A HUMAN ACTION CLASSIFIER FROM 4-D DATA (3-D+TIME)
Based on an Invariant Body Shape Descriptor and Hidden Markov Models

Massimiliano Pierobon, Marco Marcon, Augusto Sarti and Stefano Tubaro
Image and Sound Processing Group, Dipartimento di Elettronica e Informazione - Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Keywords: Action Recognition. Action Classification. Gesture Recognition. Gesture Classification. Human Motion Analysis. Video Surveillance. Human Machine Interaction. Computer Vision. Multiple View Volumetric Reconstruction. Voxel Based Representation of Human Body. 3D Shape Description. Principal Component Analysis. Pattern Recognition. Context-Dependent Recognition. Hidden Markov Models.

Abstract: Many human action definitions have been provided in the field of human computer interaction studies. These distinctions could be considered merely semantical as human actions are all carried out performing sequences of body postures. In this paper we propose a human action classifier based on volumetric reconstructed sequences (4-D data) acquired from a multi-viewpoint camera system. In order to design the most general action classifier possible, we concentrate our attention in extracting only posture-dependent information from volumetric frames and in performing action distinction only on the basis of the sequence of body postures carried out in the scene. An Invariant Shape Descriptor (ISD) is used in order to properly describe the body shape and its dynamic changes during an action execution. The ISD data is then analyzed in order to extract suitable features able to meaningfully represent a human action independently from body position, orientation, size and proportions. The action classification is performed using a supervised recognizer based on the Hidden Markov Models (HMM) theory. Experimental results, evaluated using an extensive action sequence dataset and applying different training conditions to the HMM-based classifier, confirm the reliability of the proposed approach.

1 INTRODUCTION

Gestures and actions are among the principal ways through which a human being interacts with reality. Many human action definitions have been provided in the field of human computer interaction studies. In (Nespoulous and Perron, 1986), e.g., four different dichotomies were defined in order to provide a classification of different types of human action, namely the act-symbol (with material purpose or communicative), opacity-transparency (having cultural dependent or universal meaning), semiotic-multisemiotic (autonomous or supported by other communication channels) and centrifugal-centripetal (intentional or not). These distinctions could be considered merely semantical if we provide a lower level definition of human action, namely, a sequence of body posture. Thus, a particular set of postures can form a time pattern that conveys information about the action performed.

When humans are involved in action classification the input data that they receive is merely visual: the postures performed by the actor are recognized on the basis of images. A group of action recognition researchers took this consideration as the starting point to develop *vision-based systems* using input from camera devices for their automatic classification purpose. E.g. in (Aggarwal and Cai, 1997) and in (Gavrila, 1999) it is possible to find exhaustive and yet valid surveys of the possible directions that can be followed in vision-based human motion studies and human action recognition.

Despite the ability of the human brain to recognize postures only on the basis of image data, information on body joints configuration is 3-D in nature. The natural way of dealing with posture representa-

Pierobon M., Marcon M., Sarti A. and Tubaro S. (2007).
A HUMAN ACTION CLASSIFIER FROM 4-D DATA (3-D+TIME) - Based on an Invariant Body Shape Descriptor and Hidden Markov Models.
In *Proceedings of the Second International Conference on Signal Processing and Multimedia Applications*, pages 396-403
DOI: 10.5220/0002143303960403

tion is, thus, in the 3-D environment (Mikić et al., 2001). In the work presented in this paper we use a multi-camera input device and a 3-D Visual-Hull reconstruction technique (Laurentini, 1994) in order to provide volumetric information to the system (see Subsect. 2.1). In this way, problems such as viewpoint dependence, motion ambiguities and self-occlusions are inherently solved before the body posture tracking stage.

Frame-by-frame 3-D representations of the scene (4-D data) in terms of voxels (volumetric pixels) have been the input data from which extracting posture-dependent features (Cuzzolin et al., 2004). In the SubSect. 2.3 we introduce a method for performing the tracking of body postures throughout an action sequence, mainly based on the dynamic adaptation of the technique used by Cohen and Li (Cohen and Li, 2003) for static posture estimation. Through experimental sessions, we developed a technique able to extract a posture-dependent signal, independent from actor's position, orientation, size and voxel-set resolution.

