

# RBM-based Silhouette Encoding for Human Action Modelling

Manuel J. Marín-Jiménez

Dept. of Computer Science and Numerical Analysis  
University of Córdoba  
Córdoba, Spain  
mjmarin@uco.es

Nicolás Pérez de la Blanca, M. Ángeles Mendoza  
Dept. of Computer Science and Artificial Intelligence  
University of Granada  
Granada, Spain  
nicolas@ugr.es, nines@decsai.ugr.es

**Abstract**—In this paper we evaluate the use of Restricted Boltzmann Machines (RBM) in the context of learning and recognizing human actions. The features used as basis are binary silhouettes of persons. We test the proposed approach on two datasets of human actions where binary silhouettes are available: ViHASi (synthetic data) and Weizmann (real data). In addition, on Weizmann dataset, we combine features based on optical flow with the associated binary silhouettes. The results show that thanks to the use of RBM-based models, very informative and shorter feature vectors can be obtained for the classification tasks, improving the classification performance.

**Keywords**—Restricted Boltzmann Machines; binary silhouettes; human actions.

## I. INTRODUCTION

In the last decade different parametric and non-parametric approaches have been proposed in order to obtain good video sequence classifiers for HAR [1], [2]. Nevertheless, video-sequence classification of human motion is a challenging and open problem, at the root of which is the need of finding invariant characterizations of complex 3D human motions from 2D features.

Different *middle-level* features have been proposed in the recent past years [3], [4], [5]. Nevertheless, the choice of good features for traditional classifiers is nearly an art. Very recently, Hinton [6] and Bengio [7] suggested the interest of using hierarchical non linear models (RBM, DBN) where the inputs are very simple features and the model estimates all intermediate set of features.

Hinton [8], [9] introduced a new algorithm allowing to learn high level semantic features from raw data by using Restricted Boltzmann Machines (RBMs). Variants of this algorithm have shown to be very successful in some image classification problems, where the raw data distributions are represented by the pixel gray level values [10], [11] or binary values [6].

**Contributions of this paper.** This paper introduces the use of models based on Restricted Boltzmann Machines in the problem of human action recognition based on binary silhouettes. A thorough experimental study is carried out on two widely used datasets: ViHASi and Weizmann. In addition, this paper shows that the combination of simple



Figure 1. Goal of this work. We aim to evaluate the capacity of generative models to learn silhouettes of persons performing different actions. Top row: actions included in Weizmann dataset. Bottom row: actions included in ViHASi dataset.

binary silhouettes with aHOF features [12] improves the discrimination in a kNN-based classification framework.

**Outline of the paper.** This paper is organized as follows. In section Sec. II, RBM models are introduced. In section Sec. III, the experiments and results are discussed. Finally, section Sec. IV contains the conclusions of this work.

## II. RESTRICTED BOLTZMANN MACHINES

A Restricted Boltzmann Machine (RBM) is a Boltzmann Machine with a bipartite connectivity graph (Fig. 4). That is, an undirected graphical model where only connections between units in different layers are allowed. A RBM with  $m$  hidden variables  $\mathbf{h}_i$  is a parametric model of the joint distribution between the hidden vector  $\mathbf{h}$  and the vector of observed variables  $\mathbf{x}$ , of the form  $P(\mathbf{x}, \mathbf{h}) = Z^{-1} \cdot e^{-E(\mathbf{x}, \mathbf{h})}$  where

$$E(\mathbf{x}, \mathbf{h}) = -\mathbf{b}^T \mathbf{x} - \mathbf{c}^T \mathbf{h} - \mathbf{h}^T \mathbf{W} \mathbf{x}$$

is a bilinear function in  $\mathbf{x}$  and  $\mathbf{h}$  with  $\mathbf{W}$  a matrix,  $\mathbf{b}$  and  $\mathbf{c}$  vectors, and  $Z = \sum_{\mathbf{h}} e^{-E(\mathbf{x}, \mathbf{h})}$  being the partition function (see [7]). It can be shown that the conditional distributions  $P(\mathbf{x}|\mathbf{h})$  and  $P(\mathbf{h}|\mathbf{x})$  are independent conditional distributions, that is

$$P(\mathbf{h}|\mathbf{x}) = \prod_i P(\mathbf{h}_i|\mathbf{x}), \quad P(\mathbf{x}|\mathbf{h}) = \prod_j P(\mathbf{x}_j|\mathbf{h})$$

