21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Non-Markovian Dynamic Time Warping

Seiichi Uchida[1], Masahiro Fukutomi[1], Koichi Ogawara[2], Yaokai Feng[1]
[1]*Kyushu University, Japan* [2]*Wakayama University, Japan*

Abstract

This paper proposes a new dynamic time warping (DTW) method, called non-Markovian DTW. In the conventional DTW, the warping function is optimized generally by dynamic programming (DP) subject to some Markovian constraints which restrict the relationship between neighboring time points. In contrast, the non-Markovian DTW can introduce non-Markovian constraints for dealing with the relationship between points with a large time interval. This new and promising ability of DTW is realized by using graph cut as the optimizer of the warping function instead of DP. Specifically, the conventional DTW problem is first converted as an equivalent minimum cut problem on a graph and then edges representing the non-Markovian constraints are added to the graph. An experiment on online character recognition showed the advantage of using non-Markovian constraints during DTW.

1. Introduction

Dynamic time warping (DTW) between two sequential patterns, $X = x_1, \dots, x_t, \dots, x_T$ and $Y = y_1, \dots, y_\tau, \dots, y_T$, is formulated as the optimization problem of the warping function $\tau = u_t$, which is shown as a path in Fig. 1 (a) and matches x_t to $y_\tau (= y_{u_t})$. DTW is often called dynamic programming matching or elastic matching and has widely been employed for compensating nonlinear timing fluctuation in sequential patterns, such as speech, handwriting, gesture, human activity, etc.

DTW is often implemented with some constraints in order to exclude unexpected warping functions which result in unnatural matching, or so-called over-fitting. A typical constraint is the monotonicity and continuity constraint, expressed as $0 \le u_{t+1} - u_t \le \epsilon$, where ϵ is a small integer (often 2). This constraint realizes a monotonic and continuous (i.e., smooth) warping function as shown in Fig. 1 (a).

An important fact about the monotonicity and continuity constraint is that it is a *Markovian constraint*, which restricts the relationship between u_t and u_{t+1}. It is well-known that under Markovian constraints the globally optimal warping function can be derived by dynamic programming (DP) efficiently with $O(T^2)$ computations. In other words, Markovian constraints have been forcibly employed for utilizing DP as the solver of the DTW problem. In fact, most DTW problems have been solved by DP (or its stochastic extension, HMM) under some Markovian constraints.

However, many sequential patterns have *non-Markovian* characteristics. Consider a handwriting trajectory of "0". The position of its ending point should be constrained by the starting point in order to form a closed-circular shape of "0". This example clearly shows that our writing process is obviously non-Markovian. In fact, we can find non-Markovian characteristics everywhere — we are often referring to not only the latest point but also long past points. Similarly, it is highly probable that nonlinear timing fluctuation also has some non-Markovian characteristics. For dealing this characteristics, we need to incorporate non-Markovian constraints which restrict the relationship between u_t and u_{t+n} $(n > 1)$. Unfortunately, as discussed above, the conventional DP-based DTW cannot deal with them[1].

The main contribution of this paper is to realize a non-Markovian DTW, which can incorporate non-Markovian constraints, such as $\alpha \le u_{t+n} - u_t \le \beta$. To the authors' best knowledge, this is the first realization of the non-Markovian DTW. The key idea for this realization is to use graph cut instead of DP. In the proposed method, the DTW problem is re-formulated as a graph cut problem. Since graph cut does not assume any Markovian characteristics of the problem, it is possible to incorporate non-Markovian constraints in addition to the Markovian constraints.

2. Related Work

The proposed DTW method employs graph cut instead of DP. Graph cut can also obtain a globally optimal solution efficiently and thus has been applied to

[1] Precisely speaking, it is "theoretically" possible to deal with the relationship between u_t and u_{t+n} in DP by treating n variables $(u_t, u_{t+1}, \dots u_{t+n})$ as a single high-dimensional variable. However, the computational complexity of the resulting DP algorithm becomes an exponential order of n [1] and thus the algorithm is computationally intractable even with small n.

