# Independent Viewpoint Silhouette-based Human Action Modelling and Recognition

Carlos Orrite, Francisco Martínez, Elías Herrero, Hossein Ragheb, Sergio Velastin

# Independent Viewpoint Silhouette-based Human Action Modelling and Recognition

Carlos Orrite[1], Francisco Martínez[1], Elías Herrero, Hossein Ragheb[2], and Sergio Velastin[2]

[1] Aragon Institute for Engineering Research, University of Zaragoza, SPAIN
[2] Digital Imaging Research Centre, Kingston University, UK

**Abstract.** This paper addresses the problem of silhouette-based human action modelling and recognition independently of the camera point of view. Action recognition is carried out by comparing a 2D motion template, built from observations, with learned models of the same type captured from a wide range of viewpoints. All these 2D motion templates, are projected into a new subspace by means of the Kohonen Self Organizing feature Map (SOM). A specific SOM is trained for every action, grouping viewpoint (spatial) and movement (temporal) in a principal manifold. This approach enables the interpolation of data "between different viewpoints" and, at the same time, to establish motion correspondences between viewpoints without considering a mapping to a complex 3D model. Every new 2D motion template gives a distance to the map, related to the probability that motion feature belongs to that particular action. Action recognition is accomplished by a Maximum Likelihood (ML) classifier over all specific-action SOMs. We demonstrate this approach on two challenging video sets: one based on real actors making 11 complex actions and another one based on virtual actors performing 20 different actions.

## 1 Introduction

With the ubiquitous presence of video data and the increasing importance in a wide spectrum of real-world applications such as visual surveillance, human-computer interfaces, gaming and gesture-based control, and event detection for smart environment, it is becoming increasingly demanding to automatically analyze and understand human motions, from large amounts of video data. Machine learning for vision-based motion analysis is the research field that tries to join aspects from motion analysis such as detection, tracking and object identification with statistical machine learning techniques. This paper addresses the problem of silhouette-based human action modelling and recognition independently of the camera point of view. There are hundreds of papers dedicated to human pose estimation and motion capture, but there are basically two main schools of thought: model based top-down approaches and model-free bottom-up strategies. Model-based approaches presuppose the use of an explicit model of a person and basically match a projection of the human body with the image observation.

Bottom-up methods do not use such an explicit representation, but directly infer human pose/action, from the image features previously extracted. Basically, an example-based method or a learning-based approach is followed from a database of exemplars.

This paper focuses on bottom-up methods where action recognition is achieved by comparing a 2D motion template, built from observations, with learned models of the same type captured from a wide range of viewpoints. 2D templates have the great advantage over 3D models of being directly observable in the image. In addition, humans are able to recognize actions from a single viewpoint. Johansson [1], showed that the trajectories of the 2D joints provide sufficient information to interpret the performed action. However, the main disadvantage of 2D-templates is their dependence on the viewpoint. Recently, many authors have proposed a common approach consisting in discretizing the space considering a series of view-based 2D models [2], [3], [4]. This approach gives some preliminary goods results, but, there are two main problems to be addressed: spatial discontinuities due to viewpoint discretization and temporal discontinuities. The present approach tries to integrate spatial and temporal templates into a common framework combining simultaneously motion state and viewpoint changes. The novelty is that a 3D reconstruction is not required in any stage. Instead, learned 2D motion templates, captured in different viewpoints, are used to produce a temporal-viewpoint map where 2D observations are projected for recognition.

### 1.1 Overview of the work

**Feature extraction**

Human silhouette extraction from videos is easy for current vision techniques, especially in the imaging setting with fixed cameras. So, the method presented here directly relies on moving silhouettes. The input features used in this paper are based on Motion History Images (MHI), as introduced in [5]. MHIs capture motion information in images by encoding, respectively, where motion occurred, and the history of motion occurrences, in the image. 2D motion template exemplars from different viewpoints and velocity actions are generated.

**Human action modelling**

All these 2D motion templates are projected into a new subspace by means of the Kohonen Self Organizing feature Map (SOM) [6]. A specific SOM is trained for every action. In this way, the SOM provides the grouping of viewpoint (spatial) and movement (temporal) in a principal manifold. So, it enables the interpolation of data "between different viewpoints" and, at the same time, establishes motion correspondences between viewpoints without considering a mapping to a complex 3D model.