Furthermore, for the case of binary variables we get

$$P(\mathbf{h}_i|\mathbf{x}) = \sigma(c_i + \mathbf{W}_i \mathbf{x}), \quad P(\mathbf{x}_j|\mathbf{h}) = \sigma(b_j + \mathbf{W}_j \mathbf{h}) \quad (1)$$

where  $\sigma(x) = (1 + e^{-x})^{-1}$  is the logistic sigmoidal function and  $\mathbf{W}_i$  and  $\mathbf{W}_j$  represent the  $i$ -row and  $j$ -column respectively of the  $\mathbf{W}$ -matrix.

**Learning parameters: Contrastive Divergence.** Learning RBMs maximizing the gradient log-likelihood needs of averaging from the equilibrium distribution  $p(x, h)$  what means a prohibitive cost. The Contrastive Divergence (CD) criteria proposed by Hinton, [8], only needs to get samples from the data distribution  $p_0$ , and the one step Gibbs sampling distribution  $p_1$ , what implies an affordable cost. The equations for parameter updating give values proportional to the difference of averages from these two distributions. That is,

$$\Delta w_{ij} \propto \langle v_i h_j \rangle_{p_0} - \langle v_i h_j \rangle_{p_1}$$

where  $\langle v_i h_j \rangle$  means average (using the subindex distribution) on the number of times that hidden unit  $j$  is on for the visible variable  $i$ . The equations for the bias  $b_i$  and  $c_j$  are similar.

### III. EXPERIMENTS AND RESULTS

The aim of the following experiments is to study if RBM models are able to model simultaneously body poses of different actions. The experiments are carried out on two datasets: ViHASi and Weizmann. The first one, ViHASi, is composed by binary silhouettes synthetically generated. On the other hand, Weizmann dataset includes realistic data.

#### A. Experiments on ViHASi dataset

ViHASi (Virtual Human Action Silhouette) dataset [13] contains 20 actions<sup>1</sup> performed by 11 different virtual actors. The actions are generated from different camera viewpoints.

In this experiment, we are going to use binary images (silhouettes) representing different instants (poses) of the performed actions.

The original resolution of the images is  $640 \times 480$ , but for our experiments the images have been cropped and resized to a common size  $42 \times 42$  pixels. Fig. 1 (bottom row) shows typical examples of the actors and actions that can be found in this database.

Since this database contains actions performed with different points of view of the camera, we have mixed different cameras in the experiments.

The evaluation of classification performance has been carried out under a *leave-one-out* strategy on the actors. Therefore, the reported performance is the average of 11 repetitions.

**Encoding action poses with RBM.** In this experiment, we learn RBM-codes, with different layouts, for the input

<sup>1</sup>Action names: *Collapse, Grenade, HangOnBar, HeroDoorSlam, HeroSmash, JumpFromObject, JumpGetOnBar, JumpOverObject, Kicks, Knockout, KnockoutSpin, Punch, Run, RunPullObject, RunPushObject, RunTurn90Left, RunTurn90Right, StandLookAround, Walk, WalkTurn180.*

Table I  
RBMs ON VIHASI. PERCENTAGE OF CORRECT CLASSIFICATION PER FRAME. CAMERAS: C3+CN3, C6+CN6, C9+CN9.