The second stage of this research work has been mainly dedicated to the implementation of a reliable pattern recognition algorithm in the context of human action classification (see Subsect. 2.4). The similarity with the speech recognition problem can be quite obvious at this point: it is possible to consider postures as the atoms of actions in the same way as phonemes are often considered the bricks that form words. In other words, the same well-studied approaches used for speech recognition can be followed also in action recognition projects. Therefore, during this work it has been developed a context-based recognition algorithm based on the Hidden Markov Models theory (Rabiner, 1989), a technique already studied and applied in many speech recognition researches.

1.1 Possible Applications

Potential applications of this type of research projects can be easily found in the fields of automatic video surveillance systems (Collins et al., 2000) and (Maybank and Tan, 2000), human-computer gestural interaction researches (Li et al., 1998), (Segen and Kumar, 1999), (Yang and Ahuja, 1999) and (Cui and Weng, 1996), motion based medical diagnosis (Lakany et al., 1999), (Köhle et al., 1997) and (Meyer et al., 1997), robot skill learning. Automatic recognition and classification of suspicious movements (Ivanov et al., 1998) and gaits (Little and Boyd, 1998), (Shutler et al., 2000), (Huang et al., 1999) and (Cunado et al., 1998) in sensitive areas is perhaps one of the most important recent needs demanding for applications at the cutting edge of human action recognition technology. Furthermore the market of video games control devices would benefit from the development of gestures and movements control systems (Freeman et al., 1996) and some industrial products are yet in this direction (Geer, 2004).

2 OVERVIEW OF THE SYSTEM

2.1 4-d Data From Multiple View Acquisition

In order to perform a 3-D reconstruction procedure using a multiple view of the scene, the system has to distinguish the actor silhouette from the rest of the image. A background subtraction technique is used in order to provide this kind of segmentation. Once the object silhouette is extracted for each view, the so called **Visual-Hull** *volumetric reconstruction* of the scene shot by cameras is computed frame-by-frame before any tracking procedure. In this method, 3-D reconstruction is performed using the *volume intersection* approach, which recovers the volumetric description of the object from multiple silhouettes by back projecting from each viewpoint the corresponding silhouette for perspective projections (Laurentini, 1994) (Fig. 1). The intersection volume is then sampled regularly across the three dimensions in order to generate a volume made of binary voxels (ON/OFF). Body posture tracking is then computed directly on volumetric action sequence frames (Fig. 2).

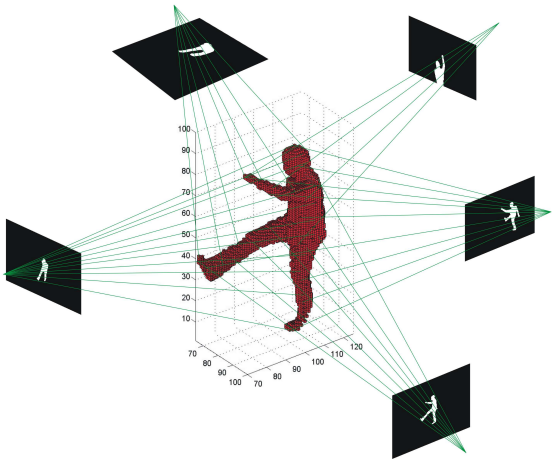

Figure 1: Volumetric intersection. Example of voxel-set creation by 3-D intersection of visual hulls projected from segmented edges.

2.2 Exploiting 4-d Data Information

The 4-D data used for our human action recognition purpose (see Fig. 2) contain multiple information, both related to space and time. Taking into account only the spatial data (instantaneous body volumetric reconstruction), it is possible to distinguish various action-dependent and action-independent information. The *body posture* (body joints configuration), *body position* and *orientation* (with respect to the reference frame of the acquisition system) belong to the first category, whereas the particular body size and proportions of the actor and the volumetric frame resolution belong to the latter. On the other hand, the time data related to the performed action can be divided into *posture sequence information*, *execution time warping*, *action iterations* (if the same action is repeated several times during an action sequence acquisition) and *action concatenation* (if different actions are executed consecutively during a sequence). During the development of an action recognizer that could be as general as possible, we considered the spatial information related to **postures** and the time information related to **posture sequence** as the lowest level data on the basis of which it is possible to classify a human action. All the other types of action-dependent data contain higher semantical information that can be exploited depending on the specific recognition task. E.g. the body orientation can be used if the actor is pointing at a particular direction, or the body position can be important if there is an interaction with the environment. Time information related to execution time warping can be related to different ways of performing a sequence of postures in different action instances and, therefore, it could be potentially used for gait analysis. Eventually, the recognition of the number of action iterations and the analysis of action concatenation can be considered as a natural extension of the system proposed in this article, in which we assume to have input sequences each one containing the execution of only one type of action, possibly repeated several times.

