Volume 51, Number 2, 2010

ACTA TECHNICA NAPOCENSIS
Electronics and Telecommunications

USING DYNAMIC TIME WARPING ALGORITHM OPTIMIZATION FOR FAST HUMAN ACTION RECOGNITION

Tamás VAJDA
Sapientia Hungarian University of Transylvania, Tîrgu Mures, Romania

Abstract: In this paper, we present an approach based on dynamic programming and neural network for recognition and matching human action. The body parts angular motion represents a human action. Each body part angular motion is represented by a one-dimensional time series. These time series are then compared separately for every body part with templates using dynamic programming (DTW). The results of the comparisons are used as input for a neural network that classifies the human action.

Keywords: *Dynamic Time Warping, action recognition.*

I. INTRODUCTION

Recognizing human action from monocular video is a challenging task. Two types of action recognition methods are widely used: state-space model and template based approaches. State space model include neural networks, Hidden Markov Models, and extension of it. On the other hand, template-matching methods compare the action to stored templates. Our method is a hybrid approach. We use template base method to classify the body part motions and state space method to merge the information of classified body part motion and classify the action.

The action recognition is challenging because of the high degree of motion. In addition, the action has a natural compositional structure, which means, that action can be categorized simultaneously into several categories, for instance running and waving. Obviously, even the transition between simple actions has temporal ambiguity and overlap. For a comprehensive survey on human action recognition, we refer to [5,11,8].

A common approach to recognize or model sequential data like human motion is the use of Hidden Markov Model (HMM) [9,12] and conditional random field (CRF) [2,3,4] on both 2D observations and 3D observations. In HMM [1] sequential data is modeled using a Markov model that has finite states. We must choose and determine the number of states in advance for a motion, but the motion can have different time length. Therefore, it is difficult to set the optimal number of state corresponding to each motion.

To resolve this problem we used instead of HMM a special variant of the dynamic time warping method to match the templates.

Other approaches make use of templates or global trajectories of motion. Using global trajectories is highly dependent from the environment where the system is built, and can separate the composed action which introduces high interclass variation making it hard to classify the motion [6,7]. Another problem of using global trajectories in action recognition is that it is very difficult to find the silence point that marks the possible beginning of a new action.

Our approach the matching procedure is done separately for every body part reducing the number of templates. We used a neural network to merge the result of body part motion matching and categorize the action.

In the following section, we will present the Dynamic Time Warping algorithm, the third section we will present the human action decomposition method applied to the detected human to extract the time series, in the forth section we introduce the constrained DTW which optimize the DTW for human motion recognition. In the fifth section we will present briefly the Neural Network used to synchronized the DTW matching results and to get an overall response to the human action. The last two section we will present our experiments and conclusions.

II. THE DYNAMIC TIME WARPING ALGORITHM

The Dynamic Time Warping compares two time series and computes the distance between them, even if the two series are shifted in time axis. Given two series *X*, and *T* equation 1, of lengths |*X*| and |T|,

$$\begin{aligned} X &= x_1, x_2, \ldots x_i, \ldots . x_{|X|} \\ T &= t_1, t_2, \ldots t_j, \ldots . t_{|T|} \end{aligned} \qquad (1)$$

To align two time series we construct a |*X*|-by-|*T*| distance matrix. The (i^{th}, j^{th}) element of the matrix corresponds to the distance between the x_i and the t_j element of the series. To get the distance between the two time series we search the warping path *W* equation 2.

Manuscript received May 8, 2010; revised June 14, 2010

$$W = w_1, w_2, ... w_K, \\ \max(|X|, |T|) \le K < |X| + |T| \tag{2}$$

where *K* is the length of the warp. Every element of the warping path is a pair of coordinates or indexes, which represent a relation between the two time series.

$$w_K = (i, j) \tag{3}$$

where *i,j* represent the index of the two time series. There are three constraints on the warp path. The first constraint is the *boundary condition*: $w_1 = (1,1)$ and $w_K = (|T|, |X|)$. That is mean that the warping path must start at the first element and must end at the last elements of both time series. On the other hand, must start on the bottom left corner, and end on opposite corner of the distance matrix.

