HUMAN ACTION RECOGNITION BASED ON SELF ORGANIZING MAP

Wei Huang, Q. M. Jonathan Wu

Department of Electrical and Computer Engineering, University of Windsor,
Windsor, Ontario, N9B3P4, Canada
Email: {huang1m, jwu}@uwindsor.ca

ABSTRACT

This paper proposes a novel neural network approach for human action recognition based on Self Organizing Map (SOM). The SOM acts as a tool to cluster feature data and to reduce data dimensionality. The key poses in action sequences are extracted by the trained SOM. After the mapping of SOM, a human action sequence is represented as a trajectory of map units. For action recognition, a longest common subsequence algorithm is utilized to match action trajectories on the map robustly. The experiments are carried out on a well known human action dataset, viz.: the Weizmann dataset. We obtain promising results which show the potential of this SOM based action recognition method.

Index Terms— Human action recognition, self organizing map, longest common subsequence matching, human silhouette, dynamic programming

1. INTRODUCTION

Automatic recognition of human actions in video is useful for many important applications such as video surveillance and monitoring, content-based video indexing and retrieval, human-computer interaction, etc.

Generally, there are two subproblems associated with the problem of human action recognition. One subproblem is extracting a set of features to model the action sufficiently. In early work, [1] used local space-time features to describe motion patterns and combine such descriptors with SVM classification schemes for the recognition task. 2D Harris corner detector was extended to a 3D Harris detector to detect local regions where the image intensities have high variations in both space and time domains. [2] presented a novel local descriptor for video sequences which is based on histograms of oriented 3D spatio-temporal gradients. [3] extended a 2D shape representation method to 3D space-time volumes by adding time domain to a Poisson equation. Space-time shape features were extracted and used to model an action. [4] combined Motion Energy Images (MEI) and Motion History Images (MHI) together to build a temporal template, which was used to represent human actions. The MEI indicated the location of motion in a video sequence, and the MHI showed the temporal history of motion.

The other subproblem is clustering the feature data and reducing the data into low dimensional space. In the past few years, the "bag of words" model has indicated its validity in action classification. [5] used a "bag of words" model to cluster local features into spatio-temporal codewords. [6] proposed a constellation-of-bags-of-features model to exploit both the geometric power of the constellation model as well as the richness of the "bag of words" model. The constellation model described the mutual geometric relationship among different body parts. A "bag of words" model was attached to each part of the constellation model. [7] introduced a 3D SIFT descriptor and incorporated this new descriptor in a bag of words paradigm in order to obtain improved performance on the task of action recognition. In [8], a set of MHIs, captured in different viewpoints, were input to each action-specific SOM to project the motion templates into a new subspace. A Maximum Likelihood classifier was used over all action-specific SOMs for action recognition. The SOM grouped both viewpoint and movement information of each action in the map to build motion correspondences between different viewpoints.

In this paper, we introduce a novel neural network approach for human action recognition based on SOM. We adopt SOM to cluster feature data and to reduce data dimensionality. Sequences of human silhouettes are selected as basic feature representations and act as inputs to SOM. Key poses in motion sequences are extracted by the trained SOM. A human action sequence is therefore represented as a trajectory of map units. For action recognition, a longest common subsequence algorithm is employed to match action trajectories on the map robustly.

2. PROPOSED ALGORITHM

In the following sections we describe our proposed algorithm in details. Section 2.1 depicts the feature representation method by using sequences of human silhouettes. Section 2.2 details Self-Organizing Map. Section 2.3 presents the action recognition scheme based on Longest Common Subsequence algorithm. Section 3 gives

experimental results. Section 4 concludes the proposed approach.

2.1. Feature representation by sequences of human silhouettes

The purpose of this stage is extracting sequences of human silhouettes to represent human action in raw video. There are several advantages of using silhouettes: 1) silhouettes contain rich shape information; 2) a series of body silhouette images is independent of the speed of movement; 3) binary silhouettes are robust to variations in clothing and lighting condition; and 4) silhouettes are easier to extract compared with tracking body parts.

