STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LV**, Number 2, 2010

ACTION RECOGNITION USING DTW AND PETRI NETS

TAMÁS VAJDA

Abstract. This paper proposes a new approach for recognition in monocular video the human behavior sequence. We use a simple to complex approach in action recognition by decomposing it to its basic elements. The human body parts motions are tracked and classified individually. The body parts motions are matched using an adapted Dynamic Time Warping (DTW) method, an approximation of DTW algorithm that has linear time and space complexity. The adapted DTW uses a three step approach and in the second step we may eliminate the most of the incorrect template which reduces the time for comparing the entire template database. The results of the DTW matching are used to activate hierarchical Petri Nets used to classify the behavior.

1. Introduction

Recognizing human behavior from monocular video sequences is one of the most promising application of computer vision. The behavior recognition has two big issues: the first one is the human tracking which represents the measurement stage and the second is the recognition stage which is the measurement processing stage. We will focus here mostly on the second issue. The behavior recognition is challenging because of the high degree of motions, the coarsest human model is represented by 28 dimensions, and missing or erroneous measurement. Due to the high degree of motion, the actions can be often classified into several categories simultaneously. Some activities have a natural compositional structure. Behavior is composed mostly from basic action units (run and hand-wave, walk and shake hands). Even the transition between simple activities naturally has temporal segments of ambiguity and overlap. The research devoted to human motion recognition is extensive, we refer to [5, 11, 8] for comprehensive surveys. A common approach to recognize or model sequential data like human motion is the use of Hidden Markov Model (HMM) on both 2D observations [9, 12] and 3D observations. In HMM

Received by the editors: December 7, 2009.

2010 *Mathematics Subject Classification.* 68T45.

1998 *CR Categories and Descriptors.* Code [**I.4.8**]: Subtopic – *Motion.*

Key words and phrases. Behavior recognition, DTW, Petri Nets.

This paper has been presented at the International Conference Interdisciplinarity in Engineering (INTER-ENG 2009), Târgu-Mureş, Romania, November 12–13, 2009.

[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. Recently, there has been increasing interest in using conditional random field (CRF) [2, 3] for learning of sequences. The advantage of CRF over HMM is its conditional nature, resulting in relaxation of the independence assumption, which is required by HMM to ensure tractable inference [4]. But all these methods assume that we know the number of states for every motion. 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 which mark the possible beginning of a new action. The main contribution of this paper is the introduction of novel action representation. Using the decomposition method we can create an environment independent representation of different actions by representing every body part motion relative to its parent. The second contribution is DTW adaptation for human motion recognition. In this paper we will also demonstrate that the Petri Nets are suitable for behavior recognition.

2. Pictorial structure based human detection and posture estimation

The first step for behavior recognition is the measurement. In our case we want to know the current configuration of the human body, its relation to other moving objects, and its relation to its environment. To achieve this goal we used the Pictorial structure method introduced by Felzenszwalb [13]. In this approach the human body is modeled by 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. Matching a pictorial structure does not involve making any decisions about location of individual parts; more work is to find a global minimum of energy function without any initialization.

The pictorial structure model can be represented as an undirected graph $G = \{V, E\}$, where V represents the body parts set and E represents the relation between parts. An instance of the object is given by its configuration $L = \{l_1 \dots l_n\}$ where l_i is the location of part v_i.

To find the best match of a body configuration within an image we find the L^* that minimizes the sum

$$\arg\max_L(\sum_{i=1}^{n} m_i(l_i) + \sum_{(v_i,v_j)\epsilon E} d_{ij}(l_i, l_j)) \tag{1}$$

where $m_i(l_i)$ measures the cost of mismatching part v_i. with location l_i and $d_{ij}(l_i, l_j)$ measures the cost of deforming the model when placing v_i at location l_i and v_j at l_j. To measure this deformation we use Mahalanobis distance. By using the statistical framework proposed by Felzenszwalb the Pictorial structures can be viewed as an energy minimization problem in terms of statistical estimation. In this framework we need the model parameters $M = (u, E, c)$ where $u = u_1 \dots u_n$ are appearance parameters, E indicates connections between parts and $c = \{c_{ij}|(v_i, v_j)\epsilon E\}$ represents connection parameters. These parameters are learned from training examples.

