple features and h is the total number of calculated ratios.

This method enables the effortless construction of a large number of features, while at the same time keeping the produced representation relatively concise. In the scenario that 10 functionals are utilized for each stage (e.g. 10 T functionals, 10 P functionals and 10 Φ functionals), one may construct up to 10x10x10 = 1000 triple features. The number of divisions performed may affect the length of the final vector.

Due to the fact that not all features in the vector share the same discriminatory power, the use of a dimensionality reduction technique is deemed suitable for the task of selecting the most discriminant features. It will also make the classification problem more tractable. In this pipeline, Principal Component Analysis (PCA) is applied on the VTF vectors, in order to construct an appropriate subset of the features that is suitable for classification. In our experiments, only a small fraction of the initial VTF vector survives this task (typically between 25 and 40 features).

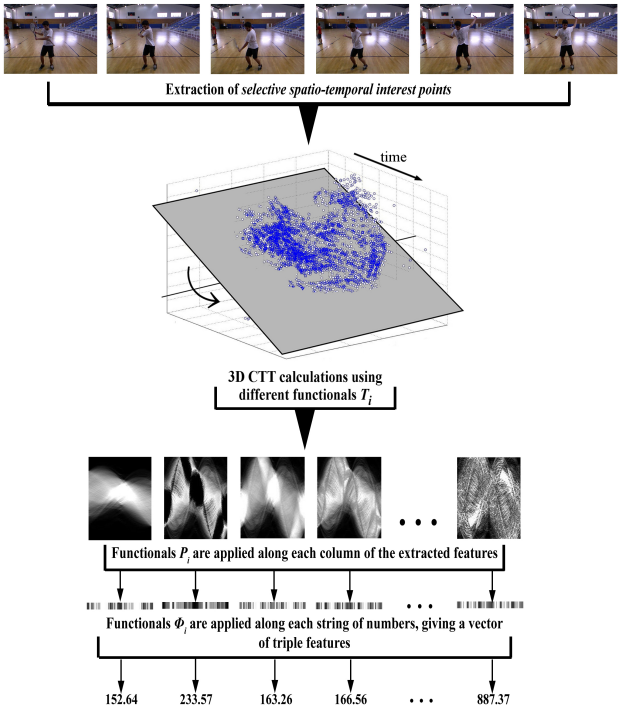

Fig. 6. Triple feature extraction from a spatio-temporal volume.

IV. Experimental evaluation

In this section, we will document the experimental procedure we followed in order to indicate the efficacy of the proposed technique on the task of human action recognition. Experimental results are reported on a set of different known and challenging datasets. The algorithm performance under different data sources is also demonstrated.

At this point and before we describe the experimental protocols used, it is important to mention that, according to [20], there is a variety of different experimental scenarios used for the same datasets among researchers working on action recognition from videos. It is also reported that methods evaluated on known datasets such as KTH [14] and Weizmann [21] may present result variations up to 10.67% when different validation approaches are applied.

In the following experiments, the leave-one-person-out cross validation protocol was used for the evaluation of performance. This protocol represents real-world conditions in the best way. In a hypothetical real world scenario, the physical activity of an unknown person is captured by a vision-based recognition system and thereafter processed and compared against a pre-processed dataset that has been used to train that system. The classification of the recorded activity is determined based on its relevance when compared to any sample of the data that comprise the training set, according to the system's specific rules. Accordingly, the leave-one-person-out protocol utilizes one person's action samples for testing, while the rest of the samples form the training set. This procedure is repeated N times, where N is the number of subjects (persons) within the dataset. Performance is measured as the average accuracy of $I = \sum_{n=1}^{N} H_n$ iterations, where H_n is the number of samples for the n^{th} subject within the dataset.

A. Experimental setup

For the experiments, three different datasets have been used. The KTH [14], the Weizmann [21] and the THETIS [22] action databases. Figures 7, 8 illustrate various samples for the different types of actions contained in the first two datasets. In the KTH dataset, six types of different actions are contained, namely walking, jogging, running, boxing, hand waving and hand clapping. These actions are performed a number of times, by 25 different persons in four different scenarios, under various illumination conditions. All video sequences were captured over homogeneous backgrounds at 25 frames per second, using a static camera.

The Weizmann video database is comprised by a set of 90 low-resolution video sequences showing nine different subjects. Each subject performs 10 natural actions such as running, walking, skipping, jumping-jack (jack), jumping forward on two legs (jump), jumping in place on two legs (pjump), galloping sideways (or side), waving with two hands (wave2), waving using one hand (wave1) and bending.