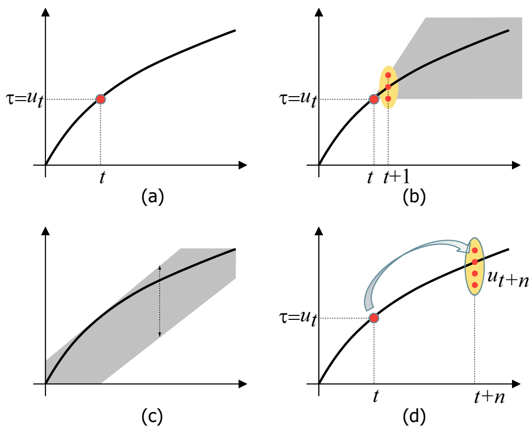

Figure 1. (a) Warping function $\tau = u_t$. (b) Markovian constraint. (c) Range limitation.(d) Non-Markovian constraint.

various optimization problems [2] including matching tasks between 1D patterns [3, 4], 2D patterns [5] and 3D patterns [6]. Coefficients for unary and binary terms in graph cut can be learned statistically from training patterns [7, 8, 9]. These facts enable graph cut to deal with recognition tasks [9, 5] where DP is often used. Since its optimization principle is totally different from DP, it has a potential to avoid the limitation of DP, such as Markovian characteristics. However, most of the previous approaches using graph cut cannot handle non-Markovian characteristics in matching or recognition tasks [3, 4, 5, 6].

Ishikawa et al. [10] have applied graph cut to stereo matching problem, which is essentially the same as DTW. The proposed method extends their method by newly introducing non-Markovian constraints. As emphasized in Section 1, non-Markovian characteristics can be observed everywhere and therefore this promising extension will give a new research direction to DTW and other nonlinear matching problems.

3. Problem Formulation of DTW

Let us start with a conventional DTW where only a Markovian constraint is imposed. The Markovian DTW is formulated as the minimization problem of the following objective function F with respect to the warping function $U = u_1, \ldots, u_T$:

$$\min F = \min_{\substack{u_1,\ldots,u_T \\ \alpha_{t,t+1} \le u_{t+1}-u_t \le \beta_{t,t+1}}} \sum_{t=1}^{T} d_t(u_t) \qquad (1)$$

where $d_t(u_t)$ denotes the local matching cost between x_t and y_{u_t}. As noted before, the monotonicity and continuity constraint $\alpha_{t,t+1} \le u_{t+1} - u_t \le \beta_{t,t+1}$ is a Markovian constraint. Since in the conventional DTW the lower bound $\alpha_{t,t+1} = 0$ and the upper bound $\beta_{t,t+1} = \epsilon = 2$ are fixed for all t, this constraint can be rewritten simply as $0 \le u_{t+1} - u_t \le 2$; however, for the latter discussion, we use the above generalized notation. The boundary conditions $u_1 = 1, u_T = T$ are also assumed during the minimization.

It is noteworthy that the Markovian constraint is rather lax. Figure 1 (b) shows the area of possible paths under the constraint after the path goes through a point (t, u_t). This figure indicates that the relationship between u_t and u_{t+n} becomes weaker for larger n and thus the conventional Markovian DTW is too flexible to avoid over-fitting. In the conventional DTW, therefore, a warp range limitation ($|t - u_t| \le \delta$) of Fig. 1 (c) has often been imposed to restrict the warping function.

4. DTW as a Graph Cut Problem

The globally optimal solution of (1) can be obtained by not only DP but also graph cut as proved in [10]. As noted before, DP treats the DTW problem as a path optimization problem on a t–τ plane as shown in Fig. 2(a). In contrast, the graph cut treats the DTW problem as a minimum cut problem on a directed graph $G = (V, E)$ of (b). Hereafter, using Fig. 2, we will show how this graph (b) is derived from (a).