The second and the last constraint is the *continuity* and the *monotonicity* merged in equation 4.

$$w_m = (i,j),\ w_{m+1} = (i',j') \\ \mathrm{i} \le \mathrm{i'} \le i+1, \mathrm{j} \le \mathrm{j'} \le j+1 \tag{4}$$

This restrict the allowable step to the adjacent cells including the diagonal cells to so that the *i, j* increase monotically in warping path. Every index from both time series must be used.

Many warping path satisfy this three conditions, however we need that path which optimizes the cumulative distance of the path elements (equation 5).

$$DTW(X,T) = \min\left\{\frac{1}{K}\sqrt{\sum_{l=1}^{K} w_l}\right. \tag{5}$$

The 1/K is normalizing the distance, for the fact that warping paths may have different length.

The best way to construct the optimal warping path is the dynamic programming method. First, the task should be split in subtasks, portions of time series. By finding the optimal solution to these subtasks, we will get the optimal solution to the entire problem. To achieve this we need to construct a cumulative distance matrix D using the following equation 6.

$$D(i, j) = Dist(X_i, T_j) + \min[D(i-1, j), \\ D(i, j-1), D(i-1, j-1)] \tag{6}$$

Every cell is computed as sum of the distance (Euclidian or other type of distance) of current element ($Dist(X_i, T_j)$) and the minimum of the cumulative distance of the adjacent cell. The cost matrix is computed from bottom up and from left to right. After the entire cost matrix is filled, a warp path must be found from left lower corner *D*(1, 1) to *D*(|*X*|, |*T*|) top right corner. The warp path is actually computed in reverse order starting at *D*(|*X*|, |*T*|) using a greedy search that evaluates cells to the left, down, and diagonally to the bottom-left and the smallest valued cells coordinate is added to the beginning of the warp path found so far. The search continues from the last added cell. The search stops when *D*(1, 1) is reached.

III. CONVERTING THE HUMAN MOTION INTO TIME SERIES

We have two convenient ways to represent motion: global trajectories, or decomposing the motion to its basic elements. Because the human motion can be compositional or concurrent, the global trajectories are not the best choice. Some action need only legs for example walk, run, jump, and some only the hand: handshaking, waving. For this reason, we decomposed the action to its basic elements – to body part motion. To make the recognition easier we track every body part individually and relative to its parents body part. Using this approach, we can use only those basic motions (body part motions) in classification that are relevant so we can easily recognize composed motion too.

Some cases, when we have low-resolution images, we cannot track separately all body part motion. In these cases, the only possibility is to track a global motion. There are many possibilities to do that. We can use Haar based detector [15] or we can us chamfer matching [16] to detect the human and its pose.

Our goal is to get the most detailed information about the configuration of the human body, and its relation to other moving objects and environment in the current frame. To achieve this goal we used bottom up approach the chamfer matching [16] in cases of the low resolution images and a top down approach the Pictorial structure method introduced by Felzenszwalb [13] and extended Ramanan [11] in higher resolution images.

In case of the higher resolution frames, the Pictorial structure approach, is modeling the human body as a collection of parts in a deformable configuration, with 'spring-like' connections between pairs of parts. These connections are modeling spatial relations between parts. Appearances and spatial relationships of individual parts can be used to detect an object. Best match of the pictorial structures depends on how well each part matches its location and how well the locations agree with the deformable model. The main advantage of the approach is that the motions of the human body parts are tracked individually and relative to its parent's body part. Using this approach, we can use only those basic motions (body part motions) in classification that are relevant and we can easy recognizing composed motion too. The first and most significant motion is the torso motion. Here we look at two elements the motion relative to the image (global motion) and the angular motion. The torso represents the root of body parts in the pictorial structure. The upper legs, and upper arms are connected to the torso and we analyze only their angular motion between -270 and +270 grades. The absolute motion is tracked between -180 and +180. The 180 and 270 represent a tampon zone. If the motion angle is above 180 or below -180, we will have two possible time series. Three events can reset one of the time series [22]: the angular motion returns quickly between -180 and 180 degree; the DTW matching for one of them has a strong result or the angle is increase above 270 or decreases below -270. (Figure 1).

