RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

REPORTE TÉCNICO
Reconocimiento de Patrones

Full-Body Human Action Recognition: State of the Art

Lic. Mabel Iglesias Ham,
Dr. C. Edel García Reyes

RT_021 **marzo 2010**

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
Reconocimiento de Patrones

Full-Body Human Action Recognition: State of the Art

Lic. Mabel Iglesias Ham,
Dr. C. Edel García Reyes

RT_021 **marzo 2010**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Index

Full-Body Human Action Recognition: State of the Art

Lic. Mabel Iglesias Ham, Dr. C. Edel B. García Reyes

Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV)
7ma ♯21812 e/ 218 y 222, Rpto. Siboney, Playa, C.P. 12200, Ciudad de La Habana, Cuba.
miglesias@cenatav.co.cu

RT_021 CENATAV
Fecha del camera ready: 30 de octubre de 2009

Abstract: Human action recognition in video sequences is particularly useful in application areas as video-surveillance, video indexing and browsing, automatic video-annotation, among others. A significant research effort has been done in this area mainly over the last two decades. However, first studies were mainly based on the use of markers attached to the human body or to a prototype, to capture accurately the motion patterns. Very far in human history when advances in computer technology were not available, there was some interest in understanding human motion for Medicine, Biomechanics and Art. In this technical report we are aiming to focus in the main publications for human action recognition, which is considered a topic included within the human motion analysis subject. We base our efforts in recently published reviews enriched by new contributions of the last years. We propose a new taxonomy, grouping the different approaches and stressing their limitations as a starting point to future contributions in the field.
Keywords: Human Action Recognition, Video-Surveillance, Space Time Volume-Set

Resumen: El reconocimiento de acciones humanas en secuencias de video es de vital importancia en aplicaciones de video-vigilancia, video-indexación y exploración, anotación de video de forma automática, entre otros. Un significativo esfuerzo de investigación se ha realizado en esta área principalmente en las últimas dos décadas. No obstante, los primeros estudios estuvieron apoyados en el uso de marcadores sobre el cuerpo humano o de un prototipo, para obtener de forma exacta los patrones del movimiento. Mucho antes en la historia cuando no existía el avance tecnológico de estos tiempos, también se mostró interés en el entendimiento del movimiento humano en áreas como el Arte, la Biomecánica y la Medicina. Este reporte técnico pretende enfocarse en las principales publicaciones sobre el reconocimiento de acciones humanas, tópico que se cosidera está incluido en el área del análisis del movimiento humano. Basamos el trabajo en estados del arte de reciente publicación enriquecidos con nuevas contribuciones de los últimos años. Proponemos una nueva taxonomía, agrupando los diferentes enfoques con las limitaciones que presentan como punto de partida para futuras contribuciones.
Palabras clave: reconocimiento de acciones humanas, video-vigilancia, volumen espacio temporal

1 Introduction

Recognizing actions of human beings from video is a highly active research area in Computer Vision due to its applications in video surveillance, human-computer interface, sports, video indexing and browsing, automatic video-annotation, among others.

Before advanced computer technology were available to perform motion analysis based on a captured image data or a model based calculation(simulation), some work was

done in areas like Medicine, Biomechanics and Art that contributed to human motion understanding[28]. Examples include studies about the characteristics of normal and pathological motion in Biomechanics. They focus mainly in locomotion giving less attention to muscle models. However, to correct the motion of a disable child through surgery, the attention is given more in the opposite direction.

In the later half of the 19th century began experimental studies in gait(Biomechanics), measuring features as the center of mass of human body used for the development of prosthesis. Later on, with the use of computers, these studies were mainly supported with marker based pose tracking (See Fig. 1), starting with the work of Johanson, 1973 [25]. Here, the cameras were fast but they only used binary information received from the highlighted markers attached to the human body, afterward used to generate a stick figure for simulations. Also, with the advent of computers, areas as Computer Vision and Computer Graphics appear with particularly useful applications in computer games, movies, automatic video surveillance, automatic video-annotation, video indexing and browsing, etc. Both areas have as main interest topic the human model for tasks as tracking, recognition of human actions and human motion modeling.

Fig. 1. Markers based human motion capture. Image taken from [28]

In general, human motion understanding deals with the analysis of global motion patterns and not with some particular parts such as the face or hands, to recognize expressions or gestures usually used in human-computer interaction applications. We would also like to differentiate some works in *vision-based human motion analysis* [37], which deals with the estimation of human body parts over time and not with motion interpretations. This is a separate field from pose recognition, which deals with classifying a pose from a limited number of classes; and also it is different from gesture recognition, which deals with interpretations of movement over time.

Pose estimation can be approached in 2D or 3D. In 3D, pose estimation is useful to be able to estimate the overall body posture in 3D in order to understand subject behavior. In 2D, it means estimation of the projection of the 3D object pose in the 2D image.

Gait recognition is included in the category of human motion understanding and aims to identify a person by its way of walking which is an important feature for several biometric applications. Also, gait has been analyzed from other perspectives, e.g. sex discrimination, or detecting abnormality in walking behavior (clinical applications), analysis of actions such as running, hopping or limping. For an overview on gait analysis and pose estimation studies you can refer to [42,43].

Here we are actually interested in the interpretations of the motion for action recognition. Besides, we are interested in videos captured from a single static camera, capturing motion with a non intrusive technique, so we reject the expensive marker based methods.

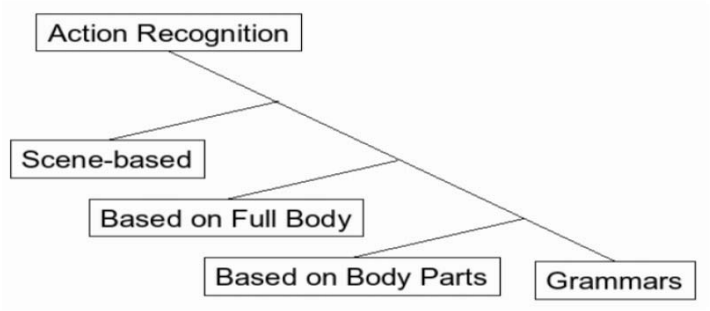

Fig. 2. Different types of action recognition approaches presented by Kruger, 2007 [29]

In [29] the authors deal with action at different levels of complexity and provide a general taxonomy of action recognition approaches (See Fig. 2). The first kind of action recognition is scene-based and it is oriented for surveillance applications to distinguish "regular" from "irregular" activities and it should be independent from the object causing the activity, e.g, cars or persons. The other approaches focus explicitly on human activities, they distinguished between the full-body based approaches and the body-part based approaches that model the human as a set of body parts. In the human activities cases it is considered to have as a basic information the general view of the whole body in time. The approaches based on body parts consider more detailed information of the human motion and they are used to detect a more complex set of actions. With the whole body approach simple actions like running, walking and jumping as their speed and direction can be detected. Finally, an approach based on action primitives and grammar is used to interpret actions as a composition on the alphabet of these action primitives. The primitives are detected and an iterative process is used to find the sequences of primitives for a novel action.