The nodes V are comprised of two special nodes (source and sink) and other nodes, each of which corresponds to a grid point (t, τ) on the 2D-space of (a). Specifically, as shown in (c), for every grid point (t, τ), a pair of nodes $(v^0_{t,\tau}, v^1_{t,\tau})$ are prepared in G.

The edges E are comprised of several groups having different roles. First, the edge between $v^0_{t,\tau}$ and $v^1_{t,\tau}$ has a weight $d_t(\tau)$. Consequently, the cost for the path in (a) is equivalent to the cost for the cut in (c). Second, the edges connected to sink or source nodes in (d) have an infinity weight and represent boundary conditions. For example, the two cuts in (d) represent the cases of $u_1 = 3 \ne 1$ and $u_T = 2 \ne T$, respectively. They never happen because they have an infinite cost by cutting those edges. Third, the edges in (e) and (f) also have an infinite weight and play a role to exclude unexpected cuts (like “NG” cuts in these figures).

The monotonicity and continuity constraint with the lower bound $\alpha_{t,t+1}$ and the upper bound $\beta_{t,t+1}$ is also represented by edges with an infinite weight. The edges in (g) are for $\alpha_{t,t+1} = 1$ (where, for a simpler visibility, $\alpha_{t,t+1}$ is fixed at 1 instead of 0) and the edges in (h) are for $\beta_{t,t+1} = 2$.

Now we have all the elements, that is, nodes and edges of the graph G of (b). In short, the nodes and the edges of (c) are prepared for representing the local costs and the edges of (d)-(h) are prepared for regulating possible cuts equivalent to possible paths. Thus, we can

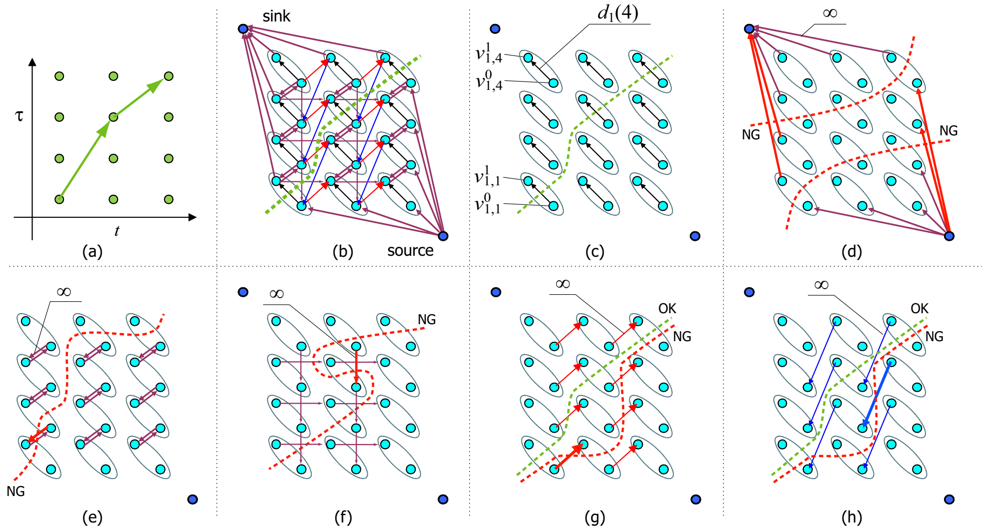

Figure 2. (a) DTW as a path optimization problem. (b) DTW as a graph cut problem. (c) Edges for $d_i(u_i)$. (d) Edges for boundary constraints. (e) and (f) Edges for excluding unexpected cuts. (g) Edges representing $\alpha_{t,t+1} = 1$. (h) Edges representing $\beta_{t,t+1} = 2$.