We do not intend to address the foreground detection issue in this work, so the silhouette masks provided in [3] are directly used in our experiments. To cope with size and position variation, silhouette images are centered and normalized by keeping the aspect ratio of the silhouettes so that the resulting images do not distort the moving shape and have same dimensions. Accordingly, each 2D silhouette image can be converted to a 1D vector with the same dimension in a row-scan manner. The silhouette sequences are then represented as: $X_k = \{x_1, x_2, ..., x_{N_k}\}$, where $x_i (i = 1,2,..., N_k)$ is 1D vectors and N_k is the length of the sequence. Some normalized example silhouettes of different actions performed by the subject "daria" are shown in Figure 1.

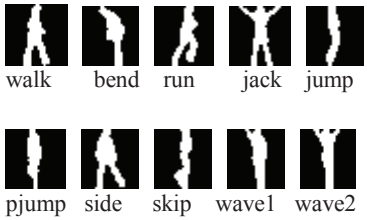

Fig. 1. Normalized example silhouettes of different actions performed by the subject "daria".

It is a difficult task to learn the action characteristics by analyzing data in the high dimensional image space. In this paper, we adopt SOM to cluster feature data and to reduce data dimensionality for recognition of human actions in video. The dimension reducing property of the SOM allows replacing the original high dimensional data points by 2-dimensional points. At the same time the neighborhood preservation property of the SOM allows to cluster the feature data into mapped data points, which are utilized in a subsequent trajectory comparing stage.

2.2. Self-Organizing Map

The Self-Organizing Map (SOM) was introduced by T. Kohonen [9]. The SOM is one of the most popular unsupervised learning neural networks and has been successfully applied as a data mining tool with various applications in pattern recognition and image analysis.

The SOM quantizes the data space formed by the training data and simultaneously performs a topology-preserving projection of the data onto a regular low-dimensional lattice. The lattice can be used efficiently in visualization. The map is generated by establishing a correspondence between the input data and neurons located on the lattice. The correspondence is obtained by a competitive learning algorithm. The weight vectors of the neurons are modified iteratively in the training steps. When a new input arrives every neuron competes to represent it. The best matching unit (BMU) is the neuron that wins the competition. The BMU with its neighbors on the lattice are allowed to learn the input. Neighboring neurons will gradually specialize to represent similar inputs, and the representations will become ordered on the map lattice. The best matching unit *b(n)* is the weight vector $\omega_j(n)$ that is nearest to the input *x(n)* and is obtained by the following metric:

$$b(n) = \arg\min \left\| x(n) - \omega_j(n) \right\| \tag{1}$$

In general, the Euclidean distance is used. The winning unit and its neighbors adapt to represent the input by modifying their weight vectors towards the current input:

$$\Delta\omega_j = \gamma h_{jk} (x(n) - \omega_j) \tag{2}$$

where γ is a learning rate which is a monotonically decreasing function with regard to time. h_{jk} is a Gaussian function of the Euclidean distance between units *j* and *k* on the map:

$$h_{jk} = \exp(-d(j,k)^2 / \sigma^2) \tag{3}$$

In our approach, the input data of SOM are the sequences of human silhouettes. After SOM mapping, the chains of the best matching units (BMUs) for silhouette sequences produce time varying trajectories of activity on the SOM. These trajectories are therefore employed for action classification. Figure 2 illustrates the transformation of a silhouette sequence into a trajectory on the map.