H	C3+CN3		C6+CN6		C9+CN9	
	1NN	5NN	1NN	5NN	1NN	5NN
100	95.6	95.3	93.9	94.0	95.7	95.8
200	96.9	96.9	95.0	94.8	96.4	96.6
500	97.3	97.2	95.6	95.6	97.0	97.0
1000	<b>97.5</b>	<b>97.6</b>	<b>95.7</b>	<b>95.8</b>	<b>97.2</b>	<b>97.3</b>
Raw	<b>97.6</b>	97.0	<b>95.6</b>	95.5	<b>97.1</b>	97.1

frames, and evaluate the quality of the learnt features by using a classification criterium, i.e. a kNN classifier is used. Table I shows the classification performance per frame (i.e. each frame is classified in an isolated way as belonging to a single action), for cameras C3+CN3, C6+CN6 and C9+CN9, respectively. Note that the camera point of view is single in each of these experiments. The results show that, with codes of dimensionality 500, the classification performance is already similar to the one offered by the raw data (bottom row).

Table II  
RBMs ON VIHASI. PERCENTAGE OF CORRECT CLASSIFICATION PER FRAME. CAMERAS: C6+CN6 PLUS C16+CN16.

H	100	200	500	1000	1764	2000	Raw
1NN	92.9	95.0	95.7	<b>95.9</b>	95.8	95.8	<b>95.5</b>
5NN	93.0	95.1	95.7	<b>96.0</b>	95.8	95.8	95.3

In table II, we are mixing cameras C6+CN6 with C16+CN16, whose points of view are opposite. The results show that with 500 hidden units the classification performance is already higher compared to the one offered by the raw data (right most column).

#### B. Experiments on Weizmann dataset

Weizmann actions dataset [14] consists of 93 videos, where 9 people perform 10 different actions: *walking, running, jumping, jumping in place, galloping sideways, jumping jack, bending, skipping, one-hand waving* and *two-hands waving*.

Silhouettes are obtained by background subtraction [14], therefore, the noise derived from the silhouette estimation procedure makes the experimental setup more realistic.

The original size of the binary masks is around  $110 \times 75$  pixels, and there is a total of 5687 frames. For our experiments, we have resized them to a common size of  $50 \times 32$  pixels.

The evaluation of classification performance has been carried out under a *leave-one-out* strategy on the actors. Therefore, the reported performances are the average of 9 repetitions.

**Encoding binary silhouettes.** In this experiment we use

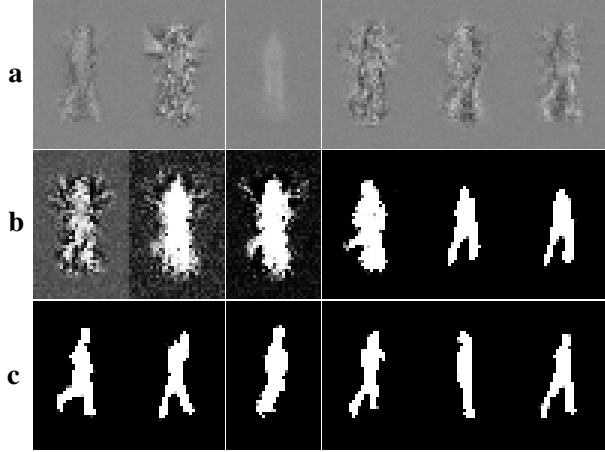


Figure 2. RBM with 800 hidden units, learnt on Weizmann dataset. (a) Example of weights learnt. (b) Evolution of a sample drawn from the model (after 1, 2, 5, 10, 50, 100 Gibbs iter.) (c) Random samples generated by the model (100 iter.).

RBM models with different number of hidden units to learn binary silhouettes of the actions performed in the Weizmann dataset. The codes generated in the hidden layer are used as feature vectors for classification in a kNN framework.

Table III  
WEIZMANN SILHOUETTES ENCODED BY RBM. PERCENTAGE OF CORRECT CLASSIFICATION PER FRAME.