**Human action recognition**

Given a video sequence corresponding to a specific action, several overlapping windows generate a stream of MHIs. This set of MHIs is continuously feeding to all action-specific SOMs. Every MHI gives a distance to the map, related to the probability that motion feature belongs to that particular action. Once all MHI

templates, corresponding to the same action, have been feed to the specific-action SOM, the individual likelihood outputs can be combined to obtain a consensus decision. This paper presents two strategies: one based on feature combination and the other on neuron action tracking by a HMM. In the latter approach, the state sequence of each SOM is used as the observation vector for the specific HMM. The activity recognition is performed by the well-known Forward-Backward algorithm. Action recognition is accomplished by a Maximum Likelihood (ML) classifier over all specific-action SOMs.

## 2 Previous work

Human action recognition and pose recovery have been studied extensively in recent years, see [7], for a survey. In this document, a brief overview of those methods, more related to our proposal, are given. Since human actions can be characterized as motion of a sequence of human silhouettes over time, silhouette-based methods have received quite a lot of attention lately [5], [8], [9], [10], [11].

Motion Energy Images (MEI) and Motion History Images (MHI) were introduced by Davis and Bobick [5], to capture motion information in images. They encode, respectively, where motion occurred, and the history of motion occurrences, in the image. This approach works effectively under the assumption that the viewpoint is relatively fixed (usually from frontal or lateral view), possibly with small variance. The lack of a view-invariant action representation limits the applications of such 2D based approaches. To overcome this limitation, the authors propose to use multiple cameras. Weinland et al. [10], extend the motion history image concept introducing the motion history volumes as a free-viewpoint representation for human actions in the case of multiple calibrated video cameras. The practicalness of their method is limited since it requires multiple test cameras as well as training cameras. More recently [11], the same authors propose a new framework to model actions using three dimensional occupancy grids, built from multiple viewpoints, in an exemplar-based HMM. In a recent work, Lv and Nevatia [8], exploits a similar idea as proposed by this paper, where each action is modelled as a series of synthetic 2D human poses rendered from a wide range of viewpoints, instead of using an explicitly 3D. The constraints on transition of the synthetic poses are represented by a graph model called Action Net. Given the input, silhouette matching between the input frames and the key poses is performed first using an enhanced Pyramid Match Kernel algorithm. The best matched sequence of actions is then tracked using the Viterbi algorithm. As they proposal is based on synthetic 2D human poses and the type of shape context descriptor used, this approach seems not to be very robust under noisy conditions, which is actually one of the strongest points in this work, as will be shown in the experimental section. The idea of dimensionality reduction to analyze human actions in a low-dimensional subspace rather than the ambient space has been proposed before [9]. Space-time points are projected into a low-dimensional space exploiting locality preserving projections (LPP). Action classification is achieved in a nearest neighbor framework using median Haus-

dorff distance or normalized spatiotemporal correlation for similarity measures. The main difference in relation to the approach introduced in this paper is that the information encoded by the specific-action SOM explicitly models camera viewpoint and temporal action, while the other use parameters to encode such information.

The use of SOM networks for human activity recognition has not received much attention in the past. It has been mainly focused on describe movement in terms of a sequence of flow vectors. Johnson et al. [12], describe the object movement as positions and velocities in the image plane. A statistical model of object trajectories is formed with two competitive learning networks. Hu et al. [13], improve this work by introducing a hierarchical self-organizing approach for learning the patterns of motion trajectories that has smaller scale and faster learning speed. Owens et al. [14], apply SOM feature map to find the flow vector distribution patterns. These patterns are used to determine whether a point on a trajectory is normal or abnormal. These works are based on some flow parameters and the use of the SOM is restricted to learn trajectories, rather than modelling the human motion from different point of views and dynamic figure evolution.