2.3 Posture-Dependent Features

2.3.1 Invariant Shape Descriptor

The core of our body posture tracking procedure is based on the method proposed by Cohen and Li in (Cohen and Li, 2003). They used the Shape Descriptor to compute features suitable for static posture recognition. Our purpose was slightly different because we needed features to perform classification of human actions. Thus, *our shape description had to represent meaningfully not only body postures, but also their frame-by-frame dynamic changes*.

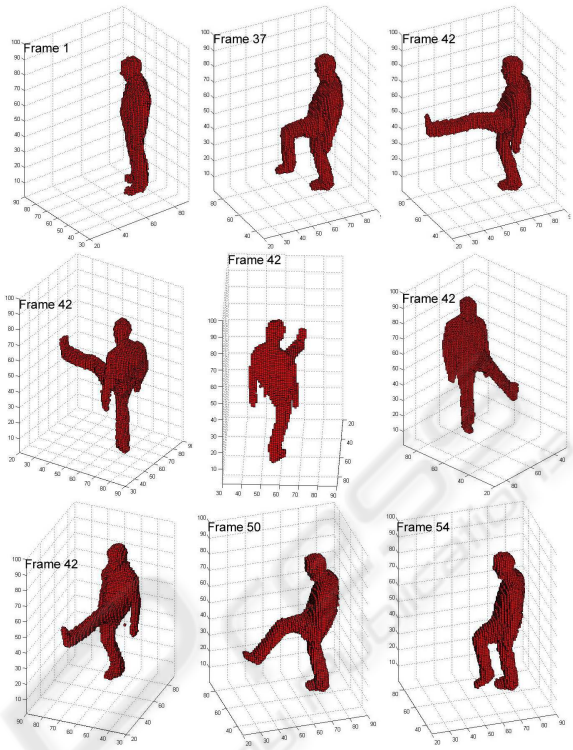

Figure 2: 4-D data. Example of 4-D data regarding a “kick” action. The frame 42 is viewed from five different perspectives.

The procedure starts from the first frame of a sequence (3-D frame from 4-D data) containing the human body volume (e.g. Frame 1 in Fig. 2). The algorithm needs a definition of a **reference shape** consisting of a *vertically oriented cylinder*. It is adapted to the *actor’s height* and its *axis passes through the body 3-D centroid* (Fig. 3 (b)).

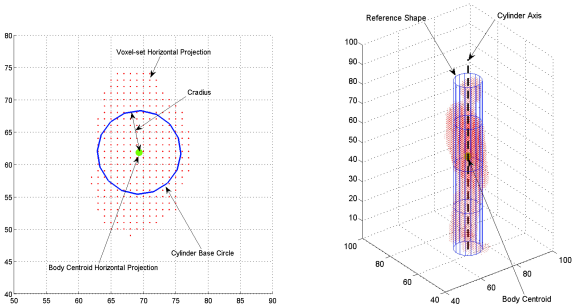

Figure 3: Example of Invariant Shape Descriptor reference shape. In (a) the body horizontal projection silhouette is used to adapt the base circle. In (b) it is shown the reference cylinder surface. Each voxel is represented only by its central point.

The use of the cylinder allows the discrimination between different orientations of the object (body) with respect to the horizontal plane. The base of the adopted cylinder is the *major circle inscribed inside the projection of the body ON-voxels on the horizontal plane* (Fig. 3 (a)). The main advantages of this choice will be explained later.

Once the reference shape surface is gauged on the current voxel-set, it is possible to apply the 3-D Shape Descriptor algorithm: the reference cylinder surface is sampled into a number S of **control points** (p_s, $s \in \{1,\ldots,S\}$). S is a user-defined parameter chosen according to computational cost and representation accuracy criteria.