$H$	100	200	400	800	1600	Raw
<b>SMax</b>	81.7	81.8	81.8	82.1	<b>82.9</b>	-
<b>1NN</b>	85.5	85.9	85.9	86.8	86.4	85.0
<b>5NN</b>	86.4	87.1	87.2	<b>87.7</b>	87.6	<b>85.5</b>

Table Tab. III shows the results of the per frame classification by using 1NN and 5NN, along with a SoftMax classifier [6] trained during the fine-tuning stage of the RBM parameters learning. The last column shows the performance achieved by the raw data (pixels of each original binary image are arranged in a single vector by concatenating the columns). The results in the table show that, even with vectors of dimensionality 100, the classification performance of the RBM-codes is higher than the offered by the raw data (1600 dims.).

Since RBM is a generative model, we have run 100 Gibbs sampling steps on one of the models learnt in the experiment summarized on table Tab. III, i.e. 800 hidden units. Figure Fig. 2.a shows 6 columns (reshaped) of the weights matrix  $W$ . In row **b** of the figure, we can see how a sample evolves through the Gibbs sampling steps. In particular, it shows the states of the sample after 1, 2, 5, 10, 50 and 100 sampling steps. Finally, the bottom row (**c**) of the figure includes random samples generated by initializing the hidden units randomly (values in range  $[0, 0.05]$ ). Note how the samples mimic human poses included in the dataset. Possibly, *skip*,

*side*, *jump*, *run*, *pjump* and *walk*.

**Sampling poses from two-classes models.** In this experiment, we train RBM models by using examples of action poses extracted from pairs of action classes. In particular, we train three models with 400 hidden units each.

Figure Fig.3 shows samples generated from models *a*) trained with *wave1* and *wave2* examples; *b*) trained with *jack* and *side* examples; and, *c*) trained with *run* and *walk* examples. They represent the states of the sampled data after 100 Gibbs sampling steps. Hidden units were randomly initialized with values in the range  $[0, 0.05]$ . Note that the visual quality of the generated samples is comparable to the original data.



Figure 3. Sampling RBM models: (a) wave2 and wave1, (b) side and jack, (c) walk and run.

**Combining silhouettes and motion.** In this experiment, each frame is represented by the vector obtained by concatenating a binary silhouette to an *aHOF* descriptor [12]. For each frame, we compute its *aHOF* motion descriptor<sup>2</sup> by using the 20 previous frames. Therefore, each frame is represented by a 2624 (1600+1024) dimensional vector (see Fig. 4).

Table IV  
COMPARISON OF SILHOUETTES PLUS *aHOF*-20 AGAINST JUST *aHOF*-20, ENCODED BY RBM, ON WEIZMANN. PERCENTAGE OF CORRECT CLASSIFICATION PER FRAME.

$H$	<i>Sil+aHOF</i>			<i>aHOF</i>	
	<b>SMax</b>	<b>1NN</b>	<b>5NN</b>	<b>1NN</b>	<b>5NN</b>
50	90.4	89.7	89.8	90.9	91.1
100	83.5	92.2	92.2	91.4	91.2
400	94.1	93.5	<b>94.0</b>	92.3	92.3
800	94.3	93.6	93.6	92.7	<b>92.8</b>
1600	<b>94.7</b>	93.2	93.6	-	-
2624	94.6	93.2	93.5	-	-
Raw	-	<b>94.2</b>	94.1	93.0	<b>93.9</b>

The classification results shown in table Tab. IV indicate that the combination of *Sil+aHOF* features improves the representation. In fact, RBM-50 codes achieves a classification performance similar to the best provided by silhouettes

<sup>2</sup>*aHOF* setup:  $8 \times 4$  cells, 8 orientations and 4 magnitude bins. As in [12]

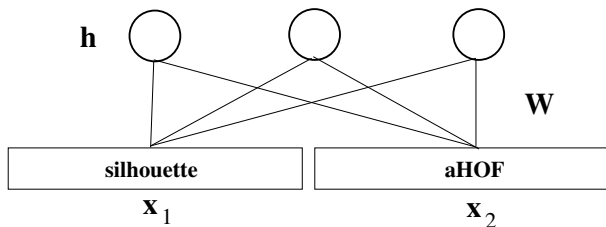


Figure 4. RBM is able to learn combined information. Observed data comes from two sources: binary silhouette and aHOF descriptor.

alone (see table Tab. III). For comparison purposes, table Tab. IV includes two extra columns (right most) with the results obtained by using just the aHOF data.

