Workshop Proceedings of the 10th International Conference on Intelligent Environments
J.C. Augusto and T. Zhang (Eds.)

doi:10.3233/978-1-61499-411-4-4

Recognizing Actions of Humans in Motion for Smart Environments

Oleg STAROSTENKO[a,1], Roberto ROSAS-ROMERO[a], Jorge MARTÍNEZ-CARBALLIDO[b], Vicente ALARCON-AQUINO[a], J. Alfredo SANCHEZ[a]
[a] *Department of Computing, Electronics and Mechatronics, Universidad de las Americas Puebla,* Cholula, *Puebla, 72810, Mexico*
[b] *Instituto Nacional de Astrofísica, Óptica y Electrónica Tonantzintla, Puebla, México*

Abstract. Development of high precision systems for recognition of human actions directly from video records is still open problem. Recently, in smart environments the recognition of dynamic actions of human in motion receives a particular interest. We propose two approaches for human action recognition. In the first approach, the envelope of 30x30 pixels is applied to enclose invariant to dimensions human silhouette separated from background. Once the area with located figure is defined, the image sequence is used as input of convolutional neural network that extracts global figure features without previous image processing. The second proposed approach is based on natural knowledge of the human figure such as proportions of body and position of feet. Together with processing global features, we extract six local features combining in this way the holistic and cluster-based approaches for representation of human figure. The input sub-sequence of previously aligned binary silhouettes from video frames is processed to concatenate local and global features into a single feature vector feeding hierarchical system of three linear support vector machines for human action classification. In order to evaluate the proposed approaches, two frameworks for recognizing human actions such as walk, jump, run, side and skip have been designed and tested on Weizmann standard and proper developed datasets achieving correct classification rate of 97-100%.

Keywords. Human action recognition, feature extraction, multi-class classifiers, convolutional neural network, machine learning

Introduction

Recently emerging intelligent applications for automatic human action recognition (HAR) may sense, interpret and predict automatically environmental conditions describing events in videos or images. Lately, the machine learning and pervasive sensing technologies for recognizing actions and activities of humans in motion are widely used as essential part of smart environments developed for video surveillance, ambient-assisted living, human-computer interaction, personal healthcare, gaming industry, pedestrian detection by intelligent vehicles, etc. [1-4].

Recognition of human activities and interpretation of ambient scenes by autonomous systems is still challenging task. Scientific communities have joined

[1] Corresponding Author. E-mail: oleg.starostenko@udlap.mx

efforts to propose and develop smart environments in context of automatic recognition of running, walking, rotating, jogging, skipping, boxing, hand clapping, waving, etc. However, despite different efforts, most of the proposed approaches to date are quite complex and usually do not provide fast and high precision ambient analysis in real time. Currently, several relevant systems for human action recognition achieve recognition rate about of 70-100%. However, these reports only partially solve this still open problem of automatic human action recognition and human activity classification.

As usual, human action recognition from video sequence consists of the following steps: *1)* image acquisition, *2)* low level image feature extraction, *3)* medium level descriptive action extraction generating a set of action primitives and *4)* high level semantic extraction from primitive actions [1], [5], [6]. The extraction and processing of significant quantity of features makes slow classification and recognition processes. Some researchers report the reduction of processing complexity operating only with one feature levels particularly, adopting knowledge based models for interpretation of global image content [7].

For this reason, in this paper we seek to collaborate in the design of two HAR approaches, which are characterized by their simplicity, high precision and speed. In the first approach we evaluate an ability of convolutional neural network to classify the human actions without previous image processing using only global information about whole body. In order to increment the precision of recognition, in the second proposal we add previous analysis of natural knowledge about human figure such as the proportions of body and the number of feet touching the ground. The analysis is characterized by processing a small fragment with human silhouette in motion, extracting reduced number of features per frame (only six features instead of hundreds of them in other methods).

This paper is organized as it follows. In section 1 the related works for human figure representation and analysis of classifiers used in contexts of HAR are discussed. In section 2 two proposed approaches for human action extraction are presented. Section 3 describes designed CNN-based and multi-class SVM classifiers. Experiments and results are presented in section 4. The evaluation and discussion of results are described in section 5 concluding the paper with analysis of contributions and definition of future works.