Using the Posterior Sampling method, we get for every frame the optimal location and configuration of body model.We used the pictorial structure on a background-subtracted image and a feature image. As result of detection we get the body's relative configuration, the absolute position of body parts.

3. Action decomposition

Human motion can be represented in many ways: silhouette, volumetric representation of motion, temporal templates or global trajectories. We have two 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 actions 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 motions. 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 the classification which are relevant so we can easily recognize 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 values represent a buffer 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: the angular motion returns quickly between -180 and 180 degrees; the DTW matching for one of them has a strong result or the angle is increased above 270 or decreased below

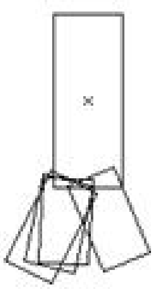

Figure 1. Relative motion of the upper leg relative to the torso

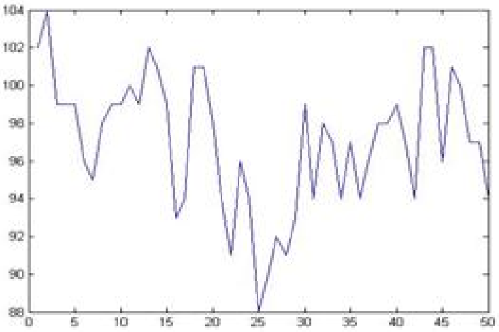

Figure 2. Full resolution time series of waving - upper arm

-270. (Figure 1). The lower legs are connected and their angular motions are relative to the upper legs. Also the lower arm angular motions are tracked relatively to the upper arm. We do not track the motion of the head. In Figure 2. a time series of motion is presented for the upper arm representing the waving action.

The most important point in the motion series are the peaks and the still (constant) points, because they mark a change in the motion direction. Knowing that the same action can be done at different speed the time between two direction changes in a body part motion is not so relevant.

4. Dynamic-Time Warping method for behavior recognition

Body part motions are time series in which every measurement is an element from the series with constant time periods between them. The best way to compare saved template motion and measurement, which can be time series of different length, is the dynamic time warping (DTW) algorithm. The DTW compares two time series where we note the template series with T and the measurement with X, of lengths $|T|$ and $|X|$,

$$\begin{gathered} X = x_1, x_2, ..., x_i, ...x_{|X|} \\ T = t_1, t_2, ..., t_j, ...x_{|T|} \end{gathered} \tag{2}$$

construct a warping path W

$$\begin{gathered} W = w_1, w_2, ..., w_|k| \\ \max(|X|, |T|) \leq k \leq |X| + |T| \end{gathered} \tag{3}$$

where k is the length of the warping path at the n^{th} element of the warping path is

$$w_n = (i, j) \tag{4}$$

where i, j are an index from time series X, and T. Every index of both time series is to be used in the warping path and need to increase monotonically in the warp path. The minimum-distance warping path is optimal, where the distance of a warping path W is

$$Dist(W) = \sum_{n=1}^{n=k} Dist(w_{ni}, w_{nj}) \tag{5}$$

$Dist(W)$ is the distance of warping path W, and $Dist(w_{ni}, w_{nj})$ is the distance between the two data point indexes in the n^{th} element of the warp path. A two-dimensional $|X|$ by $|T|$ cost matrix D is constructed, where the value at $D(i, j)$ is the minimum distance warping path that can be constructed from the two time series $X' = x_1, x_2, \ldots, x_i$ and $T' = t_1, t_2, \ldots, t_j$. The value at $D(|X|, |T|)$ will contain the minimum-distance warping path between time series X and T. To find the minimum-distance warp path, every cell of the cost matrix must be filled.

$$D(i, j) = Dist(i, j) + \min(D(i-1, j), D((i, j-1), d(i-1, j-1)) \tag{6}$$

$D(i, j)$ is the minimum warp distance of two time series of lengths i and j, if the minimum warp distances are already known for all slightly smaller portions of that time series that are a single data point away from lengths i and j. After the entire matrix is filled, the warping path is actually calculated in reverse order performing a greedy search that evaluates cells to the left, down, and diagonally to the bottom-left starting at $D(|X|, |T|)$ to $D(1, 1)$.