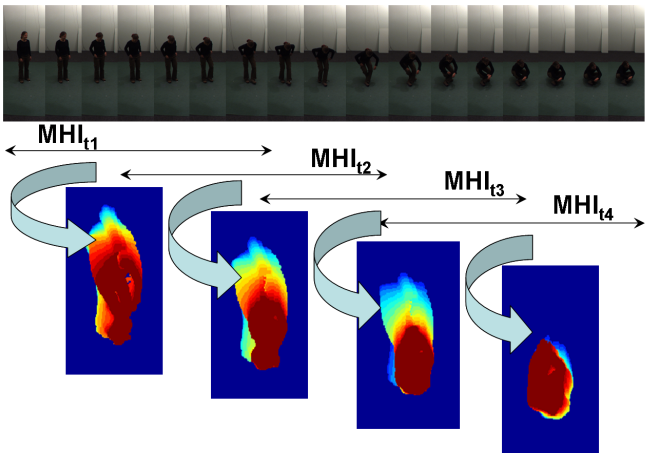


**Fig. 1.** Overlapping MHIs

## 3 Human action modelling based on motion templates

Motion templates are based on the observation that in video sequences a human action generates a space-time shape in the space-time volume. MHIs were introduced to capture motion information in images, encoding how recently motion occurred at a pixel, [5]. These images contain both the spatial information

about the pose of the human figure at any time (location and orientation of the torso and limbs, aspect ratio of different body parts), as well as the dynamic information (global body motion and motion of the limbs relative to the body). To take into account different movement velocities, several window lengths have been considered, so different MHIs have been obtained for the same kind of action, see Fig 1. Each MHI can be seen as a "piece" of an action and the link of successive pieces will constitute the proper action. As the goal is to recognize human body activities from different persons that are free to move in space, with different orientations and sizes, some normalization step has to be carried out. The location and scale dependencies are removed by centering with respect to the center of mass, and scale normalizing, with respect a predefined mask. For the point rotation, or camera point of view, the possibility could be to generate a 3D model of the body as [10]. However, this approach, using many cameras, each highly calibrated, is not feasible for many applications. In this paper we introduce a novel approach where a 3D reconstruction is not required in any stage. Instead, learned 2D motion templates, captured in different viewpoints, are used to produce 2D image information that is compared to the observations. All these 2D motion templates, encompassing temporal movement, are projected into a new subspace by means of the Kohonen Self Organizing feature Map (SOM) [6]. The SOM provides the grouping of viewpoint (spatial) and movement (temporal) in a principal manifold.

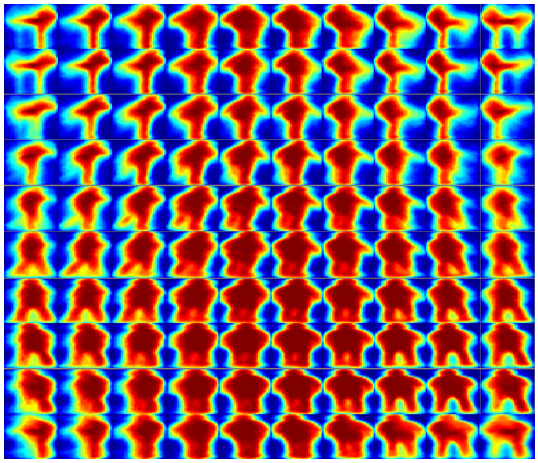

**Fig. 2.** Trained SOM for action kick

The basic idea of SOM is simple: every neuron i of the map is associated with an n dimensional codebook vector. The neurons of the map are connected to adjacent neurons by a neighborhood relation, which defines the topology of the map. The network is trained by finding the codebook vector which is most similar to an input vector. This codebook vector and its neighbors are then

updated so as to render them more similar to the input vector. In this approach, one SOM is trained for each action to recognize. The input data are the MHIs of the available viewpoints and different actors. Fig. 2 shows the representation of the codebook of the trained SOM for action kick. Each image of the figure is the codebook vector of a neuron, and as it can be seen, it corresponds to a particular MHI pattern. So, the SOM integrates spatial and temporal models into a 2D map, combining simultaneously motion and viewpoint changes. It enables the interpolation of data "between different viewpoints" and, at the same time, establishes motion correspondences between viewpoints without considering a mapping to a complex 3D model.