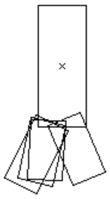

Figure 1. Relative motion of the upper leg relative to the torso

The lover legs are connected and their angular motions are relative to the upper legs. In addition, the lover arm angular motions are tracked relative to the upper arm. We do not track the motion of head.

In Figure 2. is presented a time series of motion for the upper arm representing the waving action.

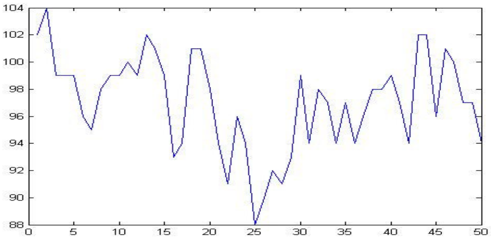

Figure 2. Full resolution time series of waving – upper arm

Most important point in motion series is the peaks and the still (constants) points and the zero crossing point, because they mark a change in the motion direction and they are stable or typical position of the human body. Knowing that the same action can be done in different speed, the time between two direction changes in a body part motion is not so relevant.

In case of the low-resolution frames, we use the chamfer matching method [16] to track and to detect the pose of the human body. Using the fast template search method introduce by the author we always can track the human body and measure the distance from the closest template class.

We can approximate the motion series using the key positions. There is a unequivocal mapping of the key position to relative position of all body parts. We always map this position when the current match has the lowest distance from the template. We count the number of frames between two consecutive best match key position and interpolate the intermediate points.

Using these two approaches, we are able to connect them and provide a general framework based on DTW to recognize the human action.

IV. CONSTRAINED DYNAMIC TIME WARPING METHODS

The quadratic time and space complexity of DTW creates the need for methods to speed up dynamic time warping. The most common method is the use of constrains, which limit the search area in the cost matrix. This constrain is important not only for speed up of the DTW but also to eliminate the problem of singularities [13,14,17].

There are other methods to speed up the computation of the DTW like the FastDTW [11], which use a recursive shrinking and refine to get the best warping path.

To compare the time series of the human motion we use an adapted version of DTW algorithm, which has a multilevel approach with following key operations:

- Shrinks a time series into a smaller time series that represents only the peak or constant vales from the time series,
- Coarse DTW - Finds a minimum-distance warping path for the shrunk series and uses that warping path as an initial guess for the full "resolution's" minimum-distance warp path
- Final DTW - Refines the warping path projected from a lower resolution through local adjustments of the warping path using Sakoe-Chiba constrain.

The first is in the coarsening step we shrink the time series. Human body part motions most significant moments are the direction changes. In our approach instead of averaging the time series, we use a heuristic selection of the data keeping only the peaks and constant vales from the series. This is done by keeping only those x_i elements from the X if the one of the next two conditions is true:

$$((x_i \le x_{i-2}) \wedge (x_i \ge x_{i+2})) \,\|\, ((x_i \ge x_{i-2}) \wedge (x_i \le x_{i+2})) \qquad (7)$$

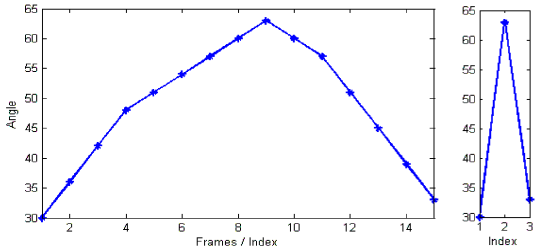

Fig. 3. The original and the shrink time series of waving - upper arm

Fig. 3. presents the original time series of waving upper arm and the shrunken series. The second step we make a classical DTW comparison of the shrunken templates and the shrunken input. Using this comparison, we can eliminate the majority of the templates and only a few templates need to be compared at higher resolution.

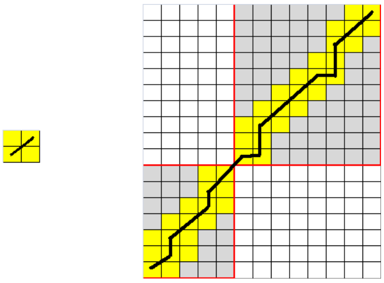

Fig. 4. The coarse and the full resolution cost matrix with warping path

Fig. 4 shows the shrunk time series cost matrix and the projection of this to the original resolution cost matrix. Projection takes a warping path calculated at a lower resolution and determines what cells in the next higher resolution time series the warping path passes through. This projected path is then used as heuristic during solution refinement to find a warping path at a higher resolution. To make it faster we use Sakoe-Chiba band constrain.