2.3. Action recognition using longest common subsequence scheme

After trajectories on the Self-Organizing Map are obtained, we are ready to implement the recognition task. An action sequence has been mapped to a trajectory on the map. When

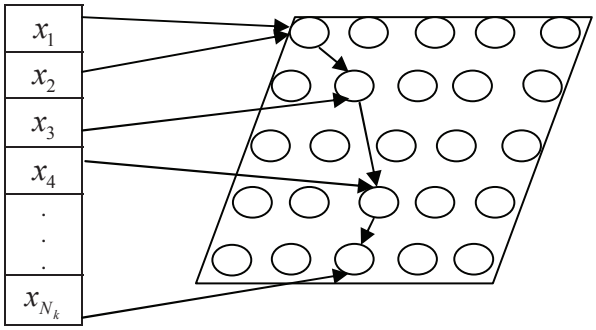

Fig. 2. Transformation of a silhouette sequence into a trajectory on the SOM.

action sequences of same class are input, the trajectories they produce are similar. Therefore, the action recognition problem has become the trajectory matching problem. We propose the action recognition algorithm based on longest common subsequence (LCS) method. LCS is used to compare two sequences by finding a longest subsequence of the two sequences: the larger the length of the common subsequence, the more similar the two sequences are. An important merit of LCS is that the subsequence does not need to be contiguous, only the order of the subsequences is maintained. This property enables the LCS create a good balance between the rigid temporal order constraint based methods and bag-of-words based methods which lose temporal information.

Given two sequences *P* and *Q,* the LCS problem can be solved efficiently with dynamic programming, as explained in [10]. Let $P = \{p_1, p_2, ..., p_m\}$, $Q = \{q_1, q_2, ..., q_n\}$. We need to find an LCS of these two sequences. If $p_m = \mathrm{q_n}$, we need find an LCS of P_{m-1} and $\mathrm{Q_{n-1}}$. Adding the number of 1 to this LCS gives an LCS of *P* and *Q*. If $p_m \neq \mathrm{q_n}$, we need solve two subproblems: finding an LCS of P_{m-1} and Q and finding an LCS of *P* and $\mathrm{Q_{n-1}}$. Whichever of these two LCS is longer is an LCS of *P* and *Q*.

We define *c[i, j]* to be the length of an LCS of the sequences P_i and $\mathrm{Q_j}$, then the optimal substructure of the LCS problem gives the following recursive formula:

$$c[i,j] = \begin{cases} 0 & \text{if } \mathrm{i} = 0 \text{ or } \mathrm{j} = 0 \\ c[i-1, j-1]+1 & \text{if } \mathrm{i,j} > 0 \text{ and } \mathrm{p_i} = q_j \\ \max(c[i, j-1], c[i-1, j]) & \text{if } \mathrm{i,j} > 0 \text{ and } \mathrm{p_i} \neq q_j \end{cases} \quad (4)$$

The values of *c[i, j]* are stored in a table *c*. Another table *b* is maintained to simplify construction of an optimal solution. The table *c* has the size of m*n and returns the length of an LCS of *P* and *Q*. Using dynamic programming, the LCS problem can be solved efficiently.

3. EXPERIMENTAL RESULTS

We evaluate our action recognition algorithm using the Weizmann human action dataset from [3]. This dataset contains 10 action classes performed by nine different persons. The actions include bending (bend), jumping jack (jack), jumping-forward-on-two-legs (jump), jumping-in-place-on-two-legs (pjump), running (run), gallopingsideways (side), skipping (skip), walking (walk), waving-one-hand (wave1), and waving-two-hands (wave2). In this dataset, actors have different physical sizes, wear different clothing and perform actions differently both in speeds and motion styles. Figure 3 shows sample frames from video sequences of the Weizmann dataset. There are totally 90 video sequences in our experiments. We adopted a leave-one-out scheme for evaluation. In the leave-one-out strategy, for each type of actions, the video sequences of one person are selected as the test data and the rest as the training data. We repeated this experiment for 40 times and obtained the average recognition rate. The parameters of the SOM are: initial learning rate $\gamma = 0.1$, $\sigma = 5$. The SOM is trained during 2000 iterations. The normalized silhouette images extracted in the feature representation stage have the same 31x31dimention and are transformed to trajectories on the 2D SOM. The LCS scheme implemented by dynamic programming is thus used to compare each pair of the motion trajectories of map units and to classify the corresponding human action sequence. We reach 88.75% average recognition rate for all actions, using an 11x11 SOM. The corresponding confusion table generated by our classification results is shown in Table 1. It can be seen in the table that the proposed approach can recognize most of actions in the dataset with satisfying accuracy. The recognition accuracy of the action "skip" is relatively low, which is because the silhouette sequences of "skip" are quite similar with the silhouette sequences of the confused actions.