For each control point p_s:

- Define a spherical coordinates system (ρ,θ,φ) with origin fixed in the p_s location where: $0 \leq \rho \leq \rho_{max}$, $0 \leq \theta \leq \pi$ rad and $0 \leq \varphi \leq 2\pi$. $\theta = 0$ corresponds to the vertical direction, $\varphi = 0$ is the direction of the segment orthogonal to the cylinder axis passing through p_s and ρ_{max} is a value higher than the maximum distance of voxels from the control points.
- Sample uniformly the polar coordinates into parts, respectively S_ρ, S_θ and S_φ. This way we obtain a set of coordinate values $\{(\rho_i,\theta_j,\varphi_k)\}$.
- Assign to p_s a 3-D spherical histogram f_s initially represented by a zero-valued matrix with $S_\rho \times S_\theta \times S_\varphi$ dimensions.
- For each elementary volume in spherical coordinates, defined by a particular $(\rho_i,\theta_j,\varphi_k)$, count how many ON-voxels are contained and store this number in the corresponding histogram location $f_s(i,j,k)$.

The 3-D *Shape Descriptor* $F(i,j,k)$ is a spherical histogram obtained summing up the corresponding values taken from all the histograms of the control points and normalizing these quantities to the maximum value obtained:

$$F(i,j,k) = \sum_{s=1}^{S} \frac{f_s(i,j,k)}{\max_{\bar{i},\bar{j},\bar{k}} \left(\sum_{l=1}^{S} f_l\left(\bar{i},\bar{j},\bar{k}\right)\right)} \qquad (1)$$

The Shape Descriptor $F(i,j,k)$ is invariant (Invariant Shape Descriptor) with respect to **body position** in the voxel-set cartesian frame of reference. The reference cylinder, in fact, *follows the body centroid movements*. Furthermore, the *use of control points lying on the cylindrical surface* allows invariance with respect to **body orientation**. The particular procedure we used to *adapt the reference cylinder to the human body* aims to make the system invariant with respect to the **body size and proportions** of the actor who is performing the posture. The *final normalization of the Shape Descriptor values* removes the proportional relation to how many voxels the body volume is made up (volumetric frame resolution) and possible effects due to **different sizes of volumes** in spherical coordinates, derived from the use of different reference cylinders.

After having computed the cylindrical surface, the cylinder follows the motion of the body centroid but its size remains unchanged for the rest of the sequence. This way we obtain an harmonious variation of features throughout the motion. In our experiments, as suggested in (Cohen and Li, 2003), we sampled the spherical coordinates ten times each, obtaining a spherical histogram that contains 1000 bins (see Fig. 4).

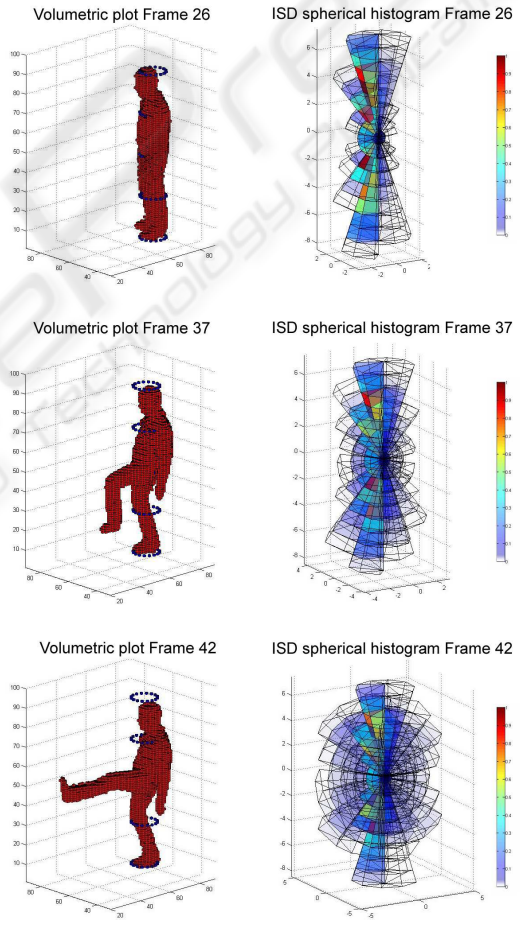

Figure 4: Invariant Shape Descriptor (ISD). Example of ISD spherical histograms computed for 3 frames of a "kick" action sequence. Control point locations (in blue) are shown in the volumetric plots (left).