In this report we will use the action hierarchy presented in [29,35] due its intuitiveness and simplicity: *actions primitives, actions* and *activities*. *Activities* are decomposed into *actions*, and *actions* are integrated by *action primitives*. We aim to deal with approaches that focus on human actions based on observations of the the whole body to identify simple actions such as running, jumping, bending, walking, boxing, hand-clapping, etc (See Fig. 3). Not the case when talking about gait recognition for biometry; or recognizing more complex actions like robbing a car or drinking coffee, that can be seen as an action interacting with the environment (See Fig. 4).

Fig. 3. Simple actions as walking, crouching and falling. Image taken from [21]

Fig. 4. Actions defined by the interaction with the environment. Image taken from [31]

The rest of the report is organized as follows. Section 2 shows a commented review on the general approaches for human action recognition, followed by the proposal of a new taxonomy(Section 3). Next, a summary with the detected open problems is provided in Section 4. In Section 5 a collection of the most used evaluation databases are enumerated. The most active conferences in the topic are grouped in Section 6. Finally, in section 7 we show conclusions of the state of the art.

2 Different Approaches for Human Action Recognition

In the study of human dynamics most of the motion can be characterized as non-rigid and piecewise rigid (articulated motion). Articulated motion is called to the motion of individual rigid parts of an object that move independent of one and another. The rigid parts apply to the rigid motion constraints, but the overall motion is not rigid.

When analyzing human motion two main approaches are considered: based on an a priori shape model or not. Shape models of increasing complexity used are stick figures, 2D and 3D contours. The stick figure approach is supported in the observation that the human motion is basically the motion of the skeleton attached with muscles, and 2D silhouettes by the frequent input of a 2D projection of the 3D human in images. The volumetric approaches approximate better the details of the human body by 2D ribbons, generalized cones, elliptical cylinder and spheres. However, they require a bigger number of parameters for computation. In [42] a detailed discussion of the two approaches is presented.

We agree with the statement in [29], saying that the use of an explicit a priori model (either 2D or 3D), is often not feasible in case of noisy and imperfect imaging conditions and that a direct pattern recognition based on the 2D data is potentially more robust. In addition, 3D human estimations are usually obtained by an estimation from images taken from different views, needing more than one camera video as input.

Stick figures and 2D shape models can be extracted from data in approaches without a priori models. Examples with stick figures include data obtained with LEDs(moving light displays, first proposed by Johansson in 1975 [26]) placed on a human body (mainly at elbows and shoulder) dressed in black in front of a black background. In addition to the reading of these spots, it is needed the computation of the relative information to create the connections in order to differentiate the actions. Markers provide a similar information than the LEDs but it doesn't need to compute the extra information. Finally, another used stick figure is obtained from skeletonization.

In the case of 2D shape models, for every body segment, ribbons, blobs, contours or templates have been used. Joints are detected among the connected or close ribbons, or in the center of the overlapping area of 2 connected contours, etc.

In the next subsections we are going to review the main methods for human action recognition published without a priori model and with the observation of the whole body in time.

2.1 Optical Flow Based

Optical flow computation [19] assumes that the intensity structures found locally in the image, remain approximately constant over time at least during small intervals of time. Some methods, compute a histogram of intensities in a window around a pixel. Then, the similarity with spatially close windows in the next frame is computed. The optical flow direction is taken as the direction of the window with maximum similarity. The optical flow information enables to know the probable displacement of the objects present in a scene. Optical flow computation can give undesired results in presence of soft or homogeneous surfaces and discontinuities, particularly for tracking purposes. However, the optical flow feature is independent from changes in appearance (persons wearing different clothing), compared to other measures like spatial or temporal gradients. An overview on optical flow techniques for human motion estimation and recognition can be found in [27].

There is a group of methods for human action recognition that use optical flow computation [10,22]. In [22], a SVM classification technique is applied over the KTH database, using a shape descriptor, optical flow and global temporal information. The probability of boundaries (Pb, [34]) is computed based on Canny edges computed per frame. It is claimed that the Pb features compared to the result from standard edge detection methods, delineate better the boundaries of objects and eliminate the effect of noise caused by shorter edge segments in cluttered backgrounds to a certain degree. Using the detected boundaries the human silhouette is localized with the Hough Transform which is applied to fit edges to the boundaries. First, the shortest edges (encoding details of the shape) and the longer edges (coarse shape information) are used to build a histogram on its orientation and location. To determine the location, histograms are computed in spatial grids (3 by 3 grid per

frame), and bins of 15 degrees are used for a total of 12 (See Fig. 5). The final shape feature per frame will have dimension 108 (12 x 9) after concatenating features for all bins and windows in the grid. To reduce dimensionality of the shape features, an entropy of every feature is measured. The highest values of entropy means the value is changing through the performing of the action. For that reason, these values are considered to differentiate the action classes better so they are kept. Then, for every action a different set of shape features are used (around 30-dimensional feature vector).

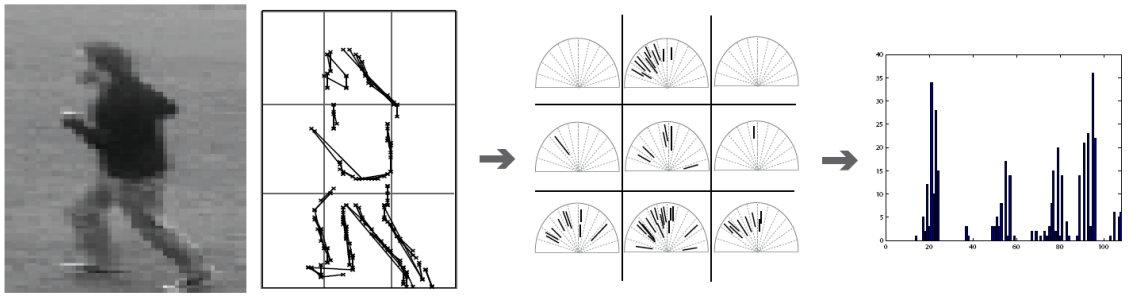

Fig. 5. Shape feature extraction with spatial and orientation histograms. Image taken from [22]

As a feature of motion, orientation histograms of optical flow values are computed in blocks in every frame, by matching it with the previous frame. Histograms are used to get a more compact representation. The histograms have information as the associated spatial position and orientation. For each spatial bin and direction $\theta \in \{0, 90, 180, 270\}$, the optical flow histogram value $h_i(\theta)$ is computed like:

$$h_i(\theta) = \sum_{j \in B_i} \psi(\upsilon_\theta * F_j)$$

, where F_j is the flow value in pixel j, B_i is the set of pixels in bin i, υ_θ is the unit vector in direction θ and ψ is a function like

$$\psi(x) = \left\{ \begin{matrix} 0 & if\ x <= 0 \\ 1 & if\ x\ > \ 0 \end{matrix} \right\}$$

One SVM classifier is trained for shape and motion features separately. Also, for every action a separate classifier is trained to detect if the features belong or not to the action class, and a voting scheme is used to take the final decision. Videos are decomposed in blocks of 7 frames with an overlapping of 3 frames, for feature computation. Every block is classified independently. An overall accuracy of 94% is shown with the method.