1. Related works

1.1. Modeling human silhouette

Several relevant reports about representation and modeling human figure may be subdivided into two general groups: one is based on holistic representation and another uses cluster-based representation. These two approaches define a number of features that have to be taken into account during representation and indexing of human silhouette in HAR applications. Particularly, the methods based on holistic representation use the information about whole human figure to characterize an action. In the case of cluster-based methods, a feature extraction process is based on the patch-by-patch inspecting parts of human figure generating significant number of independent descriptive vectors that require complex processing and interpretation.

Into each group of methods for representation of human figure there exist many models, which differ by type, quantity, dimensionality and domains of silhouette features, ability for recognizing multiple actions, invariance to perspective, complex background, variable illumination, presence of occluded bodies and others [2], [7]. It is important to mention some relevant methods, which represent human figure as: three-dimensional objects generated by grouping of silhouettes in volumes on the space–time domain [8], or as a temporal sequence of deformed centroid-centered silhouettes [9], or as a sequence of parameters from star figures represented as Gaussian mixture models [10]. Other works have used whole region of interest recognizing human actions analyzing shape information and their movement on raw images [11], generating 3D volumes described by spatial and temporal maps measuring energy through derivatives [12] and processing rectangular image patches for extraction of parts of silhouette that are more representative for recognize human action [5].

As it has been mentioned in relevant models for representation of human body, the human actions may be sub-divided into some approaches: *1)* non-parametric, when features are extracted from each frame of video (used techniques are: dimensionality reduction, template matching, motion energy analysis); *2)* volumetric that consider the video as a 3D volume intensity pixels (used techniques are: space time filtering, parts based, sub-volume matching, tensor based, Gaussian kernels, Gabor filter banks, spatial-temporal gradient); *3)* parametric that analyze temporal dynamics of the movement (used techniques are: Hidden Markov Models (HMM), linear dynamic systems, non-linear dynamical systems) [2], [5], [13].

Finally, for classification of high-level hierarchical human activities in complex scene, recognized human actions are used in such models as: graphical models (Bayesian network, dynamic belief networks, Petri nets); syntactic models (context or stochastic context free grammars) and knowledge-based models (logic reasoning, ontology, event representation) [5], [13]. Table I resumes some approaches proposed by various researchers for recognizing human actions and activities.

Table 1. Comparative analysis of approaches for recognition of human actions and activities

Authors	MHAR	HAR approach	Invariance	RL	3D	Body model
Kim et al. 2010 [6]	X	independent component analysis, linear discriminant analysis with HMM	ND	low	3D	body parts
Meng et al. 2010 [14]	X	spiking neural networks, interest point detection (corners)	no	medium	3D STD	whole body
Ji et al. 2013 [15]	X	filter responses, convolutional neural networks (CNN)	view and size	medium	3D	just head area
Thang et al. 2014 [16]	X	3D histogram skeletons	ND	medium	3D	whole body
Rad et al. 2012 [17]	X	Support vector machine (SVM)	aerial perspective	medium	3D	whole parts
Zhou et al. 2012 [18]	X	pondered strategy of feature learning	ND	medium	STV	whole body
Luo et al. 2013[19]	X	SVM and Naive Bayesian	4 views	medium	MVs	whole body
Goudelis, 2013 al.[20]	X	trace transform, radial basis functions, Kernel SVM	capturing variations	medium	3D	whole body
Lu et al. 2013 [21]	X	spectral embedding of low-level features, latent semantic learning, SVM	scale and position	medium	3D	whole body

Note: MHAR means Multiple Human Action Recognition, RL- Recognition Level; ND – Not Defined, STD – Space Time Domain; STV – Space Time Volume, MV – Multiple Views.

Principal disadvantages of the well-known models and approaches are limitations in body tracking and 3D feature extracting, great complexity of scenarios with multiple objects in motion and recognition of concurrent activities, low recognition rates of nonstationary actions, susceptibility to changing backgrounds and occlusions, loss of power of discrimination for complex activities, requirement of large number of training videos, dependence on using fixed cameras, etc. [1], [7], [13]. For example, general methods based on the holistic description are susceptible to noise, occlusion and view-point variation so that, these methods work properly on controlled environments. Cluster-based methods tend to be more robust to noise, occlusion and in some cases invariant to rigid transformations however, these methods present a significant disadvantage is that: a size of description vector is usually very large and variant according to the number of patches used for human figure representation.

Because of holistic and cluster-based representations have different strengths and weaknesses, some researchers have used a mixed representation of a human figure [22], [23]. As a consequence, we propose to use the hybrid approach for human figure description and recognition of dynamic human actions, where displacement of human in video takes place.