## 4 Action Recognition Using Activity-Specific SOMs

Action recognition is accomplished by a Maximum Likelihood (ML) classifier. Fig. 3 shows the structure of the ML classifier. Given a video sequence corresponding to a specific action, a set of MHIs are continuously feeding to all the action-specific SOMs. Every $MHI_t$, where $t \in (1, T)$, gives a distance to the map, converted to a likelihood, $P(w_k|MHI_t)$, the probability the pattern $MHI_t$ belongs to class-action $w_k$.

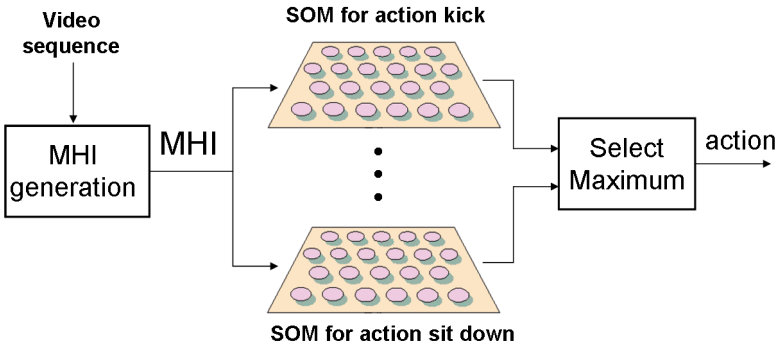


**Fig. 3.** Maximum likelihood classifier for activity specific SOM case

Once all MHI templates, corresponding to the same action, have been feed to the specific-action SOM, the individual likelihood outputs can be combined to obtain a consensus decision. Different features can offer complementary information about the movement, and combining all this information, a better result may be obtained. Different combination possibilities are evaluated in this paper, i.e., sum, product, and maximum rules. The sequence is classified as the action corresponding to the model that yielded the highest probability.

Another approach for action recognition considered in this paper is based on neuron action tracking by a HMM. In this approach an action-specific SOM is

still used, but now the action recognition relies on the dynamics of activation and not in a specific neuron. So, a tracking method along the temporal sequence is needed. The proposal is to use a HMM, as applied before for speech recognition, to concatenate the different SOM neuron activations to form the proper word-action. Following Fig. 4, continuous MHIs are fed to each action specific SOM. The state sequence of each SOM is used as the observation vector for the specific HMM. The activity recognition is done by the well-known Forward-Backward algorithm. Finally, action recognition is accomplished by ML classifier.

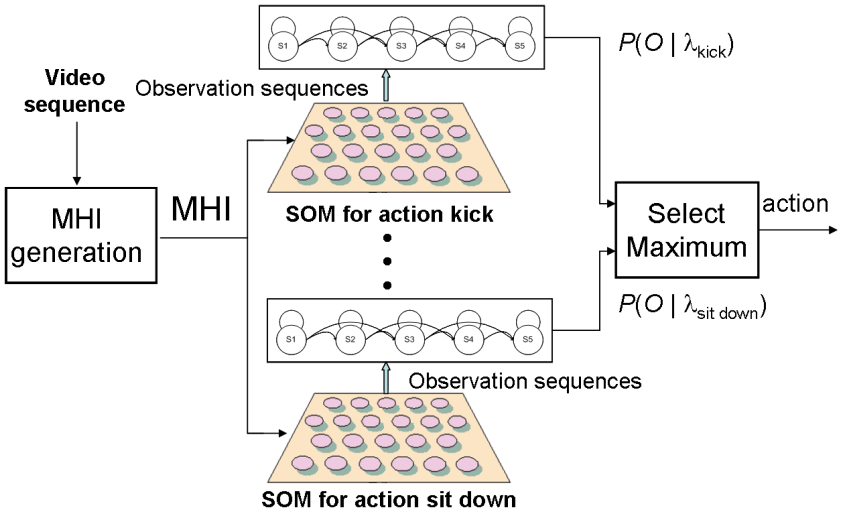


**Fig. 4.** Maximum likelihood classifier for action recognition specific SOM with HMM

## 5 Experiments

The present approach has been tested using two different datasets, one consisting on virtual actor (ViHASi) and other one, called Inria Xmas Motion Acquisition Sequences (IXMAS), with real actors.