In our experiments, we also explore the effect of different SOM sizes on the recognition performance. Figure 4 shows the performance variation with different map sizes such as 4x4, 5x5, 6x6, and so on. We compare our average recognition rates with other published results on the same dataset in Table 2. From this table, we can see that the performance of the proposed SOM based algorithm is comparable to other approaches.

The Self-organizing Map method makes the mapping statically. Then the time order of the pattern vectors does not matter in SOMs. In motion analysis, the time order of patterns is important. It is desirable to include the time order into the representation. Future research directions include incorporation of the temporal context of input streams in SOM in order to describe and classify motion patterns more effectively.

Fig. 3. Sample frames from video sequences of the Weizmann dataset.

	Walk	Bend	Run	Jack	Jump	Pjump	Side	Skip	Wave 1	Wave 2
Walk	100									
Bend		100								
Run	12.5		82.5					5		
Jack	2.5	2.5		95						
Jump	2.5				97.5					
Pjump	2.5			7.5		80	5		5	
Side						2.5	97.5			
Skip	10		20		15			55		
Wave 1						5			92.5	2.5
Wave 2									12.5	87.5

Table 1. The confusion table of action recognition by the proposed algorithm using 11x11 SOM.

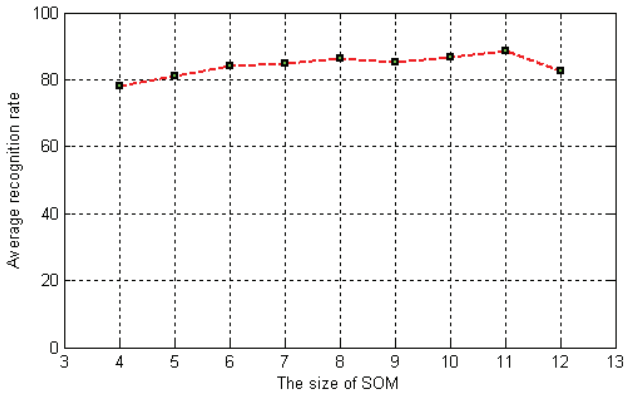

Fig. 4. The effect of different SOM sizes on the recognition performance.

Klaser et al. [2]	Niebles et al. [6]	Scovanner et al. [7]	Our method
84.3	72.8	82.6	88.75

Table 2. Comparison between our approach and others on the Weizmann dataset.

4. CONCLUSION

In this paper we introduced a novel neural network approach for human action recognition based on SOM. We demonstrated the validity of SOM for data clustering and dimensionality reduction in the human action recognition task. The key poses in motion sequences were described by the trained SOM. The human action sequences were thus represented as trajectories of map units. For action recognition, a longest common subsequence algorithm was employed to match these action trajectories on the map robustly. We tested the approach on a well known standard action dataset. The promising experimental results showed the potential of this SOM based action recognition method.

5. REFERENCES

[1] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proc. of ICPR*, pp. III: 32–36, 2004.

[2] A. Klaser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. of BMVC*, 2008.

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, " Actions as space-time shapes," in *Proc. of ICCV*, pp. 1395–1402, 2005.

[4] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.23, no.3, pp.257-267, Mar 2001.

[5] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatio-temporal words," in *Proc. of BMVC*, 2006

[6] J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *Proc. of CVPR*, 2007.

[7] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. of ACM Multimedia*, 2007.

[8] C. Orrite, F. Martinez, E. Herrero, H. Ragheb, and S. Velastin, "Independent viewpoint silhouette-based human action modelling and recognition," in *Proc. of the 1st International Workshop on Machine Learning for Vision-based Motion Analysis (MLVMA'08)*, 2008.

[9] T. Kohonen, *Self-Organizing Maps,* Springer-Verlag, 1995.

[10] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd Edition, MIT Press and McGraw-Hill, 2001.