1.2. Selection of supervised learning models for HAR classification

In two proposed approaches introduced in this paper, we exploit supervised learning models with associated learning algorithms that analyze input video and recognize patterns used for classification of human actions. Particularly, in the first approach the classification of actions is provided by convolutional neural network and in the second one the classification is performed by a hierarchical system of SVMs.

NN and SVM are typically similar learning models. If a neural network tries to find the parameters, which minimize the mean squared prediction error with respect to a set of training examples, a SVM tries to minimize both training error and some measures of hypothesis complexity. Another difference between them is that, in NN stochastic gradient descent isn't guaranteed to find the optimal set of parameters however, any decent SVM implementation will find the optimal set of parameters. On the other hand, neural networks have certain invariance to scale, position and occlusion of objects in images as well as to spatial-temporal noise that usually is presented in video records of live scenes [5], [24].

In Table 2 the results of analysis of HAR applications that use supervised learning models (NN and SVM) are summarized. Each reported application has its particular approach for feature extraction however, they share common steps such as input video processing, motion detection, low level features analysis, descriptive pattern or primitive modeling and classification by NN or SVM.

The feature extraction is not a trivial process, and ideally it will be learned during training of the weights from classification layer of NN. So, the learning filters used for feature extraction must allow reliable and precise classification. For this type of tasks there exists a particular kind of neural networks called convolutional neural networks (CNN) that manage images without preprocessing, directly extracting medium level features (descriptive primitives) and providing simultaneously their classification. This is a principal idea of the first proposed approach. In spite of this, during video acquisition the manipulating low level features is required to accelerate segmentation of foreground from background, to detect moving objects and to extract whole figure or its some important parts for particular action. This is a principal idea of the second proposed approach.

Table 2. Comparative analysis of HAR applications using supervised learning models

Authors	Application	Data source	Real Time	Effectiveness	Features
Oliveria et al. 2010 [24]	pedestrian detection	video	not defined	94.6%	combine gradient histogram version variants and LRFs provided by CNN
Szarvas et al. 2006 [25]	people recognition in 3 positions	images sequence	yes	94%	first NN learning layer is substituted by 10x10 Gabor filter; inputs are convolved images from AER electronic retina
Iosifidis et al. 2012 [26]	activity recognition	different angles of video cameras	not defined	74.1% - 100% for varying activities and camera angles	self-organized maps for human posture learning, fuzzy distances for time invariance from prototypes; multilayer perceptron applied for classification
Goudelis et al. 2013 [20]	activity recognition	video from the KTH and Weizmann datasets	yes	90.22% on KTH and 93.41% on Weizmann	information from actions is obtained by applying the trace transform; temporal information is incorporated by constructing a history trace template
Lu et al. 2013 [21]	sports, human activities	movies from KTH. You Tube action dataset	yes	76.7% on YouTube dataset, 95.3% on the KTH	semantic learning method based on sparse representation, SVM

2. Proposed approaches for human action recognition

2.1. First approach for HAR based on CNN

Although a wide range of techniques for features extraction currently exists, for this research the motion detection and figure/background subtraction approach have been used applying a mixture of Gaussian adaptive components [27]. This procedure generates sequence of binary images from input video, where objects in motion are presented as shadows over white background. To each moving objects in obtained sequence of binary frames the rectangular scaling envelope of size 30x30 pixels is applied to obtain dataset invariant to dimensions of human figure. Once the area with located figure is defined, the image sequence is used as input of convolutional neural network that extracts medium level figure features without previous image processing as it is done in holistic approaches. The described steps of the proposed approach are depicted in Figure 1.

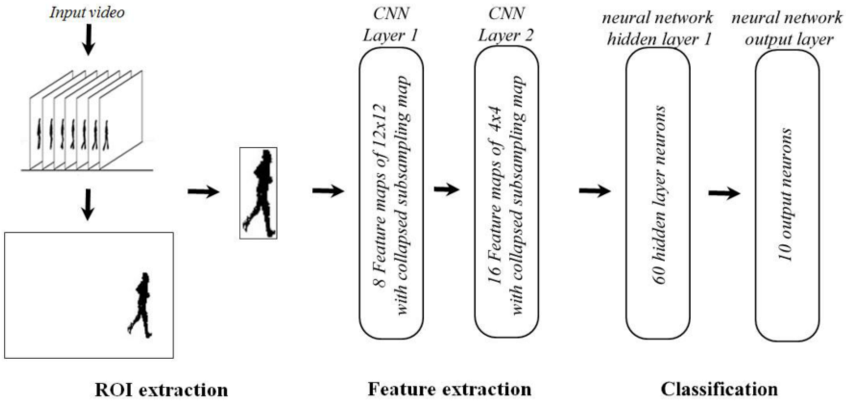
Input video
CNN Layer 1
CNN Layer 2
neural network hidden layer 1
neural network output layer
8 Feature maps of 12x12 with collapsed subsampling map
16 Feature maps of 4x4 with collapsed subsampling map
60 hidden layer neurons
10 output neurons
ROI extraction
Feature extraction
Classification