In [10] a human motion representation is proposed based on optical flow, shown in images containing the human with around 30 pixels tall. Images with this resolution can be used to recognize actions like jumping, kicking, running, etc, ex. in a football game. It is considered a medium level resolution, having the high resolution with humans around 300 pixels tall to easily differentiate the head from limbs and torso. On the other side, the low resolution images, with around a 3 pixels tall human, can be used in estimating pedestrian traffic and its not able to differentiate among too many action categories.

Optical flow is considered a noisy feature, and for that reason in some solutions a histogram of values over image regions or a smoothing of the values like in the mentioned contribution, is applied to have a more stable spatio temporal descriptor. A nearest neighbor classifier was used here to recognize among different actions from a database. The results are not compared with other methods.

2.2 Spatio Temporal Templates

Another human motion representation is called temporal templates first introduced by Bobick and Davis [4]. They are static vector-images where the vector value at each point is a function of the motion properties at the corresponding spatial location in an image sequence. Based on that, two basic representations are used, a simple two component version of the templates that consists of a motion-energy image (MEI) and a motion-history image (MHI). The MEI is a binary cumulative motion image and the MHI have at every pixel a function of its motion history. The templates are matched against the stored templates of known actions. However, the use of templates is too sensitive to the different movement durations.

In [8] a star figure enclosed by a convex polygon is built to represent the extremities of the silhouette of the human body. A video is then represented as a sequence of these star figure's parameters, which is regarded as a spatio temporal template. The centroid of the silhouette is used to join all points in the convex polygon intersecting the shape. The set of points in the polygon are clustered to eliminate redundant extremity points and to split clusters with large standard deviations (See Fig. 6). The length of the final sticks and orientations are used to model the human actions. An average precision of 88% is achieved by this solution.

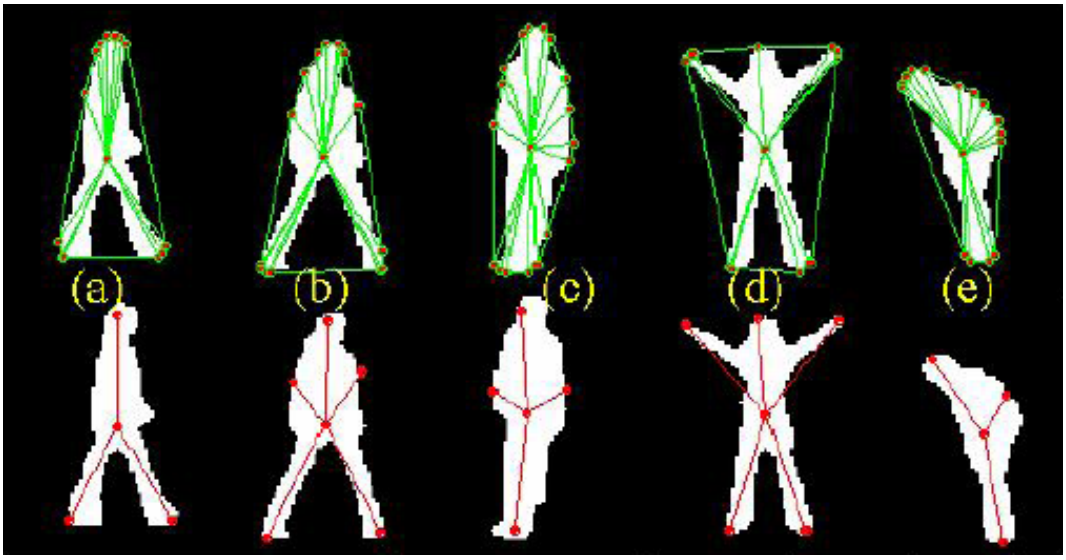

Fig. 6. Result from clustering of extremity points. Image taken from [8]

2.3 Detecting Key-Frames Approach

In a video sequence it is considered that appears a lot of repeated frames. On the other side, there is a set of frames that does not appear so frequently and that is considered that identify the action. The idea of this approach, is to find these key-frames in a video sequence as representations of the action and use them for comparison purposes [5,40,13,14].

This solution lacks information about motion, because most of the cases reduce to a single frame as a representation of the whole video. There is a limited set of actions that can be identified by a single frame. Also, a frame can be chosen in a way that presents ambiguities in the action performed even for a human.

2.4 Space-Time Shape Representation

There is a research group in human action recognition that uses volume-set as a representation of a video. This volume is obtained by concatenating 2D slices over time, converting pixels in voxels(1x1x1 unit), and creating the volumetric space 2D+time. Numerous publications show the usefulness of analyzing actions as spatio-temporal volumes[30,38,45,46,47,20,17,39] (See Fig. 2.4) of some local feature like gradient, optical flow, intensity, etc. In the space-time volume there is both spatial information about pose of the human figure at any time of the sequence and also the orientation of torso and limbs, so as dynamic information about the global motion of the body and relative motion of limbs respect to the torso. In 2007, [15], the authors used properties from the Poisson equation to extract spatio-temporal features as local spatio-temporal saliency, action dynamics, shape structure and orientation. An advantage of this representation is that are easily extendable to a broader set of actions given the appropriate set of data for training. As disadvantage it can not easily generalized changes in appearance and view-point.

Basically, the proposal takes a video (from Weizmann database, section 5) and split them in spatio-temporal cubes made of 8 consecutive frames with an overlap of 4, for two consecutive cubes. The images are first binarized with a simple thresholding technique of colors using the median background for each of the sequences. The average width of the silhouettes were 12 pixels. A translation of the center of mass of the silhouettes to a reference point is performed in order to compensate for changes of speed due to different orientation of motion respect to the camera. Also, all frames are process to have a uniform scale in space. However, still changes in speed of the real action will create more elongated patterns in the volume that could limit the recognition results.

On each of the cubes descriptive features are computed using an extension of the *Poisson Equation* used for classification of 2D silhouettes[16]. The formula assigns to every internal point of the shape the mean time required for a random walk beginning at the point to hit the boundary. The computed features have the advantage that are robust to noise in the boundary of the extracted silhouette compared to other measures like minimum distance to the boundary. In this case, the assigned value depends on various points on the boundary making it more stable to noise. In figure 8, sample volumes of different human actions with the output of computing this feature is shown. The pattern of colors for visualization goes from blue (for the lowest values) to red (for the highest values). The figure shows that pixels in the center of the figure get the higher values, and the ones

in the limbs, head and close the boundary, the lowest. In order to include the boundary pixels around the center in the region of the higher values the gradient feature is taken into account. In this way, a differentiation of the center region from the limbs is achieved.