2.3.2 Feature Selection and Dimensionality Reduction

Following the described method, an Invariant Shape Descriptor $F(i,j,k)$ is computed for each frame of an action sequence. In order to reliably reduce the dimensionality associated with the data contained in the ISD spherical histogram (1000 bins), we applied the Principal Component Analysis in the ISD data domain.

First, $F(i,j,k)$ values ($S_\rho \times S_\theta \times S_\varphi$ values) are collected in vectors (one for each volumetric frame), that are again collected into matrices (one for each action sequence) having a dimensionality of $1000 \times$ *number of frames*.

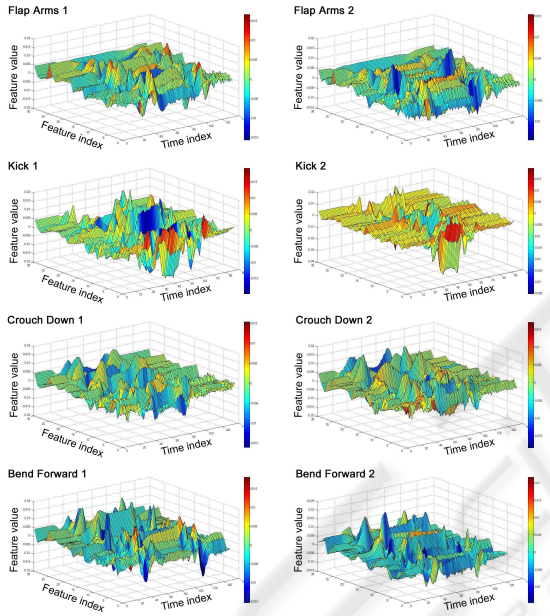

Figure 5: Action sequence feature matrices. Eight examples of action sequence feature matrices. It is possible to notice the similarities between two instances of the same actions and, on the other hand, the dissimilarities existing among instances of different actions.

Eigenvalue analysis is performed directly on the covariance of the matrices computed from the *action sequence data-set used for system training* (the ones used to train the HMM-based classifier block). Eventually, the Karhunen-Loève transform is applied projecting every computed ISD data matrix onto the first 30 principal directions (corresponding to the first 30 eigenvectors). Therefore, through a $30 \times$ *number of frames* feature matrix we represent an action sequence (eight examples are provided in Fig. 5, where each action begins and ends with the same standing up position, with arms hanging on the hips).

2.4 Posture Sequence Classification

In order to design a human action classifier able to exploit the information contained in the extracted features, it is necessary to:

- to provide a system able to classify a $30 \times$ *number of frames* feature matrix into an action category out of a predefined set;
- to provide an action classifier that aims to be insensitive to slight differences in gesture execution time warping, to the number of action repetitions and to the initial posture of the sequence;

In the recognition engine design for this research project we applied one of the most popular context-dependent recognizers to the problem of human actions classification, namely, the Hidden Markov Model classifier. Hidden Markov Models have been widely used for speech recognition applications and their practical implementations for action recognition purposes are still limited. Therefore, our references are mainly based on speech recognition applications: one of the most important for this work has been the article published by L. Rabiner (Rabiner, 1989).

The main idea behind the HMM-based classifier design is to compute a model out of an action training set (a set of sequences containing instances belonging to the same action class) and store the model parameter in a database. Once a model is computed for any available action class, these models are used as a bank of Maximum Likelihood parallel receivers able to classify new feature matrices (classification of new sequences).