Figure 1. Description of three basic steps of the first proposed HAR approach

The main characteristics of CNN are feature maps and weight sharing. Feature maps are responsible of input pattern convolution and subsampling. They consist of separate layers, which extract input features followed by subsampling layer that reduces dimensionality and guarantee properties such as invariance to translation and deformation. The weight sharing permits to feature map individual neurons to share a weight set, reducing a number of training parameters and allows parallel implementation of neural network. Subsampling reduces the resolution of feature map and sensitivity of output to shift distortions [24]. Another important feature of neural network is the local receptive field (LRF) that provides an adaptive approach to extract features. The neurons with LRF can extract elemental visual features such as corners, borders, etc. The features maps that integrate convolution and subsampling are constrained to operate equally at different parts of image. In order to provide stability, invariance to scale changes and to reduce output dimensionality the two-layered CNN are used, where the feature and subsampling maps are collapsed together achieving higher efficiency of recognition.

Hidden layer neurons perform complex neuronal pattern approximation receiving as input the feature maps from the last layer of convolution. The last NN output layer classifies the values of hidden neurons to produce an output. Ten output values indicate the classes to which belongs the pattern that was introduced to the CNN. The proposed approaches have been developed for recognizing five actions of humans particularly, in motion (walk, jump, run, side and skip) that may be used in smart environments of video surveillance, ambient-assisted living and personal healthcare support.

2.2. *Second HAR approach based on natural knowledge about human figure*

In order to improve the previous approach, additionally to processing global features, we propose to analyze reduced set of six local features combining in this way the holistic and cluster-based models for representation of human figure. This approach takes as input a sub-sequence, known as snippet, of previously aligned binary silhouettes from frames of input video. Then, for each frame the object of interest OI is segmented using background subtraction and simple method of thresholding [28]. For each frame in the input snippet there is a search for the position and dimension of the smallest rectangle, called Bounding Box (BB), which encapsulates the entire human silhouette of the region of interest (ROI) as it is shown in the Figure 2.

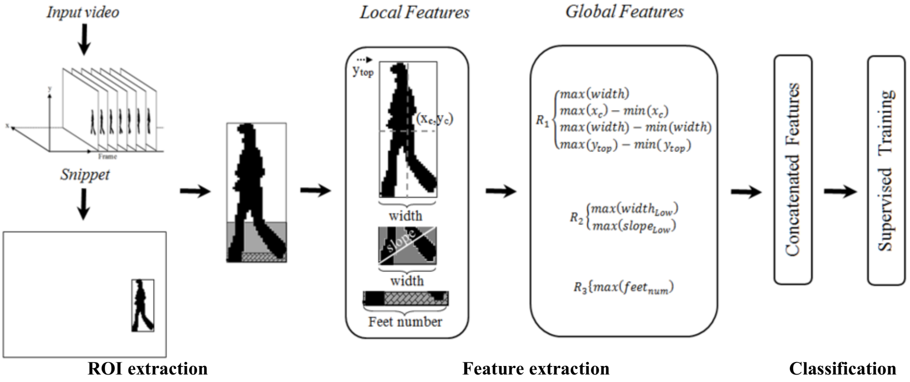

Figure 2. Description of feature extraction and HAR processes in the second proposed approach

Once the BB is defined, the OI is represented by three rectangular regions. The dimensions of a silhouette within each of three regions are the base of extracting two types of features: local and global. Six local features are computed for all frames in the snippet of interest and then their extreme values (maxima and minima) are used to compute seven global features, which are representative for the snippet. Finally, the extracted global features are concatenated into a single vector to get a feature vector per snippet. In the next stage, the feature vector is used to train a hierarchical system of linear classifiers.

Figure 3 shows a subdivision of BB into three regions and measured parameters that define local and global figure features.