We could compare the 94.7% performance achieved by our features to Gorelick et al. [14], who report a 97.8% of correct classification of subsequences of 8 frames (in jumps of 4 frames). Note that we are using silhouettes whose resolution is lower (approximately half) than the ones used by them.

The results show that the RBM codes derived from Sil+aHOF vectors, offers better results (94.0% with 5NN, 94.7% with SoftMax) than both the binary silhouettes (87.7%) and aHOF (92.8%). This behaviour also occurs in the original feature space.

If we sample from the model trained with 800 hidden units and combined features (see Tab. IV), we get poses as shown in figure Fig. 5. The visual quality of the samples seems to be lower than the ones trained with a single kind of features (e.g. Fig. 2), but they still represent the essence of the action poses shown during training.

#### IV. CONCLUSIONS

RBM models are able to learn, to generate and to represent, in a compact way, human body poses from simple pixel value representations.

In addition, the combination of static (binary silhouettes) and dynamic data (aHOF features) shows that such combination improves the classification performance offered by those features when used separately.

These properties of the RBM model open new interesting questions such as: *a)* architecture design for combining different basic features; and, *b)* how to use this model as prior distribution in recognition problems.

#### ACKNOWLEDGMENTS

This work has been granted by project Consolider Ingenio MIPRCV (CSD2007-00018) of the Spanish Minister of Science and Innovation.

#### REFERENCES

[1] T. B. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *CVIU*, vol. 104, pp. 90–126, 2006.

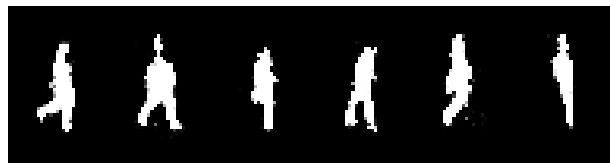


Figure 5. Samples of poses generated by an RBM model ( $h=800$ ) trained with Sil+aHOF data.

[2] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 11, pp. 1473–1488, 2008.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2nd IEEE Workshop VS-PETS*, 2005, pp. 65–72.

[4] E. Shechtman and M. Irani, "Space-time behavior-based correlation or How to tell if two underlying motion fields are similar without computing them?" *IEEE PAMI*, vol. 29, no. 11, pp. 2045–2056, 2007.

[5] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *CVPR*, 2008.

[6] G. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313(5786), pp. 504–507, 2006.

[7] Y. Bengio, "Learning deep architectures for AI," Dept. IRO, Universite de Montreal, Tech. Rep. 1312, 2007.

[8] G. Hinton, "Training product of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14(8), pp. 1711–1800, 2002.

[9] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[10] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large database for recognition," in *Comp. Vision and Patt. Rec.*, 2008.

[11] J. Yang, R. Yan, Y. Liu, and E. P. Xing, "Harmonium models for video classification," *Stat. Anal. Data Min.*, vol. 1, no. 1, pp. 23–37, 2008.

[12] M. Marín-Jiménez, N. P. de la Blanca, M. Mendoza, M. Lucena, and J. Fuertes., "Learning action descriptors for recognition," in *WIAMIS 2009*. London, UK, 2009.

[13] H. Ragheb, S. Velastin, P. Remagnino, and T. Ellis, "ViHASi: Virtual Human Action Silhouette data for the performance evaluation of silhouette-based action recognition methods," in *Workshop on Activity Monitoring by MCSS*, Sept. 2008.

[14] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE PAMI*, vol. 29, no. 12, pp. 2247–2253, December 2007.