2.4.1 The Application of Hidden Markov Models

Starting from HMM theory (Rabiner, 1989), we defined HMM parameters suitable for modeling a given action sequence (see Fig 6):

- The **states of an action model** N are associated semantically to the **principal body postures** that form an action. In this first implementation of the system, we preferred to maintain a fixed number of states. After some experiments, we found that 5 is a suitable state number given the action class considered so far.
- **The type of the observation per state** S_j is represented by a shape descriptor feature vector that ranges continuously in a 30-dimensional space.

$$O_t = \mathbf{x}_t = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{30} \end{bmatrix}_t \qquad 30 = \text{features number} \tag{2}$$

- **The state transition probability matrix** $A = \{a_{i,j}\}$, that is the probability of having a particular posture in the next time instant given the previous one:

$$a_{i,j} = P(q_t = S_j | q_{t-1} = S_i) \qquad 1 \leq i, j \leq N \quad (3)$$

- **The observation probability density function of having the feature vector $\mathbf{x}_t$ in state S_j**:

$$p(\mathbf{x}_t | S_j) \qquad 1 \leq j \leq N \quad (4)$$

is parameterized through a **Gaussian Mixture Model** (GMM):

$$p(\mathbf{x}_t | S_j) = \sum_{m=1}^{M} c_{jm} \, \mathrm{N}(\mathbf{x}_t, \mu_{jm}, \sigma_{jm}) \quad (5)$$

where the number of gaussian mixtures is chosen empirically according to the multi-modality of the $p(\mathbf{x}_t | S_j)$ probability density function of the training data-set.

- **The initial state probability vector** $\Pi = \{\pi_i\}$. We decided to keep the values π_i fixed during the Baum-Welch procedure, an iterative procedure based on the Expectation Maximization mechanism, suitable for finding the HMM parameter values such that the likelihood of the training sequences having the HMM model assigned to that action class is locally maximized (Rabiner, 1989). Moreover, we considered **equal initial probability for all the states**:

$$\pi_i = P(q_1 = S_i) = \frac{1}{N} \qquad 1 \leq i \leq N$$

In fact, *in this work we consider a gesture as a sequence of postures, but independently from the point of the sequence the action performed by the actor in the scene begins*.

3 EXPERIMENTAL RESULTS

In order to perform a full evaluation of the classification system performance, we collected up to 500 4-D action sequences and for each one we extracted the corresponding $30 \times$ *number of frames* feature matrix by means of ISD spherical histogram computation and Karhunen-Loève transformation for dimensionality reduction. The entire data-set included the combination of:

- **10** different action classes (see Tab. 1).
- **5** different actors
- **10** different instances performed by each actor for any action class

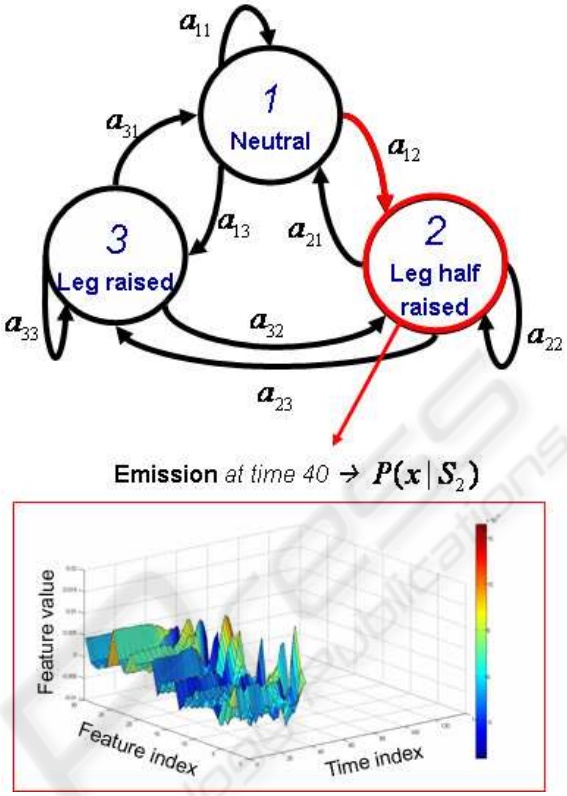

Figure 6: HMM graphical representation. This three-state ($N = 3$) Hidden Markov Model is acting as a stochastic source of an action sequence feature matrix. The tree state model is having a transition from state 1 to state 2 at time 40, therefore the observation vector at time 40 is emitted according to the pdf $p(\mathbf{x}|S_2)$.

Table 1: Different action classes used for system evaluation.