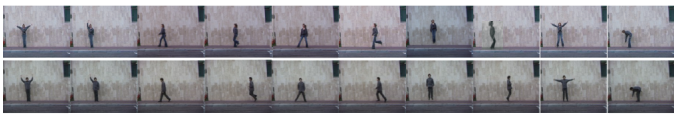

Fig. 7. Weizmann dataset: samples of the wave1, wave2, walk, pjump, side, run, skip, jack, jump and bend actions.

Fig. 8. KTH dataset: samples of the walking, jogging, running, boxing, hand waving and hand clapping actions respectively.

The THETIS set is comprised of 12 basic tennis shots performed by 31 amateurs and 24 experienced players. All videos have a resolution of 640x480 and have been captured using a Kinect sensor placed in front of the subjects. Each shot has been performed at least 3 times, resulting in 8734 (single period cropped) videos, converted to AVI format. The shots performed are the following: backhand with two hands, backhand, backhand slice, backhand volley, forehand flat, forehand open stands, forehand slice, forehand volley, service flat, service kick, service slice and smash. The modalities of the dataset used in these experiments are RGB, Depth and 3D Skeleton videos. Samples from the THETIS dataset are illustrated in Figure 9.

The action videos of the aforementioned datasets were scaled up or down, whenever deemed necessary. Then they

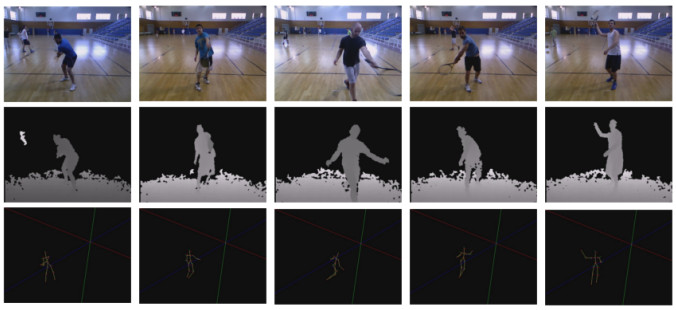

Fig. 9. Action samples from the THETIS database for the backhand, flat service, forehand flat, slice service and smash moves. Top row: RGB samples, middle row: depth samples, bottom row: 3D skeleton samples.

were used as input for the proposed pipeline, in order to create feature vectors. These were in following fed to a series of gaussian radial basis function based Support Vector Machines (SVMs). To experiment on the variations in the results produced by using different values for the plane rotation θ step, the pipeline was tested with step value 9^o and 6^o. One can intuitively determine that the smaller this step is, the closer it gets to the continuous form of the proposed transform. This may offer more robust features, although at the cost of time efficiency. Finally, PCA is performed on the produced vectors, in order keep the subset of the most discriminant features and reduce dimensionality.

Human action recognition is a multi-class problem. In order to follow the leave-one-out protocol, we formulate the problem as a general case of binary classification. In essence, we trained 6 class-specific SVMs, one for each class of the KTH dataset, using an one-against-all protocol. Similarly, we trained another 10 and 12 class-specific SVMs for the classes of the Weizmann and the Thetis dataset respectively. The final decision was made by assigning each testing sample to a class $\mathcal{C}_a$, according to the distance d of the testing vector from the support vectors of the specific class. However, since the purpose of the evaluation process was to test the generalization of the proposed pipeline in a more broad fashion, we tested every sample against all the class-specific SVMs and recorded both the successful classifications and the false positives. The results of this experimental procedure can be found in IV-B.

TABLE I
CONFUSION MATRIX OF THE PERFORMANCE OF THE PROPOSED METHOD ON THE KTH DATASET

	boxing	handclapping	handwaving	jogging	running	walking
boxing	**1**	0	0	0	0	0
handclapping	0.0009	**0.9991**	0	0	0	0
handwaving	0	0	**1**	0	0	0
jogging	0	0	0	**1**	0	0
runing	0	0	0,01	0	**1**	0
walking	0	0	0	0	0	**1**

TABLE II
CONFUSION MATRIX OF THE PERFORMANCE OF THE PROPOSED METHOD ON THE WEIZMANN DATASET

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	**1**	0	0.1111	0	0	0	0	0	0	0
jack	0	**1**	0.0455	0	0	0	0	0	0	0
jump	0.0588	0.0294	**0.9706**	0.0294	0.1176	0.0294	0.0588	0.0294	0.0588	0.0294
pjump	0.0303	0.0303	0.0303	**0.9697**	0.0303	0.0303	0.0303	0.0303	0.0303	0.0303
run	0.1053	0.1579	0.2105	0.1053	**0.9474**	0.0526	0.1579	0.0526	0.1053	0.0526
side	0	0	0	0	0	**1**	0.0357	0	0	0
skip	0.0526	0.0526	0.1053	0.0526	0.0789	0.0526	**0.9474**	0.0526	0.0526	0.0526
walk	0.0455	0.0455	0.1364	0	0.0455	0	0.0455	**1**	0	0
wave1	0.0476	0.0476	0.0476	0	0.0952	0.0476	0.0952	0.0476	**0.9524**	0.0476
wave2	0	0	0	0	0	0	0	0	0	**1**

B. Results

Comparative results on the action recognition task can be found in Table III. In addition, Tables I, II show the confusion matrices generated by the presented method on the vastly investigated KTH and Weizmann datasets. Rows depict the accuracy achieved (correct answers/all answers) by all trained SVMs on a certain action class. Columns, on the other hand, show the performance of individual, class-specific SVMs on all action classes.