The Final DTW step is a refinement that finds the optimal warping path in the neighborhood of the projected path, where the size of the neighborhood is determined locally by distance between two consecutive points in shrunk series and the difference between the length of the template series and the input series.

This will find the optimal warping path through the area of the warping path that was projected from the lower resolution.

V. ACTION CLASSIFICATION USING NEURAL NETWORKS

The DTW matching of the input series give us for every body part a set of response. These are the probabilities of matching a class of template. We have to synchronize responses and make a final decision about the overall human action.

Neural networks (NNs) are nonlinear models, which makes them flexible in modeling real world complex relationships. Furthermore, NNs are data driven self-adaptive methods in that they can adjust themselves to the data without any explicit specification of functional or distributional form for the underlying model.

Although many types of neural networks can be used for classification purposes, our focus is on three network types, specifically the Learning Vector Quantisation (LVQ), Radial Basis Function (RBF) and feedforward multilayer networks or MultiLayer Perceptron (MLP) NNs, which are the most widely studied and used neural network classifiers.

We use the result of DTW as input for Neural Networks. That is mean, that we equal input neurons in the neural network as, template classes. We normalized the output of the DTW matching between 0 and 1. This is necessary because we have to compare with DTW different length of time series.

For every behavior type, we used an output neuron. That make possible to handle composed behavior like waving and walking. Using this approach has also a drawback by reducing the speed of the network and by increasing the number of free parameters. The strait influence of the increased free parameters is that we need to create bigger training set.

VI. EXPERIMENTS

We used the detected position and the configuration of the human body the pictorial structure method [21].Using this method we can measure the speed of torso, and to track the relative motion of the body parts relative to their parents.

Using the human motion decomposition, we extract the one dimensional time series for every body part. These time series are compared to the saved templates using the constrained DTW[20], classic DTW[13-14] and the FastDTW [11] . Using the constrained DTW the majority of the templates are eliminated at early stage if the distance between the coarse variant of the series is bigger than a threshold.

To construct the templates database we have annotated and saved 5 different actions from 20 different videos. For every body part, we compared the saved motion series with the adapted DTW. If the difference between them were too big, we dropped. If they were similar, we choose the median series from them.

The matching results were used as input for three type of Neural Network: LQV, RBF and MLP.

Using output of the DTW, a dataset has been compiled. In order to follow the proper steps in the design of a test bench system, this dataset was split in a training subset (75% of the samples) and a testing subset (25%). Five different behaviors are represented in the compiled dataset by about twenty measurement results each (values corresponding to eight body part motion time series). The NN training has been executed in Matlab, using the embedded functions of this environment.

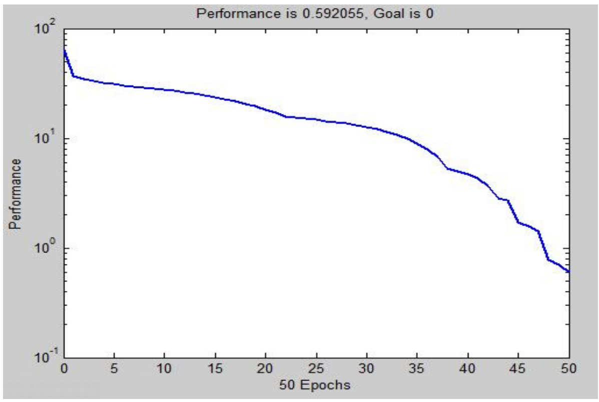

Fig. 5. Training chart of the RBF NN

The evolution of the classification accuracy of the RBF NN and the LVQ NN during the training phase is presented in Fig. 5. and Fig. 6., respectively.