### 5.1 IXMAS dataset

The Inria Xmas Motion Acquisition Sequences (IXMAS) [15], contains 11 actions for instance, each performed three times by 12 actors. The acquisition was achieved by five cameras and the actors freely change their orientation for each acquisition in order to demonstrate view-invariance.

A test was carried out by the leave-one-person out (LOO) cross validation. At each iteration, 3 samples corresponding to one single action class performed

by one single actor and captured from 4 camera views were tested, zenithal-viewpoint camera were not used. Testing is done for each of the 4 cameras in isolation. It is worth mentioning that for each iteration of LOO new SOMs are learning for every action.

Every sequence generates 10 MHIs. To combine the output of every action-specific SOM for these MHIs, we follow different combination rules. Table 1 shows the number of of misclassifications (out of 144 samples per action class), and average recognition rate (%) per action applying the sum rule. Similar results are reported by applying the product and the maximum rules.

**Table 1.** Sum rule applied to the 10 outputs given by the action-specific SOM

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 64 | 40 | 66 | 3 | 7 | 14 | 0 | 66 | 43 | 25 | 32 |
| 55.5 | 72.2 | 54.1 | 97.9 | 95.1 | 90.2 | 100.0 | 54.1 | 70.1 | 82.6 | 77.7 |

The confusion matrix for all actions is showed in Table 2. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. As it can be noticed, actions 1, 3 and 8 are similar in some point. The same happens with actions 2, in relation to actions 1 or 3. On the other hand, actions 4 and 11 are quite similar in the initial instant of movement, so there are some misclassifications between both.

**Table 2.** Confusion matrix (144 samples per action)

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 80 | 17 | 3 | 1 | 2 | 9 | 3 | 13 | 14 | 2 | 0 |
| 8 | 104 | 2 | 0 | 2 | 3 | 1 | 9 | 10 | 5 | 0 |
| 14 | 13 | 78 | 0 | 1 | 2 | 3 | 24 | 5 | 1 | 3 |
| 0 | 0 | 0 | 141 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 0 | 0 | 0 | 3 | 137 | 0 | 0 | 0 | 2 | 0 | 2 |
| 0 | 2 | 1 | 0 | 3 | 130 | 5 | 2 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 144 | 0 | 0 | 0 | 0 |
| 16 | 2 | 21 | 0 | 3 | 10 | 3 | 78 | 6 | 1 | 4 |
| 9 | 4 | 3 | 0 | 1 | 5 | 3 | 9 | 101 | 8 | 1 |
| 1 | 2 | 0 | 0 | 6 | 2 | 3 | 0 | 11 | 119 | 0 |
| 0 | 0 | 0 | 18 | 3 | 0 | 10 | 0 | 0 | 1 | 112 |

In relation to other works using the same dataset, this papers approach achieves an overall recognition rate of 77.27% over 12 actors. In [10], Weinland et al. report a higher action classification rate (93.3%), over 10 actors on the

same dataset. They use all five cameras to build visual hulls and classify actions based on a 3D action representation. As mentioned before, the practically of their method is limited since it requires multiple test cameras as well as training cameras. In [8], Lv et al. report an overall action recognition rate of 80.6% over 10 actors, quite similar to our approach. This work is tested exclusively on this dataset and no other reference is given in relation to the system's behavior to noise features, as is presented in a further subsection.

To evaluate the performance of the HMM as neuron action tracker, several left-to-right HMMs, with different numbers of hidden states (from 2 to 10) were tested. The results for the 5-state HMM approach are shown in Table 3. As one can see, these results are quite poor, in relation to those given in Table 1. The explanation for this bad achievement can be found in the different action point of view, given as a result of several trajectories difficult to model by a HMM.

**Table 3.** Average recognition rates for a 5-state HMM

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 32.6 | 25.0 | 29.8 | 59.0 | 52.7 | 56.2 | 65.9 | 39.5 | 40.2 | 40.2 | 47.2 |

### 5.2 ViHASi dataset

Virtual Human Action Silhouette (ViHASi) data [16], provides a large body of synthetic video data generated for the purpose of evaluating different algorithms on human action recognition which are based on silhouettes. This dataset exhibits an interesting property as is the different fitting clothes worn by the actors. The data consist of 20 action classes, 9 actors and up to 40 synchronized perspective camera views split into two sets of 20 cameras. The cameras are located around two circles in a surround configuration, with 18 tilt angle between neighbor cameras, where camera numbers are assigned in the anti-clockwise direction starting with V1 for one 45 slant set and V21 for 27 slant set.