As seen in Table III, the feature extraction pipeline based on 3D CTT and the Selective STIPs achieves impressive accuracy on all examined datasets and is on a par with or even outperforms other published methods in the field. Particularly on the KTH database, the presented technique achieves a notable 99.98% accuracy, which is validated by the corresponding confusion matrix (Table I). In a noticeable comparison, the pipeline proposed by Yuan et al. in [19], which is based on features extracted using a formulation of the 3D Radon transform and a combinatorial STIP + BoVW method, achieves 95.49% accuracy on the same dataset. Results on the Weizmann, where accuracy of 96.34% was achieved, suggest the existence of a slight confusion of the class-specific classifiers, especially of the ones dedicated to the ”jump” action.

There is considerable difference in performance between the three data types in the THETIS set. In the comparative results of Table III, one can see that the proposed technique outperforms all other methods tested on this dataset, namely the HOG-HOF based method on STIPs, by Laptev et al. [11], the dense trajectories approach of Wang et al. [24] (on the Depth and Skelet3D subsets) and the dynamic phases based approach of Vainstein et al. [23] (on the RGB subset). Although the results on the Skelet3D subset cannot be considered unsatisfactory, one can notice a drop in accuracy. To a certain point, this drop was expected, considering the nature of the data. Skeletons are an oversimplified model of the moving human body with minimum surface, hindering accurate STIP acquisition. However, compared to the results of another prominent STIP based method reported in [11], this method seems to be an outright improvement. On the other hand, in the Depth subset, the results are close to perfect. Noteworthy, though, is that 100% accuracy can be reported on the RGB subset. This indicates the suitability of the presented method on RGB and gray-scale imaging.

TABLE III
CLASSIFICATION PERCENTAGES (%) ACHIEVED BY DIFFERENT PUBLISHED METHODS ON THE KTH, WEIZMANN AND THETIS DATABASES.

Method	Dataset				
	KTH	Weizmann	THETIS-Skelet3D	THETIS-Depth	THETIS-RGB
3D CTT - SSTIP based VTFs	**99.98**	**96.34**	**86.06**	**98.03**	**100**
Selective STIPs + BoVW [12]	96.35	99.5	-	-	-
Dense Trajectories: MBH [24] (rep. in [22])	92.32	-	46.84	51.59	-
Dense Trajectories: Combination [24] (rep. in [22])	90.65	-	50.78	54.32	-
Dense Trajectories: Trajectory [24] (rep. in [22])	86.98	-	53.08	57.5	-
History Triple Features [15]	93.14	95.42	-	-	-
Yuan et al. [19]	95.49	-	-	-	-
Vainstein et al. [23]	-	-	-	-	86.44
Kumar and John [4]	94.62	95.69	-	-	-
Liu et al. [8]	96.7	-	-	-	-
Laptev et al. [11] (reported in [22])	92.99	-	54.4	60.23	-

V. CONCLUSION

In this paper, we propose an extension of the Trace transform to the 3D space and a new combinatorial scheme for feature extraction, applied on a mesh of spatio-temporal interest points extracted from an action video. The method blends the newly formulated 3D Cylindrical Trace transform with Selective Spatio-temporal Interest Points in a triple features extraction methodology, which manages to provide a higher level of discrimination in the task of characterizing different types of actions. The proposed pipeline has the ability to create distortion invariant and temporally sensitive action representations. Experimental results on different challenging datasets indicated that the produced features show robustness in noise, illumination variation, translation and scaling issues. At the same time, the method provides the ability of adaptation to various types of applications and their particular conditions.

ACKNOWLEDGMENT

This work was supported by the 4-year EC funded project iRead (Jan 2017-Dec 2020). This project has received funding from the European Union's Horizon 2020 innovation programme under grant agreement No 731724.

REFERENCES

[1] C. Sun, I. N. Junejo, M. Tappen, H. Foroosh, Exploring sparseness and self-similarity for action recognition. IEEE Transactions on Image Processing, 24(8), 2015, pp. 2488-2501.

[2] J. Aggarwal, L. Xia, Human activity recognition from 3d data: A review, Pattern Recognition Letters 48 (2014): 70-80.