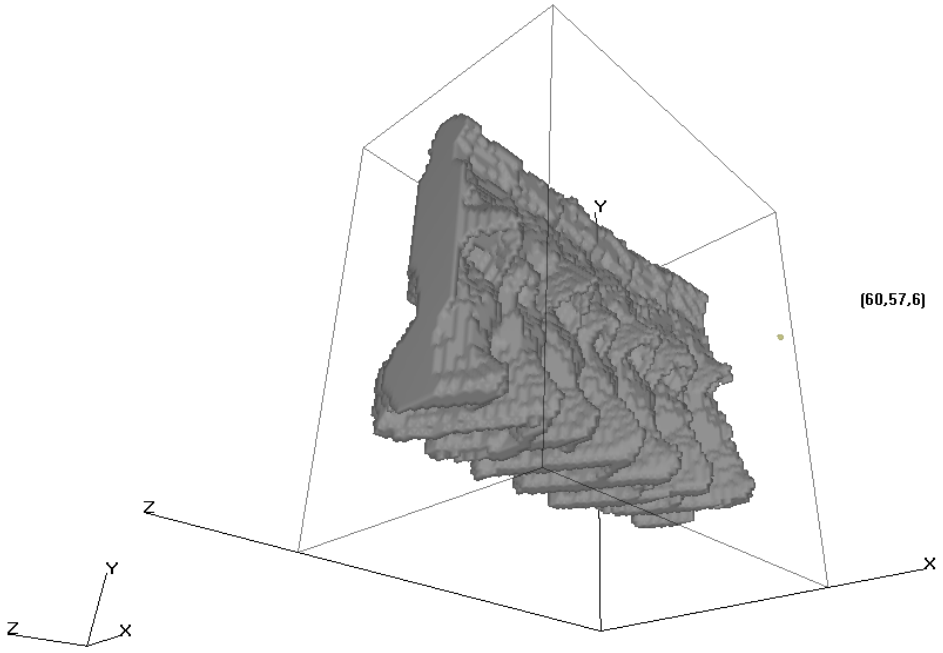

Fig. 7. Volume set of a walking action performed by an actor of the database published by [15]. The volume is visualized using Voxelo[36]

The time domain creates new shapes that don´t occur in the spatial one such as spatio temporal *stick*, *ball* and *plate*. Similarity to these new shapes is estimated from the values obtained by the *Poisson Equation*, using the eigenvalues and eigenvectors of a 3 x 3 Hessian matrix. The final similarity is weighted by the deviation to each of the three principal axes directions to identify from temporal, vertical and horizontal sticks of plates. Also, the saliency feature is extracted in order to identify fast moving parts of an action. These features are then combined in a global representation of the action by means of weighted moments(126 dimensional feature vector), and used for action classification and clustering.

In the case of action classification the procedure is to take every sequence and remove from the database all its spatio-temporal cubes. In the database still are included the cubes from other 8 different persons performing the same action. Then each of the extracted cubes is compared with the remaining ones in the database, and is assigned to the action of the most similar one using a nearest neighbor approach with the Euclidean Distance. The results obtained by the authors are shown in figure 9 with the more conflicting cases between actions skip(a3) and jump(a5). In general it misclassified 20 from 923 cubes resulting in 2.17 percent error rate.

For action clustering purposes they defined a variant of the Median Hausdorff Distance between two video sequences. The formula uses the similarity between every pair of cubes

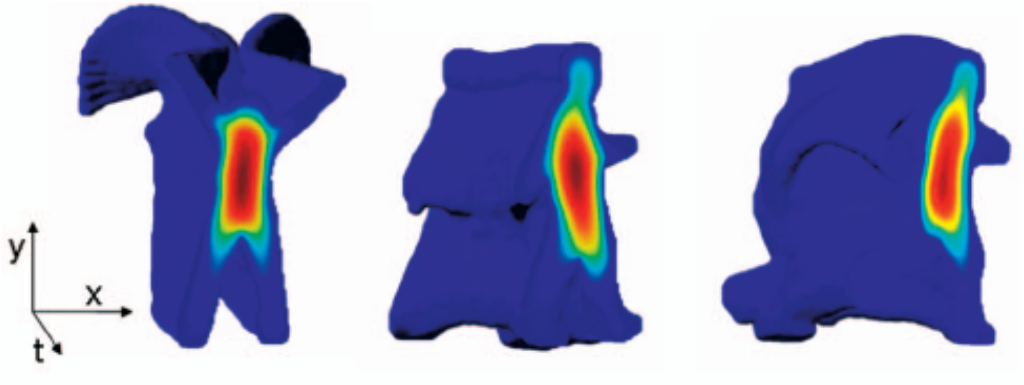

Fig. 8. Sample image from [15] with values extracted from the *Poisson Equation* for volume sets of actions

of the sequences and uses median criteria to allow more flexibility in cases or occasional occlusions or imperfections in the space-time shape. Then applied a spectral clustering technique with the misclassification of only 4 of the sequences. Sometimes a single space-time cube contains enough information to identify an action and this fact is shown in experiments for detecting a particular movement in a ballet dancing. However, in more general applications the combination of information from all the spatio-temporal cubes will give better results.

This approach doesn´t consider sequential order of the spatial shape in time for the representation. For that reason, symmetrical actions performed in backwards would not be differentiated. In experiments was shown that the solution was robust to minor occlusions, certain degree of scale and viewpoint, irregularities in the performance of an action and that was not limited to cyclic actions. However, this proposal is not suitable in presence of strong variations in viewpoint and occlusions since the spatio-temporal shape would vary considerably.

2.5 Space-Time Interest Points

Space-time interest points have been proposed to compensate variations in relative motion between an object and the camera when computing space-time descriptors[32]. This method uses interest points to adapt features to the local velocity of the image pattern to make it stable to different amounts of camera motion.

This approach is based on the fact that features as optical flow and spatio-temporal gradients, depend on the spatial resolution of the pattern and the motion relative to the camera. It follows trying to find a video representation more stable to different image acquisition conditions. These representations are based on the location of space-time interest points, which are stable locations corresponding to moving 2-dimensional image structures at moments of non-constant motion (See Fig. 2.4). These locations are computed maximizing a measure of the local variations in the volume-set (2D+time).

A local velocity and scale estimate is measured on each interest point location. Then, their locations are adapted to obtain invariant interest points to the previous variations.

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
a1	100	0	0	0	0	0	0	0	0	0
a2	0	98.0	2.0	0	0	0	0	0	0	0
a3	0	2.9	97.1	0	0	0	0	0	0	0
a4	0	0	0	100	0	0	0	0	0	0
a5	0	0	10.8	0	89.2	0	0	0	0	0
a6	0	0	0	0	0	100	0	0	0	0
a7	0	0	0	0	0	0	100	0	0	0
a8	0	0	0	0.9	0	0.9	0	94.8	3.5	0
a9	0	0	0	0.9	0	0	0	1.9	97.2	0
a10	0	0	0	0	0	0	0	0	0.9	99.1

Fig. 9. Errors obtained by authors in[15] for action classification. From a1 to a10 are the labels for walk, run, skip, jack, jump, pjump, side, wave1, wave2 and bend actions respectively

Resulting from this, we can observe more stable corresponding interest point positions in videos taken with different motion velocities between the camera and the human (See Fig. 11).

Interest points are affected by camera view point product of the variations in shape of the projected 2D figure of the human. See for example the 2D moving surface of a moving person walking toward the camera would not show much from the legs motion that created the interest points in the previous figure. However, it has been used as a stable feature to velocity and scale variations in the video capture process.

In the neighborhood of the computed interest points, descriptors are computed to compare local events. Euclidean distance was used to compare descriptors between matched points showing improvements when using velocity adapted interest points over two methods (1-without velocity adaptation, 2- with only one iteration of velocity adaptation).