The first test was carried out by the leave-one-out cross validation over all 20 actions (C1 to C20), all actors (A1 to A9). At each iteration, 12 samples corresponding to one single action class performed by one single actor and captured from 12 camera views were tested. Different overlapping window sizes were evaluated. Table 4 shows the number of errors and the average recognition rate per action applying the sum rule for a window size given as a result the division of the action sequence into 5 MHIs. Similar results are reported by applying the product and the maximum rules. The average recognition rate is 98.48%, i.e., 40 misclassifications out of 2640 test samples.

It can be seen in Table 4 that most of the misclassification corresponds to actor A9, which is in fact a virtual child and no other child actor exists in the training data. Size normalization carried out in every silhouette was not enough

**Table 4.** Number of misclassifications (out of 240 per actor) and average recognition (%) against test actors

| A1 | A1b | A2 | A2b | A3 | A4 | A5 | A6 | A7 | A8 | A9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 2 | 1 | 2 | 6 | 0 | 2 | 2 | 23 |
| 99.5 | 99.5 | 100 | 99.1 | 99.5 | 99.1 | 97.5 | 100 | 99.1 | 99.1 | 90.4 |

here, since the child morphology is different to the rest of the actors. Table 5 shows the number of errors and the average recognition rate per action. In this case, action C13 gives the worst score, see [16] for more details about type of action and camera configuration.

**Table 5.** Number of misclassifications (out of 72 samples per action class) and average recognition (%) for each action class

| C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 6 | 1 | 0 | 0 | 0 | 2 | 0 |
| 100 | 100 | 100 | 95.4 | 99.2 | 100 | 100 | 100 | 98.4 | 100 |

| C11 | C12 | C13 | C14 | C15 | C16 | C17 | C18 | C19 | C20 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 4 | 11 | 0 | 6 | 1 | 6 | 1 | 2 | 0 |
| 100 | 96.9 | 91.6 | 100 | 95.4 | 99.2 | 95.4 | 99.2 | 98.4 | 100 |

The final experiment carried out was to test the robustness of the proposal in the presence of noise. For this purpose, silhouettes were corrupted by 'salt and pepper' noise. A parameter ($\alpha$) points out the percentage of the pixels in each silhouette image that are swapped from '1' to '0'. In order to evaluate the system under unlearning conditions, the actors and the camera views used for training were different from those used for testing, for all 20 actions. The actors corresponding to the training data were A1 and A2, while those corresponding to the test data were A3, A4, A5, A6, A7, A8 and A9. The actions performed by A1 and A2 were captured from, 8 camera views (V2, V5, V7, V10, V12, V15, V17 and V20) corresponding to the first camera set, and 8 second camera views (V22, V25, V27, V30, V32, V35, V37 and V40) corresponding to the second camera set. Test data contain 24 novel camera views. Actors (A3, A4, A5, A9) were captured from the first set of 12 camera views (V1, V3, V4, V6, V8, V9, V11, V13, V14, V16, V18 and V19) while those performed by the other 3 of the rest actors (A6, A7, A8) were captured from the second set of 12 camera views (V21, V23, V24, V26, V28, V29, V31, V33, V34, V36, V38, V39). Test data then contain 1680 actions, 240 per test actor. This experiment is more

challenging compared to the previous one since the actors and the camera views in the test data are novel.

Table 6 lists the average recognition rate using different values of $\alpha$. The average recognition rate without noise is 85.83%, i.e., 238 misclassifications out of 1680 test samples. This lower percentage, in relation to the one give by the previous experiment, is mainly due to the low number of examples used for the SOM training stage, (only 2 actors, 8 view points for one slant set and 8 view points for the other slant set, per actor). In spite of this, results are quite impressive, since the recognition rate hardly goes down even for a noise level of about 25%. As a matter of fact, for 5% the results are even better. This positive achievement is due to the kind of motion template features used, as well as the good behavior of the neural approach to deal with corrupted patterns. However, the main problem may be to extract a silhouette from a noise image to get the MHI pattern.