[3] Cheng, G., Wan, Y., Saudagar, A. N., Namuduri, K., Buckles, B. P. (2015). Advances in human action recognition: A survey. arXiv preprint arXiv:1501.05964.

[4] S. S. Kumar, M. John, Human activity recognition using optical flow based feature set. In Proc. of IEEE International Carnahan Conference on Security Technology (ICCST), 2016, pp. 1–5.

[5] Deboeverie, F., Roegiers, S., Allebosch, G., Veelaert, P., Philips, W. (2016, September). Human gesture classification by brute-force machine learning for exergaming in physiotherapy. In Computational Intelligence and Games (CIG), 2016 IEEE Conference on (pp. 1-7). IEEE.

[6] S. Baysal, M. C. Kurt, P. Duygulu, Recognizing human actions using key poses, Pattern Recognition, International Conference on (2010) 1727–1730.

[7] L. Liu, L. Shao, X. Zhen, X. Li, Learning discriminative key poses for action recognition, IEEE transactions on cybernetics 43.6 (2013): 1860-1870.

[8] A.A. Liu, Y.T. Su, P.P. Jia, Z. Gao, T. Hao, Z.X. Yang, Multipe/single-view human action recognition via part-induced multitask structural learning, IEEE transactions on cybernetics 45.6 (2015): 1194-1208.

[9] A. Mansur, Y. Makihara, Y. Yagi, Inverse dynamics for action recognition, IEEE transactions on cybernetics 43.4 (2013): 1226-1236.

[10] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, A. Del Bimbo, 3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold, IEEE Trans. on Cybernetics 45.7 (2015): 1340-1352.

[11] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Conference on Computer Vision & Pattern Recognition (CVPR), IEEE, 2008.

[12] B. Chakraborty, M. Holte, T. Moeslund, J. Gonzalez, Selective spatio-temporal interest points, Computer Vision and Image Understanding 116 (3) (2012) 396–410.

[13] J. Nijhar, N. Bianchi-Berthouze, G. Boguslawski, Does Movement Recognition Precision Affect the Player Experience in Exertion Games?, in: Proceedings of INTETAIN 2011, pp 73–82.

[14] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: A local SVM approach, in: Proc. ICPR, 2004, pp. 32–36.

[15] G. Goudelis, K. Karpouzis, S. Kollias, Exploring trace transform for robust human action recognition, Pattern Recognition 46 (12) (2013) 3238–3248.

[16] S. R. Deans, The Radon Transform and Some of Its Applications, Krieger Publishing Company, 1983.

[17] A. Kadyrov, M. Petrou, The Trace transform and its applications, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 811–828.

[18] A. Averbuch, Y. Shkolnisky, 3d fourier based discrete radon transform, Applied and Computational Harmonic Analysis 15 (1) (2003) 33 – 69.

[19] C. Yuan, X. Li, W. Hu, H. Ling, S. Maybank, 3d R transform on spatio-temporal interest points for action recognition, in: IEEE Conference on Computer Vision & Pattern Recognition (CVPR), 2013, pp. 724–730.

[20] Z. Gao, M.-Y. Chen, A. G. Hauptmann, A. Cai, Comparing evaluation protocols on the KTH dataset, in: Proc. of the First Int. Conf. on Human Behavior Understanding, HBU'10, Springer-Verlag, 2010, pp. 88–100.

[21] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: The 10th IEEE International Conference on Computer Vision (ICCV'05), 2005, pp. 1395–1402.

[22] S. Gourgari, G. Goudelis, K. Karpouzis, S. Kollias, Thetis: Three dimensional tennis shots a human action dataset, in Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 676-681, 2013.

[23] J. Vainstein, J. Manera, P. Negri, C. Delrieux, A. Maguitman, Modeling video activity with dynamic phrases and its application to action recognition in tennis videos, in: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Springer, 2014, pp. 909–916.

[24] H. Wang, A. Klaser, C. Schmid, C.-L. Liu, Action recognition by dense trajectories, in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 3169–3176.

[25] N. Bianchi-Berthouze, K. Isbister, Emotion and Body-Based Games: Overview and Opportunities, in K. Karpouzis, G. N. Yannakakis (eds.), Emotion in Games: Theory and Praxis, Springer, pp. 235–255.

[26] G. Caridakis, K. Karpouzis, M. Wallace, L. Kessous, N. Amir, Multimodal users affective state analysis in naturalistic interaction, Journal on Multimodal User Interfaces, Vol. 3(1), pp. 49–66.

[27] G. Caridakis, J. Wagner, A. Raouzaiou, Z. Curto, E. Andre, K. Karpouzis, A multimodal corpus for gesture expressivity analysis. Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, p. 80, 2010.