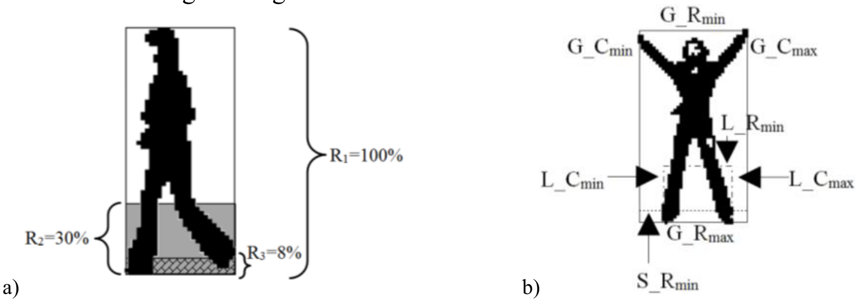

Figure 3. a) representation of a silhouette by BB by three rectangular regions, b) measured parameters

Proposed six local features computed from BB regions are: three features from region 1 (width, centroid abscissa coordinate x_c and upper edge coordinate y_{top}), two from region 2 (width and diagonal slope) and one from region 3 (number of feet planted on ground). Six local features are computed for each frame using parameters and dimensions of the regions R1, R2 and R3, presented in Figure 2 and Figure 3. For example, parameters $G\_C_{min}, G\_C_{max}, G\_R_{min}, G\_R_{max}$ define features of R1 such as width (Eq. 1), centroid abscissa (Eq. 2) and top edge (Eq.3) respectively.

$$Width_{local} = abs(G\_C_{max} - G\_C_{min}) \tag{1}$$

$$x_c = G\_C_{min} + abs\left(\frac{G\_C_{max} - G\_C_{min}}{2}\right) \tag{2}$$

$$y_{top} = G\_R_{min} \tag{3}$$

In the similar way, the parameters $L\_C_{min}, L\_C_{max}, L\_R_{min}, L\_R_{max}$ define features of R2: width (Eq.4) and diagonal slope (Eq.5). Finally, $G\_C_{min}, G\_C_{max}, S\_R_{min}, G\_R_{max}$ define number of feet ($feet_{number}$) in R3.

$$Width_{low} = abs(L\_C_{max} - L\_C_{min}) \tag{4}$$

$$Slope_{low} = abs\left(\frac{G\_R_{max} - L\_R_{min}}{L\_C_{max} - L\_C_{min}}\right) \tag{5}$$

Seven global features $G\_F(1-7)$ are obtained from three regions. Unlike local features, global features are computed per snippet through an analysis performed on all the *n* frames of a snippet. Considering that *i* is the frame index and that one snippet contains *n* frames, four global features from R1 are maximum width (Eq.6), maximum horizontal shift (Eq.7), maximum width minus minimum width (Eq.8) and maximum vertical shift (Eq.9). Two features from R2 are maximum width (Eq.10) and maximum diagonal slope (Eq.11). Last one from R3 is maximum number of feet touching the ground (Eq. 12).

$$G\_F1 = \max_{i\in\{1,..n\}} (Width_{local}[i]) \tag{6}$$

$$G\_F2 = abs\left(\max_{i\in\{1,..n\}} x_c[i] - \min_{i\in\{1,..n\}} (x_c[i] \right) \tag{7}$$

$$G\_F3 = abs\left(\max_{i\in\{1,..n\}} Width_{local}[i] - \min_{i\in\{1,..n\}} (Width_{local}[i] \right) \tag{8}$$

$$G\_F4 = abs\left(\max_{i\in\{1,..n\}} (G\_R_{min}[i]) - \min_{i\in\{1,..n\}} (G\_R_{min}[i]) \right) \tag{9}$$

$$G\_F5 = \max_{i\in\{1,..n\}} (Width_{low}[i]) \tag{10}$$

$$G\_F6 = \max_{i\in\{1,..n\}} (Slope_{low}[i]) \tag{11}$$

$$G\_F7 = \max_{i\in\{1,..n\}} (feat[i]) \tag{12}$$

The computed global features are concatenated to get a single column vector per snippet $v = [G\_F1, G\_F2, ...G\_F7]^T$.