In order to compare two image sequences a greedy matching strategy is performed, were a pair of correspondent interest points are selected repeatedly, based on the minimum euclidean distance between their features. The final distance is then the sum of the partial euclidean distances from the N best matches. Later, a Nearest Neighbor strategy is applied to recognize the action in the video sequence.

For human action recognition in general, videos were used simulating different velocities in camera capture process from videos taken from a stationary camera. If the recognition with 1-NN corresponds to a video with the same action, then it was considered correct. Experiments in this approach shows a recognition rate of about 80%.

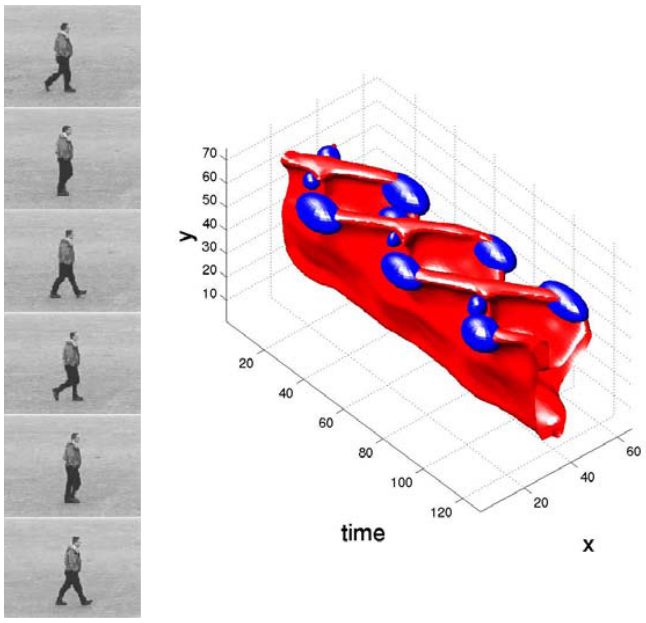

Fig. 10. Moving surface of a walking person silhouette. Interest points are shown with ellipsoids. Image taken from[32]

2.6 Bag-of-Words Approach

These methods represent actions as sets of *words*, where a *word* refers to the pose in a frame or any other local feature in the video sequence [11,23]. Also, images could be represented as a set of regions without any spatial relation for object recognition or image retrieval tasks.

Usually, methods similar to the ones applied for document retrieval are used for matching. In this way, only vectors with the frequency that every *word* has in the document (video) are used as representation. In [38] an example is presented based on space-time interest points used in an SVM (Support Vector Machine) based method. The interest points are related with motion events in the video and can be adapted to size, speed and frequency of the moving pattern. The video is then represented as a set of interest points, and around them features are computed like histograms that are the input of a SVM for recognition. The database use for evaluation is the KTH (See Section 5), containing 6 human actions performed by 25 persons in 4 different scenarios(outdoors, outdoors with scale variations, outdoors with different clothes and indoors) in 2391 sequences.

In [23], a human body is represented by a set of oriented rectangles in the spatial domain. The orientations of these rectangles in time are considered to determine the action performed. The method proposed is called *bag-of-rectangles* to represent human actions in video, because of the similarities with the named *bag-of-words* approach. Histograms are built to encode the spatial distribution of the rectangles, instead of detecting the configuration of parts. Different approaches are used with this representation including SVM, and frame by frame voting, matching frames independently. The authors report a 100% acuracy for the Weizmann database (See Section 5).

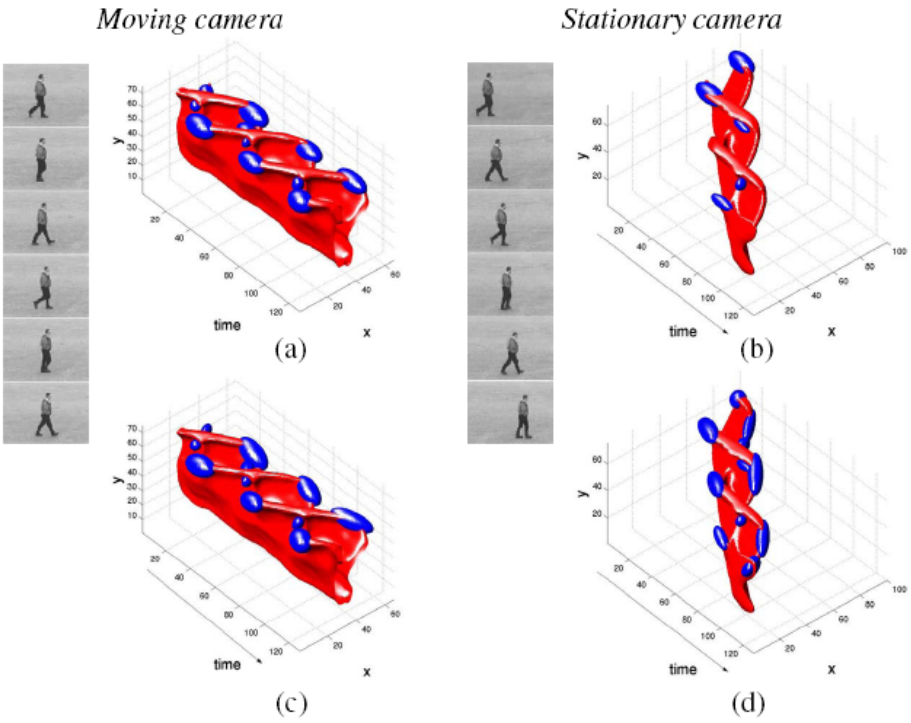

Fig. 11. Interest points computed from videos of the walking person, captured from a moving and stationary camera. (a-b): without velocity adaptation, (c-d): velocity adapted. Image taken from[32]

Based on this approach there is a recent contribution in [18], which represents actions as a document giving importance to the temporal characteristics of actions (pose sentences). Then, for matching of two action sequences the authors use string matching techniques obtaining 92% of performance. The use of the order in the words here improves the differentiation between actions as walking-running, which *bag-of-words* approach confuses. Yet, not all the bag-of-words approaches behaves worst than this solution, neither this contribution improves approaches like based on spatio-temporal shapes. A similar solution is presented in [21] with posture classification rate of 95.16% and bejavior analysis for abnormal activity detection analysing 10 behavior types, of 94.5%. The different behaviors include gymnastics, walking. squatting, stooping, sitting, laying, falling, picking up, jumping and climbing from a personal database of 450 video sequences.

In [9], a similar solution is presented with purposes of categorization of action sequences. The first step is to binarize the video sequences assuming that a background substraction technique can be applied and the result is a sequence of silhouettes. These silhouettes are normalized to a uniform size and are centered, such that changes in scale because of different distances to the camera don't affect the feature extraction process. Then, every frame is represented as a raw vector and a nonlinear dimensionality reduction is applied to reduce the high dimensionality of silhouette features. The reduction is made by means of a kernel PCA (Principal Component Analysis) with a Gaussian kernel (parameter set to 1000 in experiments) on all frames from all videos in the database.