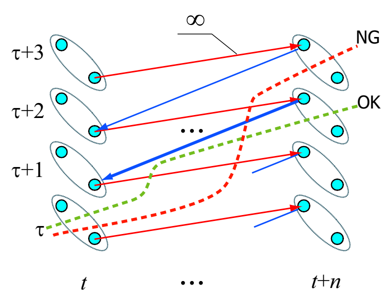

Figure 3. Non-Markovian constraint $\alpha_{t,t+n} = 1, \beta_{t,t+n} = 2$ as edges of G.

understand the minimum cut of (b) is equivalent to the optimal path of (a). In other words, the solution by some graph cut algorithms is exactly the same as the solution by DP. Most graph cut algorithms, such as Edmonds-Karp algorithm, require polynomial-order computations and are still efficient for Markovian DTW.

5. Introducing Non-Markovian Constraints

As described in Section 1, non-Markovian constraints restrict the relationship between u_t and u_{t+n} ($n > 1$). Figure 1 (d) shows an example of non-Markovian constraint. It forces a warping function taking the value τ at t to pass through one of four values as u_{t+n} at $t+n$. It is important that such a non-Markovian constraint [2] is written as $\alpha_{t,t+n} \leq u_{t+n} - u_t \leq \beta_{t,t+n}$. It is also important that this constraint does not restrict $u_{t+1}, \ldots, u_{t+n-1}$ directly.

By comparison between Fig. 1 (b) and (d), we can expect that non-Markovian constraints can regulate the warping function in more various ways. In fact, non-Markovian constraints are very flexible; for example, it is possible to put such a constraint between $t = 4$ and 9 but not between $t = 7$ and 11. In Section 6, we will introduce a learning method to design the non-Markovian constraints automatically for a specific DTW task.

Similar to the Markovian constraints, non-Markovian constraints are also introduced as edges of graph G. If we want to restrict the relationship between u_t and u_{t+n}, we need to put several edges with an infinite weight between nodes at t and $t + n$. Figure 3 shows a non-Markovian constraint $1 = \alpha_{t,t+n} \leq u_{t+n} - u_t \leq \beta_{t,t+n} = 2$.

By obtaining the minimum cut of the resulting graph, we can have the globally optimal solution of non-Markovian DTW. We do not need to make any special care for the graph cut algorithm; just like the Markovian case, the algorithm can provide the optimal solution for the graph with edges for non-Markovian constraints. An important fact is that the introduction of non-Markovian constraints affect computational complexity of DTW just linearly. Considering Footnote 1, this is a distinct advantage of the proposed method over the conventional DP-based implementation.

6. Learning Non-Markovian Constraints

The purpose of introducing non-Markovian constraints for sequential pattern recognition with DTW is generally to allow all possible warping functions be-

[2] In the extra page of our camera-ready version, we will show an extended version of non-Markovian constraints, which depend not only on t and n but also on τ.

Table 1. Recognition rate (%).

	Average	Class "6"	Class "1"
Markovian	88.7	83.3	73.3
Non-Markovian	89.4	86.3	74.7

tween patterns of the same class and to exclude the other warping functions for suppressing over-fitting by DTW. Based on this fact, the non-Markovian constraints can be simply learned as follows. First, the allowable warping functions U are collected by the conventional Markovian DTW between patterns of the class c. Then, for each timing pair t and $t+n$, the maximum and minimum values of $u_{t+n} - u_t$ are set as $\alpha_{t,t+n}, \beta_{t,t+n}$ for non-Markovian DTW of class c.

7. Experimental Results

To evaluate the effect of non-Markovian DTW, an online character recognition experiment was conducted using the Ethem Alpaydin Digit dataset which is comprised of online handwritten digit samples ("0"– "9"). Each sample was simply represented as a sequence of two-dimensional pen-tip coordinate. The length T was about 50. The dataset was decomposed into 700 training samples per class and 300 test samples (X) per class. From the training samples, 21 reference patterns (1$\sim$4 samples per class) were determined by a clustering method. Mahalanobis distance was used for the local cost $d_t(u_t)$. The cost $\min F$ was calculated between a test sample and each of the 21 reference patterns, and the class of the minimum cost reference pattern was considered as the recognition result. Note that for improving the recognition accuracy of the conventional Markovian DTW, a warp range limitation ($|t - u_t| \leq \delta$) was introduced with the optimized parameter δ.