3. Design of classifiers for the proposed HAR approaches

In the first proposal the convolutional neural network shown in Figure 1 has been used for extraction of global feature and NN with two hidden layers has been used for action classification. The proposed configuration is defined on base of the technique proposed by Peemen [30], in which the feature and subsampling maps are collapsed together for achieving higher efficiency of classification. However, it was proposed to use two convolutional layers (subsampling factor of 2), which reduces the image dimensionality without excessive feature loss, and Kernel of size 5 that is correlated to subsampling factor avoiding redundant computing. The number of feature maps was defined empirically testing several combinations that give faster convergence with the lower number of iterations providing the highest recognition precision. Finally, the number of neurons was reduced until network converges. This reduction eliminates the weight redundancy correlation and provides better generalization and recognition.

According to obtained precision rates the configuration 1 has been selected for the first proposed classifier (see Figure 1). Six configurations used for CNN tests are shown in Table 3. The recognition precision is defined from the Stanford standard dataset used for training and for generalization.

Table 3. Configurations of tested CNNs and obtained recognition precision (trained on Stanford database available at *http://vision.stanford.edu/Datasets/40actions.html)*

Configuration	Feature map of CNN layer1	Feature map of CNN layer2	Hidden layer neurons	Recognition precision
1	8	16	60	99.0%
2	10	24	70	96.98%
3	10	24	60	97.18%
4	10	24	100	96.84%
5	10	24	50	96.84%
6	10	24	150	97.08%

The complexity of approach for performing HAR is determined by network parameters particularly, for CNN layers by Eq.13:

$$O((m * n * k_2) + (p * m * q * o) + (p * r)) \tag{13}$$

where m is the number of feature maps and n - size of feature map of CNN layer 1; c is Kernel size; r - number of feature maps and p – size of feature map of CNN layer 2, finally, q - number of connections between CNN layer 1 and 2.

The complexity of the classification process by NN with hidden layers is determined by Eq.14:

$$O((s * t) + (u * s)) \tag{14}$$

where s is the number of neurons in the hidden layer, t - number of input neurons defined by $(r * p^2)$, u - number of output neurons specified by the number of actions to classify. These two layers define the complexity of the proposed approach for feature extraction and classification of human actions.

In the second approach, classification is performed by a hierarchical system of classifiers, which consists of a structure of three levels of classification. At the highest level of the hierarchy there is one perceptron, whose purposes are to enable/disable the outputs of second level support vector machine (SVM) classifier. Additionally, it will be used for future discrimination between two sets of recognized actions: the set of actions carried out at the same place and the set of actions, where a displacement takes place. Thus, this structure is used to separate a complex decision-making process into a collection of simple decisions and to extend next versions of approaches with recognition of actions without human displacement.

In this paper we report recognition of actions only with humans in motion using two SVMs. The tasks of SVM of second level are to enable/disable the outputs of SVM at the third level of classification and to recognize sub-sets of walk/jump activities separating run/side/skip activities for third level SVM. Each SVM classifier is independently trained with a set of samples obtained from Weizmann database [8]. Each multi-class SVM delivers its response through an output bus (vector). The SVM output bus feeds a multiplier, which allows or prevents this bus to contribute in the final action-tagging output vector. Also, a multiplier delivers its response through an output bus, where each single bus line (vector entry) feeds unit-step activation function. All activation function outputs are concatenated into an action-tagging output vector with 10 entries (recognized actions). There is one entry per action tag and during classification one single tag (entry) is activated.

4. Results and discussion

The proposed method was evaluated on the most often cited database in the human action recognition literature, the Weizmann database [8]. This database consists of 90 low resolution video sequences (fifty 180x144 interlaced frames per second) showing nine different persons performing 10 different actions. Once all the possible snippets of aligned binary silhouettes are obtained, each one of them is used to feed the hierarchical system of classifiers during the training and recognition stages.

It is important to mention that the precision rates presented in Table 3 are very high because the same classified patterns were used during the network training. In order to measure more exactly the network performance, we used new patterns that did not participated in classifier training process. Additionally, for five actions of humans in motion (walk, jump, run, side and skip) our proper dataset of 30 videos with resolution of 320x240 (30 fps) of 3 different persons performing these actions have been recorded providing more than 4000 images per activity.

We carried out experiments using two standard protocols. The first protocol is based on the Leave-One-Out Cross-Validation (LOOCV), where video sequences, which belong to one person, are used for testing; and video sequences, which correspond to the remaining eight persons, are used for training. The procedure is repeated for nine possible permutations, and experimental results are reported as the average outcomes from those permutations. In the second protocol two sequences from three persons are used to test the classifier, whereas sequences from the other six persons are used to train the system of classifiers (relation between training and classification is 60% and 40% respectively).