At this point video frames are represented with a 25-dimensional vector. This parameter is selected adhoc and is very likely that different action conditions would require different dimensionality in the reduced vector. A k-means clusterization method is applied on the actual vectors to obtain a set of k labels (20, 25 and 30). Now, every video is a sequence of labels where usually neighboring frames should have the same label. Taking this in consideration, to remove noise they eliminate from the sequence isolated labels. It could be one, two, or more consecutive isolated frames with the same label depending on the parameter *ii*. (See Fig. 12) The parameter *ii* should be adjusted for a particular set of video sequences depending of the speed of the performed action. In some cases 3 consecutive frames could be very small and very likely to be noise but 5 no, and in other cases 10 or 20 frames could be considered still noise if the camera capture rate is very high or if the speed of the action is slow enough.

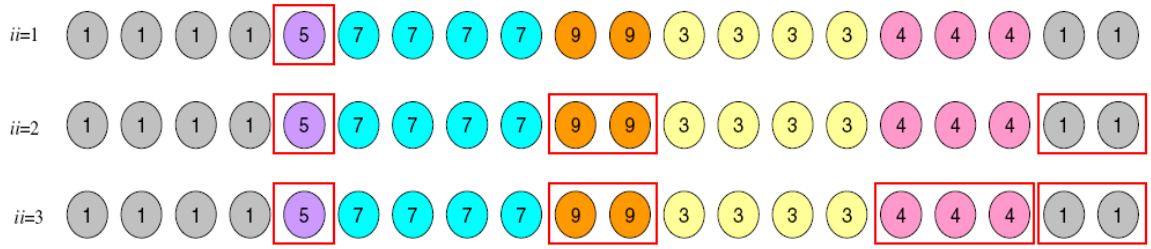

Fig. 12. Label sequences with samples of removed noise with different parameter ii. Image taken from [9]

In this way, every frame was assigned to a label (*word*) but the sequence order is still considered. To compare two sequences a Levenstein edit distance is applied. This is particularly useful here because it considers the order in the occurrence of the movements, but also that the same action could be recorded in different starting points, particularly with cyclic actions like walking and running.

Using this distance, a matrix M is computed taking in consideration the duration of the sequences compared: $M_{i,j} = d_{ij}/max(T_i, T_j)$, such that T_i and T_j are the respective lengths of sequence i and j. Taking the lengths in consideration the errors included because of different speed in the performing of the actions are reduced, mainly in cases that both compared sequences are performed with the same speed. However, when the compared pair are with very different speed it could lead to undesired distances. Consider for example the sequence of figure 12, and one of the same action but performed very fast such that we can only see one frame of each label like 1579341. Here, the edit distance would be the number of frames we need to add (13 repeated in the figure). Considering we don't remove any frame from the sequence, the normalized result would be $13/20 = 0.65$ which would be a very big value of dissimilarity for two sequences of the same action.

The matrix M is used to perform spectral clustering to reduce the number of classes to c. The affinity matrix is computed from the distance matrix like $A_{i,j} = exp(-M^2_{i,j}/\tau^2)$, where τ is a scale factor. Then, with the diagonal matrix of A (D), the normalized laplacian matrix L is built like $L = D^{-1/2}AD^{-1/2}$. The bigger c eigenvectors are selected from this matrix and with them a matrix called E is built. E is normalized per raws such that they

have unit length:

$$G_{ij} = E_{ij}/(\sum_j E_{ij}^2)^{1/2}$$

The set of normalized raws are clustered in c groups by $c-means$. Finally, every video is assigned the group label of its corresponding raw in G. In experiments c is selected to be the number of actions the system wants to recognize.

This approach compared to the frequency based one that don't consider the sequence order and its similar to the ones used for document retrieval, is very slow. However, the accuracy is improved as it was shown in the referred publication.

2.7 Shape Flow Representation

A flow line is the space-time line of a tracked object point over time. A shape flow is an assembly of flow lines representing the shape and motion of the object in the video sequence. As an advantage, shape flows are independent from appearance. Flow lines are individually unreliable, but the shape flow is a more robust motion representation.

This approach is closer to the volumetric solutions as tries to find a pattern of the motion in the volume set and find a feasible matching strategy [24]. Flow pattern lines are matched and also the spatial consistency is checked using a Delaunay graph of randomly selected flow lines starting from edge points in the first frame of the video. Every flow line needs to have the same number of sample points equal to the number of frames. The similarity between two flow lines is computed by means of the Euclidean Distance of corresponding points, after a spatial normalization of its coordinates. However, because of computational issues, the authors just selected a few lines from the shape flow for comparison and created a Delaunay graph forcing anyhow to apply optimization techniques to reduce computational times.

3 New Proposed Taxonomy

The main approaches for full-body human action recognition are discussed in this report. They can be distributed in 2 main categories depending on the style of representation. The first is the one that considers a video as the whole set of frames, and the second reducing the number of frames to get a more compact representation (See Fig. 13).

In the group of methods that reduce the cardinality of the representation is included the *Key Frames* approach, that extracts from the whole sequence a reduce set of representatives of the action (in particular only one frame). Also, the spatio-temporal templates and the eigenshapes are included. Reducing the cardinality the representation looses temporal information.

The second group, includes the *bag-of-words* approach. This approach considers the whole sequence but with the particularity that every frame is treated independently. Features are extracted from each of them, or any local feature is extracted from the sequence that is considered usually without any sequential order. The optical flow sequence or shape

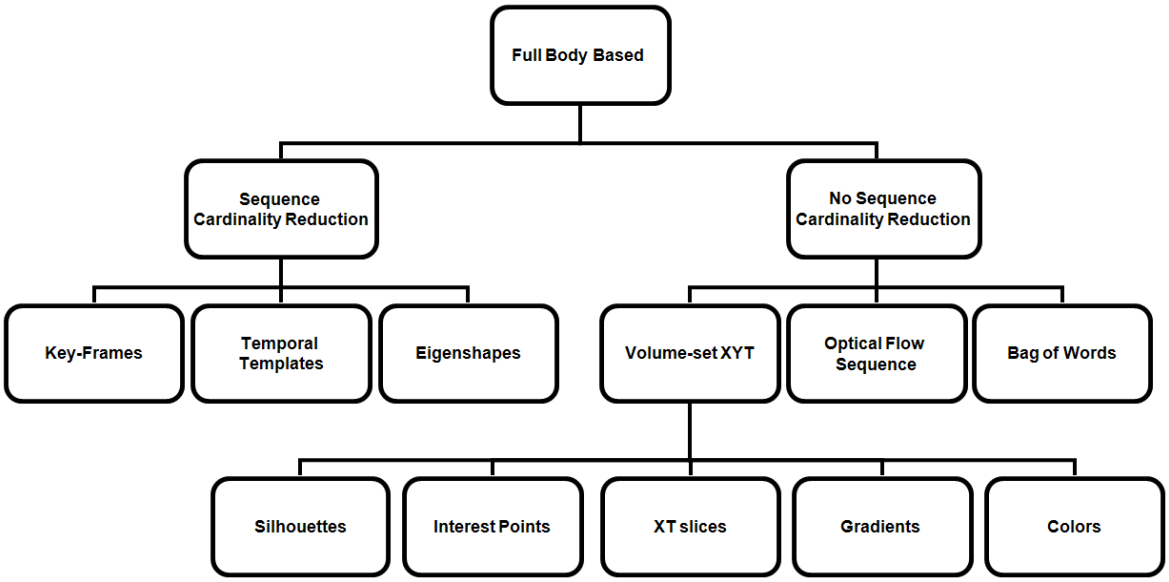

Fig. 13. Taxonomy of the main methods for full-body human action recognition

flow sequence consider the set of flow lines in the whole sequence as a representation of the action. Finally, volumetric representations are also included.