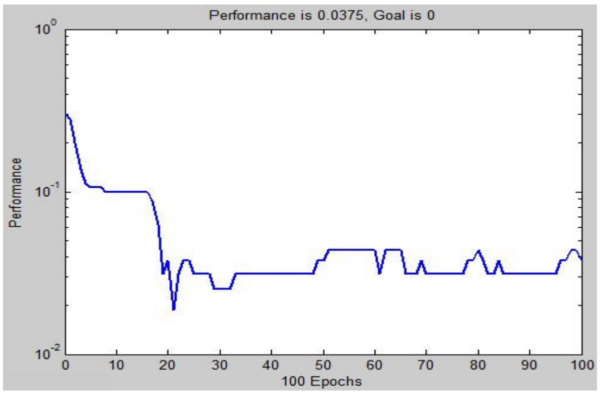

Fig. 6. Training chart of the LVQ NN

Table 1 summarizes the results obtained at matching using different type of DTW.

Table 1. Matching the time series

DTW	Processing Time (100 time series/millisecond)	Classification accuracy
classic DTW	9976	98,3%
FastDTW	1954	97,2%
constrained DTW	987	98,2%

Table 2 summarizes the actual results obtained with these methods, proving, that using NN to recognize human action from monocular video after special DTW processing is a viable solution.

Table 2. NN classification results

NN type	Neurons Inp./Hidd./Out.	Classification accuracy after max.100 epochs
LVQ	40/80/5	85%
RBF	40/5	75%
MLP	40/40/5	76%

We need to mention that we compute this accuracy percentage by comparing a manual labeling with the output of the Neural Network. This means that measurement summarized in table 2, measures not only the errors of the Neural Network but also the errors provided by the DTW matching. We used this method to measure also the capacity of the Neural Networks to handle erroneous data. Most of the cases not all the input are erroneous only some of them. The system was tested on indoor and outdoor environment and was no significant difference on the system behavior in these two cases. The difference was provided by the behavior of the human pose detection method. Using the motion decomposition, we were able to recognize also the composite action like standing and handshaking. In figure 7 and 8 we present the output of the system.

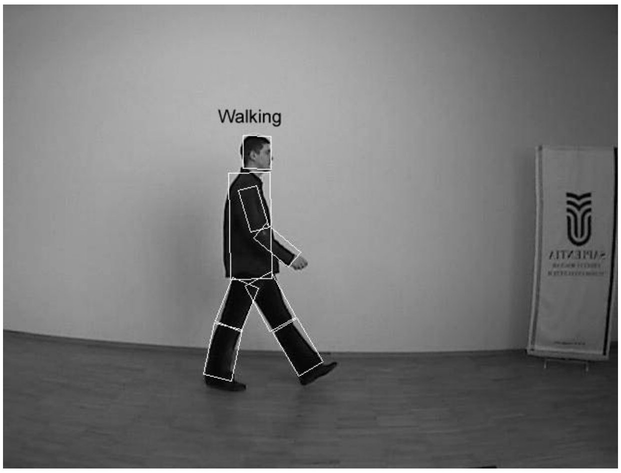

Figure 7. Output of the system single person

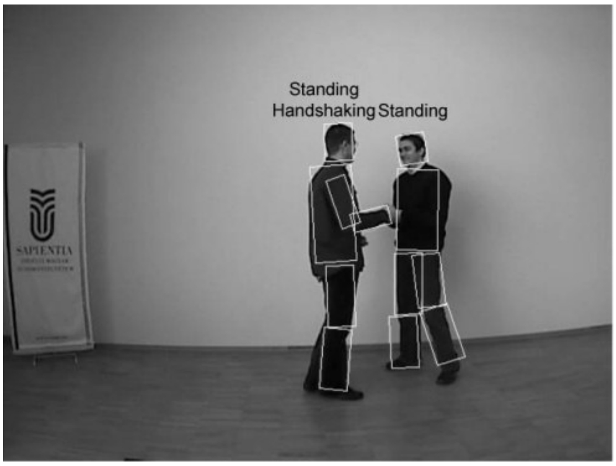

Figure 8. Output of the system two people

VII. CONCLUSION

We have presented two improvements for human action recognition: an efficient representation of motion by decomposing to its basic elements and a constrained DTW algorithm enhanced for human motion recognition purpose. Both ideas can be future improved. In case of the body part motions the angular motion can be decomposed to two time series one which contains the low frequency variation and one with high frequency time series. The low frequency will represent the position, and the high frequency will represent the short action of the body part. We also proved that the constrained DTW introduced by the article is the fastest method keeping the performance of the classical DTW.