The results are reported as the average of running tests from 84 possible permutations. For proper developed dataset the similar tests have been carried out. Table 4 resumes the obtained results of tests for recognition of five actions of humans in motion with standard and proper video datasets.

Table 4. Results of tests for recognition of action of human in motion

Actions	**Weizmann dataset**				**Proper developed dataset**			
	CNN LOOCV	**CNN 60-40**	**SVM LOOCV**	**SVM 60-40**	**CNN LOOCV**	**CNN 60-40**	**SVM LOOCV**	**SVM 60-40**
walk (%)	99.0	95.5	100	100	98.2	93.65	100	99.2
jump (%)	92.8	88.0	100	98.4	92.2	89.2	100	99.4
run (%)	96.7	89.8	100	99.3	95.6	86.85	99.6	98.9
side (%)	98.1	93.1	100	99.8	97.7	92.4	99.5	100
skip (%)	98.5	92.5	100	100	98.0	91.5	100	100

Comparison of the average precision of the proposed approaches versus those obtained from other well-known methods evaluated on the same Weizmann dataset is shown in Table 5. All methods were validated using the same (LOOCV) protocol.

Table 5. Comparison of HAR average precision of our approaches with other methods on Weizmann dataset

Authors:	**Gorelick [8]**	**Guo [9]**	**Marín [11]**	**Minhas [23]**	**Schindler [28]**	**Wang [22]**	**Proposed (CNN)**	**Proposed (SVM)**
precision (%)	97.54	98.68	98.1	99.9	99.6	100	97.0	100

In terms of the HAR precision the proposed approaches are quite competitive with relevant well-known methods. In the worst case of the 60%-40% test, the correct classification rate lies in range of 91.8 – 99.5% for first and second proposed approaches respectively.

5. Conclusion

After evaluation of the proposed approaches, it is important to conclude that they have enough merit to be used as high precision tool for recognizing actions of humans in motion achieving 97-100% of correct classification rate without previous complex image processing.

Resuming the principal contributions of this work, the proposed approaches are based practically on extraction of only global features of human figure that reduces significantly computational complexity and processing time; only six local features are used per frame (well-known systems use from 50 to 150 times more features in their feature vector [11], [22], [23], [28]); the description vector length is fixed and it does not depend on the resolution of the image; the input image must not be scaled to a fixed size and perfectly segmented silhouettes are not required; the computing of features includes simple operations such as addition, subtraction, multiplication, and division as it has been shown in equations of section 2; developed classifiers provides high precision recognition (up to 99.5% in the worst case of 60%-40% test).

For future works some researches may be suggested such as: to provide extension of number of human actions and interpretation of their activities, to adjust the approaches for efficient recognition of multiple persons in motion applied such to children as to adults, to verify time invariance of approaches to varying motion rate and illumination conditions, to test the proposed approaches using records of persons with excessive clothes that impede tracking of hands and legs in motion, to implement the proposed approaches as embedded system of smart environments particularly, for video surveillance, ambient-assisted living, personal healthcare support, pedestrian detection and others.

Acknowledgment

This research is partially sponsored by European Grant: CASES - Cloud Computing-based Advisory Services towards Energy-efficient Manufacturing Systems.

References

[1] A.A. Chaaraoui, P. Climent Pérez, F. Flórez Revuelta, A review on vision techniques applied to human behaviour analysis for ambient-assisted living, *Expert Systems with Applications* **39**, 12 (2012), 10873-10888.

[2] J.K. Aggarwal, M.S. Ryoo, Human activity analysis: A review, *ACM Computing Surveys* **43**, 3 (2011), 16-23.

[3] O. Duchenne, I. Laptev, J. Sivic, F. Bach, J. Ponce, Automatic annotation of human actions in video, *Proc. of 12th IEEE Int. Conf.on Computer Vision*, Kyoto (2009), 1491-1498.

[4] S. Regazzoni, A. Cavallaro, Y. Wu, J. Konrad, A. Hampapur, Video analytics for surveillance: Theory and practice, *IEEE Signal Processing Magazine* **27**, 5 (2010), 16-17.

[5] E. Alvarez-Valle, O. Starostenko, Recognition of Human Walking/Running Actions Based on Neural Network, *Proc. of Int. Conf. on Electrical Engineering, Computing Science and Automatic Control*, Mexico (2013), 239-244.

[6] T. Kim, Z. Uddin, Silhouette-based human activity recognition using independent component analysis, linear discriminant analysis and Hidden Markov Model, *New Developments in Biomedical Engineering*, INTECH Open Access Publisher, 2010.