The difference of the optical flow and the volumetric solutions is that optical flow is a value extracted from the transition from frame to frame, and the volumetric features are computed inside of the volumetric shape of sequential set of frames itself. Included in this group are the spatio-temporal interest points feature computation, the spatio-temporal shape features, the XT slices which are obtained in the XYT cube near the human ankle as a signature of the walking pattern, and the ones that correlates a segmentation of the cube by features like colors or gradients [44].

4 Open Problems

Recognition of human actions in a general setting is a very hard problem. Actual solutions focus in particular databases and restrict the solution to particular conditions. Real scenes contain a set of factors that make human motion understanding a tough task [3] and remain today as open problems, like changes in appearance, changes in motion (from camera and object), illumination, moving background, several objects in the scene, occlusions etc. Human's properties such as it is a highly articulated, self-occluded and non-rigid object, plus the temporal and spatial variations of the action make it a challenge from the academic point of view.

Traditional action recognition methods have as limitations that some need to measure optical flow which is difficult because of soft surfaces and discontinuities [33,10,22]. Another group use tracking of features that fail in self occluding situations, change of appearance, scale, etc. Also, there is some work that bases the recognition in the detection of key frames or eigenshapes from silhouettes of foreground [12], but they lack information about

the motion. There is a last group which uses periodicity analysis [6,7] and because of that are limited to cyclic motions.

Another major obstacle in action recognition from images is the variability of the visual data under changing viewing directions. Many activities have a strong spatio temporal pattern of appearance that is used for recognition. However, if the object rotates with respect to the camera the projected shape is different and so the spatio temporal pattern. Some works [11], try to extract features invariant to the 2D projection, considering the features as properties of the 3D real human action. This is not a very exploited issue and remains as an open problem.

5 Available Databases

In the following a list of the actual databases for evaluation:

1. KTH dataset (2003, [38,2]) (See Fig. 14). The dataset consists of 2391 low resolution videos (160x120, 25fps) of 25 subjects and 4 different recording conditions of the videos. There are 6 actions in this dataset: boxing, handclapping, handwaving, jogging, running and walking.
2. Weizmann database (2005, [15,1]). This database contains low resolution videos (180x144, 25 fps) of ten actions, each performed by nine different subjects. The actions are illustrated in Fig. 15.
3. Personal databases like the one used in [41]. It contains 100 videos from 10 actions performed by the same person. The actions include pick up object, jog in place, push, squash, wave, kick, bend to the side, throw, turn around and talk on the cell phone. (See Fig. 16)

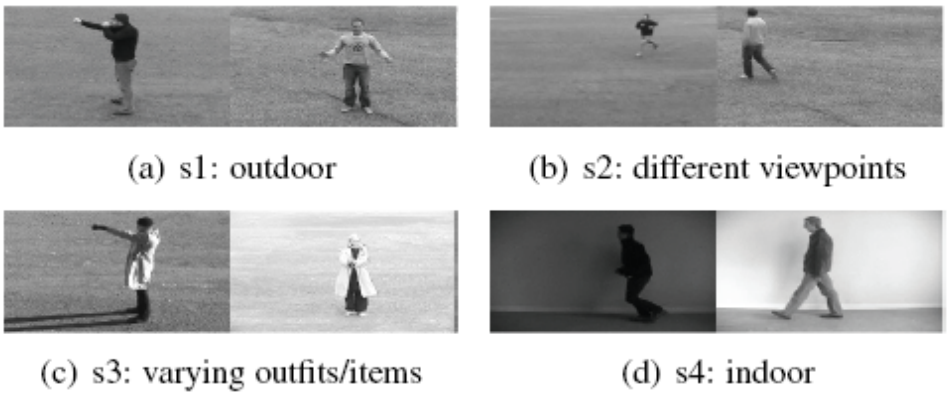

Fig. 14. Samples from database KTH for videos in different conditions. Image taken from[22]

Most of the papers are using mainly the above databases, while others uses personal databases.

Fig. 15. Ten actions from database Weizmann: Bending, Jumping jack, Jumping, Jumping in place ('Pjump'), Gallop sideways ('Side'), Running, Skipping, Walking, Wave one hand ('Wave1') and Wave two hands ('Wave2')

Fig. 16. Samples from database in[41]. Image taken from [9]

6 Main Scientific Conferences

In this section we present a list of the actual main events and journals showing an increasing number of contributions in the subject of human action recognition or motion analysis in video sequences:

1. European Conference on Computer Vision
2. IEEE International Workshop on Visual Surveillance
3. International Conference on Computer Vision
4. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance
5. International Journal of Computer Vision
6. Computer Vision and Pattern Recognition
7. Computer Vision and Image Understanding
8. Image and Vision Computing
9. EURASIP Journal on Image and Video Processing (http://www.hindawi.com/journals/ivp/si/vmar.html)

However, there is a number of important contributions on the field that appears in conferences as the International Conference on Pattern Recognition and journals as the Pattern Recognition and Pattern Analysis and Machine Intelligence.

7 Conclusions

The recognition of human actions in video sequences is a challenge from the academic point of view because of human's properties such as it is a highly articulated, self-occluded and non-rigid object. From the application point of view also represents a challenge mainly when approaching the problem as a non-invasive solution. In the last decades have been and increasing interest in these applications and some progress in understanding human actions and behaviors.

Problems such as changes in appearance, motion, clutter, lighting, occlusions, scale remain affecting actual solutions and are still open problems for the computer vision community. The use of space-time volumes has prove useful to represent both spatial information of the body in the scene in every time through the video sequence, and also the orientation of torso and limbs, so as dynamic information about the global motion of the body and relative motion of limbs respect to the torso.

In this technical report we are reviewing some of the main approaches for human action recognition, considering the view of a full-body single human, and restricting the set of actions to be independent from the interaction with the environment. The results of many studies do not use benchmarks so it is hard to replicate these studies, in other cases the experiments are performed in constrained environments, and comparative results with other methods are missing. Actually this is a very active area of research and a lot of work is still needed to be done to obtain an acceptable general solution to the problem.