We also show that using Neural Network is a good choice to synchronize the matching result and to provide a final action type even if this action is a composite one.

REFERENCES

[1.] L. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc. IEEE, 77(2):257–286, 1989.

[2.] J. Laffey, A, McCallum, and F Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”, Pmc. 18th ICML, pages 282-289, 2001.

[3.] Sminchisescu, C.; Kanaujia, A.; Zhiguo Li; Metaxas, D.; “Conditional models for contextual human motion” recognition Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, Page(s): 1808 - 1815 Vol. 2, 17-21 Oct.

[4.] Okada, S., Hasegawa, O., “Motion Recognition based on Dynamic-Time Warping Method with Self-Organizing Incremental Neural Network” Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, pages 1-4.

[5.] J. Aggarwal and Q. Cai. Human Motion Analysis: A Review. *CVIU*, 73(3):428–440, 1999.

[6.] M. Black and A. Jepson. A probabilistic framework for matching temporal trajectories:Condensation-based recognition of gestures and expressions. In ECCV, 1998.

A. Blake, B. North, and M. Isard. Learning Multi-Class Dynamics. NIPS, 11:389–395, 1999.

[7.] Bobick and J. Davis. The recognition of human movement using temporal templates. In PAMI, 2001.

[8.] M. Brand, N. Oliver, and A. Pentland. Coupled Hidded Markov models for complex action recognition. CVPR, 1996.

[9.] D. Gavrila. The Visual Analysis of Human Movement: A Survey. CVIU, 73(1):82–98, 1999.

[10.] S. Gong and T. Xing. Recognition of group activities using dynamic probabilistic networks. ICCV, 2003.

[11.] S. Salvador & P. Chan. “FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space” *KDD Workshop on Mining Temporal and Sequential Data*, pp. 70-80, 2004.

[12.] Pedro F. Felzenszwalb, Daniel P. Huttenlocher. s.l “*Pictorial Structures for Object Recognition..*” Intl. Journal of Computer Vision, 2005.

[13.] C. Ratanamahatana and E. Keogh, "Three Myths about Dynamic Time Warping", In Proc of SIAM Intl. Conf. on Data Mining, Newport Beach, California, 2005.

[14.] H. Sakoe, and S. Chiba, Dynamic Programming Algorithm Optimization for Spoken Word Recognition, In IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-26, 1978.

[15.] Tamás V. and Lőrinc M. “*General framework for human object detection and pose estimation in video sequences”*, In 5th IEEE International Conference on Industrial Informatics, vol.1, pp 467 – 472, 23-27 June 2007.

[16.] Tamás Vajda, Emőke Szatmári, Sergiu Nedevschi “*Human Body Detection and Tracking in Video Sequences Using Chamfer Matching”,* Intelligent Computer Communication and Processing, 2007 IEEE International Conference on Intelligent Computer Communication and Processing, 6-8 Sept. 2007, p. 141-146

[17.] E. Keogh, M. Pazzani (2001) „Derivative dynamic time warping□, Proceedings of the First SIAM International Conference on Data Mining, Chicago, USA, 2001.

[18.] Ying Xie, Bryan Wiltgen,” Adaptive Feature Based Dynamic Time Warping” International Journal of Computer Science and Information Security Vol. 10 No. 1 pp. 264-273, 2010.

[19] Tamás Vajda, László Bakó, Sándor Tihamér Brassai. Using dynamic programing and Neural Network to Match Human Action,11th International Carpathian Control Conference ICCC 2010, May 26-29 2010, Eger Hungary, pp 231-234.

[20] T. Vajda, Action Recognition Using DTW and Petri Nets, Studia Universitatis Babes-Bolyai Series Informatica, Volume LV, Number 2 (June 2010), pp 69-78

[21] Tamas Vajda Behavior Recognition Using Pictorial Structures and DTW 2010 IEEE International Conference on Automation, Quality and Testing, Robotics, Mai 28-29 2010, vol3, pp 198-201

[22] Tamás Vajda Action Recognition Based on Fast Dynamic-Time Warping Method IEEE 5th International Conference on Intelligent Computer Communication and Processing, ICCP 2009, Aug 27-29. 2009, pp. 127 – 131