Whichever of left, down, and diagonally adjacent cells has the smallest value is added to the beginning of the warping path found so far, and the search continues from that cell. The search stops when $D(1, 1)$ is reached. To speed up and to avoid the marginalized warping path a slop constrain is introduced by Sakoe-Chiba [14]. Our adapted DTW version uses a merged version of Fats DTW introduced by Stan Salvador and Philip Chan [12] and the Sakoe-Chiba band constrain[14]. The adapted version of DTW algorithm uses a multilevel approach with following key operations:

(1) Shrink — Shrinks a time series into a smaller time series, that represents only the peak or constant values from the time series;
(2) 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;
(3) Final DTW — Refines the warping path projected from a lower resolution through local adjustments of the warping path using Sakoe-Chiba constrain.

There are some major enhancement compared to the Stan Salvador and Philip Chan's FastDTW algorithm. The first is in the coarsening step. The FastDTW only computes an average of the neighborhood values and run several times to produce many different resolutions of the time series. Using this method to shrink the time series we may lose important information and get a low compression of the data. Compared to other time series in the series representing human body part motions the most significant moments are the direction changes. The shrinking average operation smooths the time series causing loss of information. In our approach instead of averaging the time series we use a heuristic selection of the data keeping only the peaks and constant values from the series. This is done by keeping only those x_i elements from X if the one of the next two conditions is true:

$$((x_i \leq x_{i-1}) \wedge (x_i \geq x_{i+1}))||((x_i \geq x_{i-1}) \wedge (x_i \leq x_{i+1})) \tag{7}$$

The resulting time series is a smaller one or equal to the original time series and we lose very low amount of information.

Figure 3 represents 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 template and only few template need to be compared at higher resolution.

Figure 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

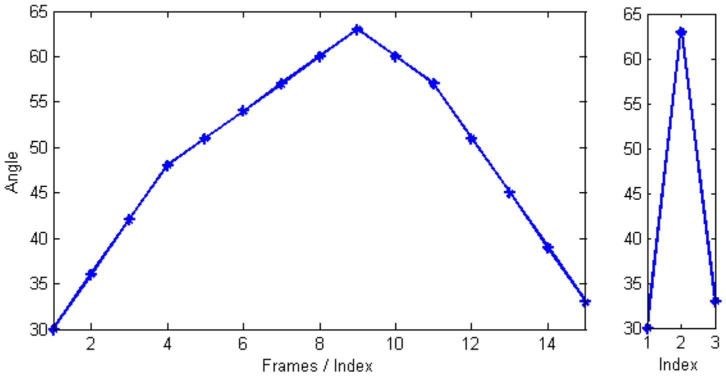

FIGURE 3. The original and the shrink time series of waving - upper arm

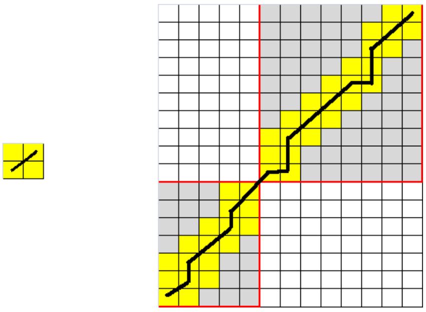

FIGURE 4. The coarse and the full resolution cost matrix with warping path

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, finds the optimal warping path in the neighborhood of the projected path, where the size of the neighborhood is determined locally by the 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.

The motion templates are computed using a dataset of labeled motion. For these motions the shrunken time series is computed. We choose as starting point the median shrunken time series of a motion. Using this series we compute a mean of all time series. The resulted template will be of the same length as the median series.

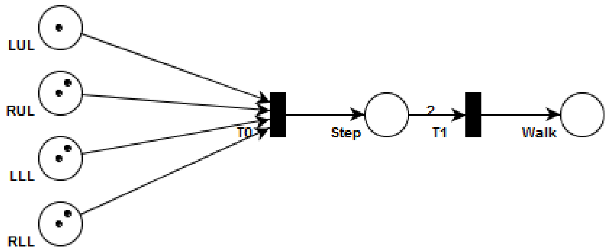

Figure 5. Output of the system single person

5. Behavior Recognition using Petri Nets

Human motion has two types: short timescales and basic motion (sustained (running, walking, jogging, etc.- typically periodic), punctuate (jump, punch, kick), parametric (reach, etc.)), and long timescale and complex composite motion (walking and waving, reading a book, etc.) The motion structure is a hierarchical one. The complex behavior is composed of simple motion, and the motion is composed of simple movements — "basic action". The use of Petri Net was proposed by many researchers but they applied it only to complex behavior based on the recognized basic action. We proposed the extension of this Petri Networks to recognize the basic action too, and the input for this network is represented the output of the DTW. The DTW comparison only categorize the body part motion into the classes. Each class has an associated place in the network. If the DTW categorizes a body part motion in a class the associated place get a token. Using the Petri nets synchronization procedure we can decide about the actual basic activity.