References

1. Actions as space-time shapes (action database). *http://www.wisdom.weizmann.ac.il/ vision/SpaceTimeActions.html*.
2. Recognition of human actions (action database). *http://www.nada.kth.se/cvap/actions/*.
3. J. K. Aggarwal. Problems, ongoing research and future directions in motion research. *Mach. Vision Appl.*, 14(4):199–201, 2003.
4. A. Bobick and J. Davis. Real-time recognition of activity using temporal templates. In *WACV '96: Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV '96)*, page 39, Washington, DC, USA, 1996. IEEE Computer Society.
5. Stefan Carlsson and Josephine Sullivan. Action recognition by shape matching to key frames. In *IEEE Computer Society Workshop on Models versus Exemplars in Computer Vision*, 2001.
6. Chee Seng Chan, Honghai Liu, and David Brown. An effective human motion classification approach using knowledge representation in qualitative normalised templates. In *FUZZ-IEEE*, pages 1–6, 2007.
7. Chee Seng Chan, Honghai Liu, and David J. Brown. Recognition of human motion from qualitative normalised templates. *J. Intell. Robotics Syst.*, 48(1):79–95, 2007.
8. Duan-Yu Chen, Sheng-Wen Shih, and Hong-Yuan Mark Liao. Human action recognition using 2-d spatio-temporal templates. In *ICME*, pages 667–670, 2007.
9. Joanna Cheng, Liang Wang, and Christopher Leckie. Dual clustering for categorization of action sequences. In *ICPR*, pages 1–4. IEEE, 2008.
10. A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *Proc. Int. Conf. Computer Vision ICCV03*, 2003.
11. Ali Farhadi and Mostafa Kamali Tabrizi. Learning to recognize activities from the wrong view point. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 154–166, Berlin, Heidelberg, 2008. Springer-Verlag.
12. R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky. Behavior classification by eigendecomposition of periodic motions. *Pattern Recognition*, 38(7):1033–1043, 2005.

13. J. Gonzàlez, J. Varona, F.X. Roca, and J. J. Villanueva. On-line human activity recognition for video surveillance. In *Proc. IX National Symposium on Pattern Recognition and Image Analysis (SNRFAI'2001)*, volume 2, pages 255–260, Castelló, Spain, May 2001.
14. J. Gonzàlez, J. Varona, F.X. Roca, and J. J. Villanueva. A human action comparison framework for motion understanding. In *In 6th Catalan Conference for Artificial Intelligence (CCIA'2003)*, volume 100, pages 168–177, Palma de Mallorca, Spain, October 2003.
15. Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, December 2007.
16. Lena Gorelick, Meirav Galun, and Achi Brandt. Shape representation and classification using the poisson equation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):1991–2005, 2006.
17. Matthias Grundmann, Franziska Meier, and Irfan Essa. 3d shape context and distance transform for action recognition. In *The 19th Inter. Conf. on Pattern Recognition, ICPR08*, page ?, 2008.
18. Kardelen Hatun and Pýnar Duygulu. Pose sentences: A new representation for action recognition using sequence of pose words. In *The 19th Inter. Conf. on Pattern Recognition, ICPR08*, page ?, 2008.
19. Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.
20. Pei-Chi Hsiao, Chu-Song Chen, and Long-Wen Chang. Human action recognition using temporal-state shape contexts. In *The 19th Inter. Conf. on Pattern Recognition, ICPR08*, page ?, 2008.
21. Jun-Wei Hsieh, Yung-Tai Hsu, H.-Y.M. Liao, Chih-Chiang Chen, Yuan Ze Univ, and Chung-Li. Video-based human movement analysis and its application to surveillance systems. *Multimedia, IEEE Transactions on*, 10(3):372–384, 2008.
22. Nazli Ikizler, Ramazan Gokberk Cinbis, and Pinar Duygulu. Human action recognition with line and flow histograms. In *ICPR*, pages 1–4. IEEE, 2008.
23. Nazli İkizler and Pinar Duygulu. Human action recognition using distribution of oriented rectangular patches. pages 271–284. 2007.
24. Hao Jiang and David R. Martin. Finding actions using shape flows. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 278–292, Berlin, Heidelberg, 2008. Springer-Verlag.
25. G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211, 1973.
26. G. Johansson. Visual motion perception. *Science American*, pages 76–88, 1975.
27. S. Ju. Human motion estimation and recognition (depth oral report). Technical report, University of Toronto, 1996.
28. Reinhard Klette and Garry Tee. Understanding human motion: A historic review. Technical Report CITR-TR-192, Communication and Information Technology Research (CITR), Technical Report Series, ISSN 1178-3553, The University of Auckland, New Zealand, 2007.
29. V. Krüger, D. Kragic, A. Ude, and C. Geib. The meaning of action: A review on action recognition and mapping. *Advanced Robotics*, 21(13):1473–1501, 2007.
30. I. Laptev and T. Lindeberg. Local Descriptors for Spatio-Temporal Recognition. In *European Conference on Computer Vision, Workshop "Spatial Coherence for Visual Motion Analysis"*, 2004.
31. Ivan Laptev. Action class detection and recognition in realistic video. *Int. Conf. Computer Vision ICCV07*, 2007.
32. Ivan Laptev and Tony Lindeberg. Velocity adaptation of space-time interest points. In *ICPR '04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 1*, pages 52–56, Washington, DC, USA, 2004. IEEE Computer Society.
33. M. Lucena, J.M. Fuertes, and N. Perez de la Blanca. Using optical flow for tracking. In *Iberoamerican Congress on Pattern Recognition, CIARP03*, volume 2905 of *LNCS*, pages 87–94. Springer-Verlag Berlin Heidelberg, 2003.
34. David R. Martin, Charless C. Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):530–549, 2004.
35. Thomas B. Moeslund, Adrian Hilton, and Volker Kruger. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.*, 104(2):90–126, November 2006.
36. Javier Sánchez Pelaez and Pedro Real. El modelador volumétrico voxelo. 2003.
37. Ronald Poppe. Vision-based human motion analysis: An overview. *Comput. Vis. Image Underst.*, 108(1-2):4–18, 2007.

38. Christian Schüldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local svm approach. In *In Proc. ICPR*, pages 32–36, 2004.
39. Yuping Shen, Nazim Ashraf, and Hassan Foroosh. Action recognition based on homography constraints. In *The 19th Inter. Conf. on Pattern Recognition, ICPR08*, page ?, 2008.
40. Xavier Varona, Jordi Gonzàlez, F. Xavier Roca, and Juan J. Villanueva. Automatic selection of keyframes for activity recognition. In *AMDO '00: Proceedings of the First International Workshop on Articulated Motion and Deformable Objects*, pages 173–181, London, UK, 2000. Springer-Verlag.
41. Ashok Veeraraghavan, Rama Chellappa, and Amit K. Roy-Chowdhury. The function space of an activity. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 959–968, 2006.
42. Jessica Junlin Wang and Sameer Singh. Video analysis of human dynamics – a survey. *Real Time Imaging*, 9:321–346, 2003.
43. L. Wang, T. Tan, H. Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 25(12):1505–1518, 2003.
44. Rahul Sukthankar Yan Ke and Martial Hebert. Spatio-temporal shape and flow correlation for action recognition. In *Visual Surveillance Workshop*, June 2007.
45. A. Yilmaz and Mubarak Shah. Actions sketch: A novel action representation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, volume 1, pages 984–989, 2005.
46. Alper Yilmaz and Mubarak Shah. Actions as objects: A novel action representation, 2005.
47. Alper Yilmaz and Mubarak Shah. A differential geometric approach to representing the human actions. *Computer Vision and Image Understanding*, 109(3):335–351, 2008.

RT_021, marzo 2010
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: DCG Matilde Galindo Sánchez
RNPS No. 2142
ISSN 2072-6287
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. No. 21812 e/218 y 222, Rpto. Siboney, Playa;
Ciudad de La Habana. Cuba. C.P. 12200
Impreso en Cuba