1	flap arms	6	push with hands
2	kneel down	7	crouch down
3	kick	8	bend forward
4	raise arm	9	push with elbow
5	hide face	10	two-step walk

The entire data-set (500 action sequences) has been divided into two *disjoint subsets*: the **train data-set**, that is used to train the Hidden Markov Models for each action class and the **test data-set**, that is used to test the system recognition ability and it is the complementary of the train data-set with respect to the entire data-set. Two subsequent phases have to be performed in order to evaluate the action classifier:

Training phase. the sequences included in the train data-set are are divided into their corresponding action class. For each action class a Hidden Markov Model procedure is performed following the Baum-Welch Expectation-Maximization

algorithm (Rabiner, 1989) and suitable model parameters are learned automatically from the given training sequences. Once this phase is completed, the system stores a set of model parameters for each learned action class.

Test phase. each test sequence is assigned to an action class on the basis of the maximum likelihood with respect to the models. In other words, each model computes the probability (likelihood) of having generated the sequence under test acting as a source (Fig 6). Then the test sequence is assigned to the action class represented by the model showing the highest likelihood.

System evaluation is then performed on the basis of **correct recognition rate** both using the train data-set and the test data-set during the training phase. When the **train data-set** is used for testing, it is possible to evaluate the Hidden Markov Models capability of adapting to the sequences used for training and acting as sources for those feature matrices. On the other hand, when **test data-set** is used, we test the generalization capability of the Hidden Markov Models of representing meaningfully new feature matrices belonging to the learned action classes.

Moreover, the correct recognition rate of the classifier is computed taking into account different experimental conditions:

- Using different percentages of train sequences (out of the 50 available for each action class)
- Using different number of actors for training
- Using Monte-Carlo method in order to select training sequences given an action class, an actor and a training sequence percentage

Experimental results are reported in Fig. 7, where different train data-set percentages (with respect to the entire data-set) include always sequences performed by all possible actors, whereas in Fig. 8 only four actors (out of five) are used for training and the correct recognition percentage is computed as the mean of the values resulting from the test using the sequences of the actor excluded from training (over all possible exclusions of one actor from the train data-set).

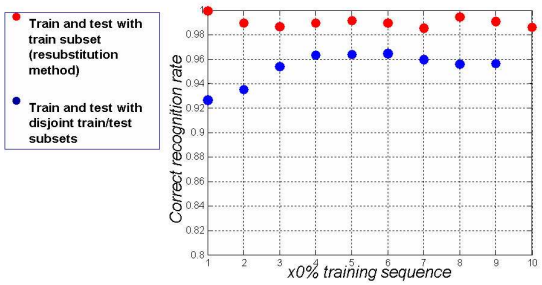

Figure 7: Correct recognition rate using all five actors for model training. The correct recognition rate is tested against different train subset percentages with respect to the entire data-set including 500 action sequences.

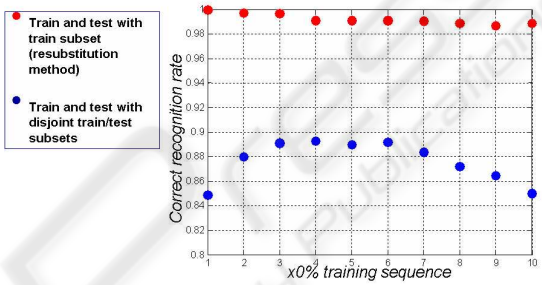

Figure 8: Correct recognition rate using only four actors (out of five) for model training. The correct recognition rate is tested against different train subset percentages with respect to the entire data-set including 500 action sequences.

4 CONCLUSION

In this paper we proposed a human action classifier based on volumetric 3-D data. Through the application of a 3-D reconstruction technique viewpoint dependence, motion ambiguities and self occlusions are inherently solved before posture tracking by a simple computational process, that follows pre-determined steps. The performance shown by the experiments have highlighted the abilities of the used Shape Descriptor not only to represent postures, as shown in (Cohen and Li, 2003), but also to be tuned up in a dynamic context (*Invariant Shape Descriptor*), providing a simple but effective method to track posture movements and slight changes in body shape during an action. The simulations that have been carried out using an Hidden Markov Model-based classifier demonstrate the ability of the Invariant Shape Descriptor histogram data to be properly used for representing action sequence features through the application of a dimensionality reduction technique (Principal Component Analysis). Possible future directions of this project could include a complete evaluation of the system in comparison to other proposed solutions in this research field.