To exemplify this we model a simple walking behavior. The step state is activated by the adapted DTW and the repeated activation of the step state activates the walking state in Figure 5. In the Petri Net the states may or may not represent a behavior. By adding new labeled states we may extend the Petri net to recognize new motion.

6. Experiments

We used the detected position and the configuration of the pictorial structure to measure the speed of torso and to track the relative motion of the body parts relative to their parents. These parameters are compared to the saved templates using the adapted FastDTW and are eliminated at an 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 4 different actions from 10 different videos. For every body part we compared the saved

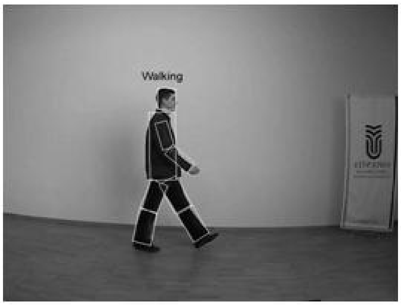

Figure 6. Output of the system single person

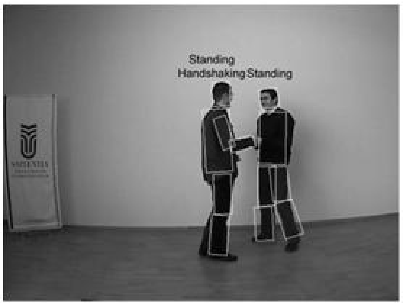

Figure 7. Output of the system two person

motion series with the adapted DTW. If the difference between them is too large they are dropped. If they are similar we choose the median series from them. To recognize the behavior a hierarchical Petri Net was used. The net was designed manually and the parameters were tuned by experiment. For experiments indoor scenes were used, with simple and composed actions. In Figures 6 and 7 we present an output of the system.

7. Conclusions

Two improvements of human action recognition have been presented: an efficient representation of motion by decomposing it to its basic elements and a FastDTW algorithm adapted for human motion recognition purpose. The angular motion representation introduced by the paper is efficient by reducing the matching problem to a 1D matching problem. By using a domain bigger than 360 degrees with the two hypothesis approach we eliminate the error introduced by the angle between +180 and -180. Using the adapted DTW the recognition is two time faster than with the FastDTW because we can eliminate many of templates at the first step at coarse comparison. Using the motion decomposition and the hierarchical Petri Nets we were able to recognize also the composite actions such as standing and handshaking.

References

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

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

[3] Sminchisescu C., Kanaujia A., Zhiguo Li, Metaxas, D., "Conditional models for contextual human motion" ICCV 2005. Tenth IEEE International Conference on recognition Computer Vision, 17-21 Oct 2005, vol. 2, pp. 1808-1815.

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

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

[6] M. Black and A. Jepson, A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. In ECCV, 1998, pp. 909-923.

[7] A. Blake, B. North, and M. Isard, Learning Multi-Class Dynamics. NIPS, 1999, pp. 389-395.

[8] A. Bobick and J. Davis, The recognition of human movement using temporal templates. In PAMI, 2001, pp. 257 - 267.

[9] M. Brand, N. Oliver, and A. Pentland, Coupled Hidded Markov models for complex action recognition. In CVPR, 1996, pp.994-1000.

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

[11] S. Gong and T. Xing, Recognition of group activities using dynamic probabilistic networks. In ICCV, vol. 2, 2003, pp. 742-750.

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

[13] Pedro F. Felzenszwalb, Daniel P. Huttenlocher. s.l, Pictorial Structures for Object Recognition Intl. Journal of Computer Vision, 2005, pp.55-79.

[14] Sakoe H. and S. Chiba, Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoustics, Speech, and Signal Proc., ASSP-26, 1978, pp. 43 - 49.

Electrical Engineering Department, Faculty of Technical and Human Sciences, Sapientia University, Târgu Mureş, Romania

E-mail address: vajdat@ms.sapientia.ro