[7] S.R. Ke, et al., A Review on video-based human activity recognition, *Computers* **2**, 2 (2013), 88-131.

[8] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, Proc. of *10th IEEE International Conference on Computer Vision*, Beijing (2005), 1-6, available at http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html

[9] K. Guo, P. Ishwar, J. Konrad, Action recognition in video by covariance matching of silhouette tunnels, *Proc. of XXII Brazilian Symp. on Computer Graphics and Image Processing*, Rio de Janiero (2009), 1-6.

[10] D.Y. Chen, Efficient polygonal posture representation and action recognition, *Electronics Letters* **47**, 2 (2011), 101-103.

[11] M.J. Marín-Jiménez, N.P. de la Blanca, M.Á. Mendoza, Human action recognition from simple feature pooling, *Pattern Analysis and Applications,* Springer-Verlag (2012), 1-20.

[12] K.G. Derpanis, M. Sizintsev, K.J. Cannons, R.P. Wildes, Action spotting and recognition based on a spatiotemporal orientation analysis*, IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**, 3 (2013), 527-540.

[13] P. Turaga, R. Chelleppa, V.S. Subrahmanian, Machine recognition of human activities: a survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 2008.

[14] Y. Meng, Y. Jin, J. Yin, M. Conforth, Human activity detection using spiking neural networks regulated by a gene regulatory network, *Proc. of Joint Conference on Neural Networks*, Barcelona (2010), 1-6.

[15] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**, 1 (2013), 221-231.

[16] T. Thang, F. Chen, K. Kotani, Extraction of discriminative patterns from skeleton sequences for accurate action recognition, *IOS Press Fundamenta Informaticae* **130** (2014), 1–15.

[17] A.H. Rad, N.M. Charkari, J.A. Nasiri, H Broojeni, Lying human activity recognition based on shape characteristics, *Proc. of 2nd Int. e-Conf. on Computer and Knowledge Engineering* (2012), 199-203.

[18] W. Zhou, C. Wang, B. Xiao, Z. Zhang, L. Ma, Learning weighted features for human action recognition, *Proc. of 21st Int. Conf. on Pattern Recognition*, Japan (2013), 1160-1163.

[19] J. Luo, W. Wang, H. Qi, Feature extraction and representation for distributed multi-view human action recognition, *IEEE Journal of Emerging and Selected Topics in Circuits and Systems* **99** (2013), 1-11.

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

[21] Z. Lu, Y. Peng, Latent semantic learning with structured sparse representation for human action recognition, *Pattern Recognition* **46** (2013), 1799–1809.

[22] J. Wang, P. Liu, M. F.H. She, A. Kouzani, S. Nahavandi, Supervised learning probabilistic latent semantic analysis for human motion analysis, *Neurocomputing* **100**, 16 (2013), 134-143.

[23] R. Minhas, A.A. Mohammed, Q.M.J. Wu, Incremental learning in human action recognition based on snippets, *IEEE Transactions on Circuits and Systems for Video Technology* **22**, 11 (2012), 1529-1541.

[24] L. Oliveira, U. Nunes, P. Peixoto, On exploration of classifier ensemble synergism in pedestrian detection, *IEEE Transactions on Intelligent Transportation Systems* **11**, 1 (2010), 16-27.

[25] M. Szarvas, U. Sakai, J. Ogata, Real-time pedestrian detection using LIDAR and convolutional neural networks, *Proc. of IEEE Intelligent Vehicles Symposium*, Japan (2006), 213-218.

[26] A. Iosifidis, A. Tefas, I. Pitas, View-Invariant action recognition based on artificial neural networks, *IEEE Transactions on Neural Networks and Learning Systems* **23**, 3 (2012), 412-424.

[27] P. KaewTraKulPong, R. Bowden, An improved adaptive background mixture model for real-time tracking with shadow detection, *Proc. of European Workshop on Advanced Video Based Surveillance Systems*, UK (2001), 1-5.

[28] K. Schindler,L. Van Gool, Action Snippets: How many frames does human action recognition require? *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, AK, USA (2008), 1-8.

[29] I.S.P. Co, The body and its proportions, *Art of Drawing the Human Body*, Spain: Parramon, 2004.

[30] M.C.J Peemen., B. Mesman, H. Corporaal. Efficiency optimization of trainable feature extractors for a consumer platform*, LNCS*, Springer, **6915** (2011), 293-304.