ACKNOWLEDGEMENTS

We wish to tank all the people that actively contributed to this project, especially those who lent themselves to the frustrating job of doing useless ac-

tions in front of eight cameras. Special tanks to Francesco Finetto who patiently carried out all the 500 acquisitions at the I.S.P.G. lab. Thanks to all the I.S.P.G. staff who made possible this research project developing and enhancing the acquisition system and the volumetric reconstruction software.

REFERENCES

Aggarwal, J. K. and Cai, Q. (1997). Human motion analysis: A review. In *IEEE Proceedings of Nonrigid and Articulated Motion Workshop*.

Cohen, I. and Li, H. (2003). Inference of human postures by classification of 3d human body shape. In *IEEE Proceedings of International Workshop on Analysis and Modeling of Faces and Gestures*.

Collins, R., Lipton, A., and Kanade, T. (2000). Introduction to the special section on video surveillance. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cui, Y. and Weng, J. (1996). Hand segmentation using learning-based prediction and verification for hand sign recognition. In *Proceedings of IEEE CS Conference on Computer Vision and Pattern Recognition*.

Cunado, D., Nixon, M., and Carter, J. (1998). Automatic gait recognition via model-based evidence gathering. In *Proceedings of Workshop on Automatic Identification Advanced Technologies*.

Cuzzolin, F., Sarti, A., and Tubaro, S. (2004). Invariant action classification with volumetric data. In *IEEE Proceedings of Workshop on Multimedia Signal Processing*.

Freeman, W., Tanaka, K., Ohta, J., and Kyuma, K. (1996). Computer vision for computer games. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*.

Gavrila, D. (1999). The visual analysis of human movement: A survey. In *Computer Vision and Image Understanding, vol.73, no.1*. Academic Press.

Geer, D. (2004). Will gesture technology point the way? In *Computer*.

Huang, P., Harris, C., and Nixon, M. (1999). Human gait recognition in canonical space using temporal templates. In *Proceedings of IEEE Vision Image Signal Processing*.

Ivanov, Y., Stauffer, C., Bobick, A., and Grimson, W. E. L. (1998). Video surveillance of interactions. In *IEEE Proceedings of the CVPR'99 Workshop on Visual Surveillance*.

Köhle, M., Merkl, D., and Kastner, J. (1997). Clinical gait analysis by neural networks: issues and experiences. In *Proceedings of IEEE Symposium on Computer-Based Medical Systems*.

Lakany, H., Haycs, G., Hazlewood, M., and Hillman, S. (1999). Human walking: tracking and analysis. In *Proceedings of IEE Colloquium on Motion Analysis and Tracking*.

Laurentini, A. (1994). The visual hull concept for silhouette-based image understanding. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, Y., Ma, S., and Lu, H. (1998). Human posture recognition using multi-scale morphological method and kalman motion estimation. In *Proceedings of IEEE International Conference on Pattern Recognition*.

Little, J. and Boyd, J. (1998). Recognizing people by their gait: the shape of motion. In *Journal of Computer Vision Research*.

Maybank, S. and Tan, T. (2000). Introduction to special section on visual surveillance. In *International Journal of Computer Vision*.

Meyer, D., Denzler, J., and Niemann, H. (1997). Model based extraction of articulated objects in image sequences for gait analysis. In *Proceedings of IEEE International Conference on Image Processing*.

Mikić, I., Trivedi, M., Hunter, E., and Cosman, P. (2001). Articulated body posture estimation from multi-camera voxel data. In *IEEE Proceedings of the Conference on Computer Vision and Pattern Recognition*.

Nespoulous, J.-L. and Perron, P. (1986). *THE BIOLOGICAL FOUNDATIONS OF GESTURES: Motor and Semiotic Aspects*. Lawrence Erlbaum Associates, Hillsdale, New Jersey London.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*.

Segen, J. and Kumar, S. (1999). Shadow gestures: 3d hand pose estimation using a single camera. In *Proceedings of IEEE CS Conference on Computer Vision and Pattern Recognition*.

Shutler, J., Nixon, M., and Harris, C. (2000). Statistical gait recognition via velocity moments. In *Proceedings of IEEE Colloquium on Visual Biometrics*.

Yang, M.-H. and Ahuja, N. (1999). Recognizing hand gesture using motion trajectories. In *Proceedings of IEEE CS Conference on Computer Vision and Pattern Recognition*.