Computational time is less than 5ms for each non-Markovian DTW. This proves the high efficiency of the proposed method, if we consider the fact that the DP-based implementation is computationally intractable even around $T = 10$.

Table 1 shows that non-Markovian DTW could provide higher recognition accuracy than conventional Markovian DTW. Figure 4 (a) shows two improved examples, "6" and "1", which were misrecognized as "8" and "9" respectively by the Markovian DTW. The warping function U by the Markovian DTW (b) shows a steep slope around its beginning part for matching patterns of different classes. This slope is allowed by the Markovian slope constraint (even with the warp range limitation) and then causes over-fitting. In contrast, this steep slope was not allowed in the non-Markovian constraints. As shown in (c), the warping function violates the non-Markovian constraints, for example,

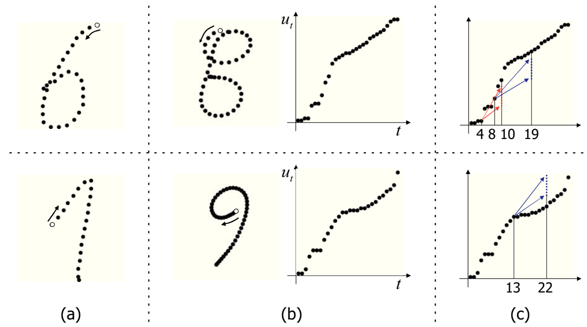

Figure 4. Two improved examples. (a) Test sample X. (b) Result by Markovian DTW. (c) Non-Markovian constraints.

$(\alpha_{4,10}, \beta_{4,10})$ and $(\alpha_{8,19}, \beta_{8,19})$ of class "8". Consequently, the steep slope, i.e., over-fitting, was suppressed in the non-Markovian DTW and these two test samples were correctly recognized.

8. Conclusion

A non-Markovian dynamic time warping (DTW) has been proposed. Use of graph cut, instead of dynamic programming (DP), as the optimizer of the warping function allows us to introduce non-Markovian constraints, which can restrict the relationship between points with a large time interval during DTW, still with polynomial computations. The usefulness of the non-Markovian DTW has been shown through an online character recognition experiment.

References

[1] U. Bertele and F. Brioschi, *Nonserial Dynamic Programming*, Academic Press, 1972.

[2] B. Price, B. Morse, and S. Cohen, "Geodesic Graph Cut for Interactive Image Segmentation," Proc. CVPR, 2010.

[3] F.R. Schmidt, E. Toppe, D. Cremers, and Y. Boykov, "Efficient Shape Matching via Graph Cuts," Proc. CVPR, 2007.

[4] F.R. Schmidt, E. Toppe, and D. Cremers, "Efficient Planar Graph Cuts with Applications in Computer Vision," Proc. CVPR, 2009.

[5] O. Duchenne, A. Joulin and J. Ponce, "A Graph-Matching Kernel for Object Categorization," Proc. ICCV, 2011.

[6] W. Chang and M. Zwicker, "Automatic Registration for Articulated Shapes," Proc. ESGP, 2008.

[7] S. Kumar and M. Hebert, "Discriminative Fields for Modeling Spatial Dependencies in Natural Images," Proc. NIPS, 2003.

[8] D. Cremers and L. Grady, "Statistical Priors for Efficient Combinatorial Optimization via Graph Cuts," Proc. ECCV, 2006.

[9] D. Anguelov, B. Taskar, V. Chatalbashev, D. Koller, D. Gupta, G. Heitz, and A. Ng, "Discriminative Learning of Markov Random Fields for Segmentation of 3D Scan Data," Proc. CVPR, 2005.

[10] H. Ishikawa and D. Geiger, "Occlusions, Discontinuities, and Epipolar Lines in Stereo," Proc. ECCV, 1998.