**Table 6.** Average recognition rates (%) adding $\alpha$ percent 'salt and pepper' random noise to the test data

| 0% | 5% | 10% | 25% | 50% |
|---|---|---|---|---|
| 85.83 | 86.0 | 85.7 | 84.8 | 75.7 |

## 6 Conclusion

In this work a new method for human action modelling and recognition from video sequences of human silhouettes was explored. MHIs were used to capture both the spatial information about the pose of the human figure at any time (location and orientation of the torso and limbs, aspect ratio of different body parts), as well as the dynamic information (global body motion and motion of the limbs relative to the body). This kind of motion features have reported good results in situations where human silhouettes were corrupted by noise. The modelling step relies on a Kohonen Self-organized feature map, trained from 2D motion templates recorded in different viewpoints and velocities. SOM approach enables the integration of spatial and temporal templates into a common framework combining simultaneously motion state and viewpoint changes. The novelty is that a 3D reconstruction is not required in any stage. It is true that better results have been reported using a 3D model. However, working with many cameras, each highly calibrated, is not feasible for many applications. Test average recognition rates given in a real dataset, as well as a virtual one, are very promising considering the complexity of the test, using a single camera, and the small amount of data used for training. The main drawback with the approach presented in this paper is that it works on isolated human action sequences or

pre-segmented sequences of the actions from the video. In reality, this segmentation is not available, and therefore there is a need to find an automatic way of segmenting the sequences.

**Acknowledgments** This work is supported by Spanish Grant TIN2006-11044 (MEyC), FEDER. Dr. Orrite received a Grant by DGA(CONAID) and CAI.

## References

1. Johanson, G.: Visual interpretation of biological motion and a model for its analysis. Percept. Psychophys. **73** (1973) 201–211
2. J. Zhang, R. Collings, Y.L.: Bayesian body localization using mixture of non-linear shape models. Proc. IEEE ICCV (2005) 725–732
3. C.S. Lee, A.E.: Simultaneous inference of view and body pose using torus manifolds. Proc. ICPR (2006) 489–494
4. G. Rogez, C. Orrite-Uruuela, J.M.d.R.: A spatio-temporal 2d-models framework for human pose recovery in monocular sequences. Pattern Recognition **41** (2008) 2926–2944
5. Aaron F. Bobick, J.W.D.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence **23** (2001) 257–267
6. Kohonen, T.: Self-organization and associative memory. Springer series in information sciences, Berlin, Springer-Verlag (1988)
7. Weiming Hu, Tieniu Tan, L.W., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. on Systems, man, and Cybernetics-part C: Applications and Reviewers **34** (2004)
8. Lv, F., Nevatia, R.: Single view human action recognition using key pose matching and viterbi path searching. IEEE CVPR (2007) 1–8
9. Wang, L., Suter, D.: Learning and matching of dynamic shape manifolds for human action recognition. IEEE Trans. on Image Processing **16** (2007) 1646–1661
10. Daniel Weinland, Remi Ronfard, E.B.: Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding **104** (2006) 249–257
11. D. Weinland, E.B., Ronfard, R.: Action recognition from arbitrary views using 3d exemplars. Proc. IEEE ICCV (2007) 1–7
12. Johnson, N., Hogg, D.: Learning the distribution of object trajectories for event recognition. Image Vis. Comput. **14** (1996) 609–615
13. W. M. Hu, D.X., Tan, T.N.: A hierarchical self-organizing approach for learning the patterns of motion trajectories. Chin. J.Comput. **26** (2003) 417–426
14. Owens, J., Hunter, A.: Application of the self-organizing map to trajectory classification. Proc. IEEE Int.Workshop Visual Surveillance (2000) 77–83
15. Rhne-Alpes', I.: The inria xmas motion acquisition sequences. https://charibdis.inrialpes.fr (2006)
16. Digital Imaging Research Centre, K.U.L.: Virtual human action silhouette data. http://dipersec.king.ac.uk/VIHASI/ (2008)