Motion Image Segmentation Using Global Criteria and DP

Takumi Kobayashi Fumito Yoshikawa Nobuyuki Otsu
National Institute of Advanced Industrial Science and Technology
Umezono 1-1-1, Tsukuba, Japan

Abstract

We propose methods for segmenting a motion sequence into motion primitives, taking into account temporal constraints (continuity along the time axis). In the proposed methods, dynamic programming (DP) is used on a motion feature sequence to allow for the effects of these constraints on the results of the segmentation. The methods do not require such a running window along the time axis, as is typical for the usual methods, and thus they can be applied to the segmentation of transient motions. The results of comparative experiments using several motion features and segmentation methods on weightlifting motion data demonstrate the effectiveness of the proposed methods.

1. Introduction

The analysis of human motion in a video sequence has become an important topic in computer vision. This task involves many computer vision problems, such as recognition [4] and segmentation [11]. Motion recognition in particular has been the primary concern of many researchers in the area. In the early years, studies for the recognition of gestures and other simple actions were carried out by applying specific object models in a carefully controlled environment for capturing images [10]. In recent years, many researchers have studied the recognition of more detailed motions, such as human gait [5], from natural motion images captured outdoors without assuming object models [4]. Note that only predefined motions are considered for the recognition task.

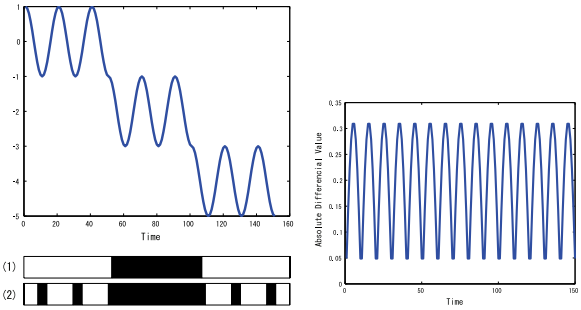
This paper addresses the problem of segmentation of motion sequences without any specific prior knowledge of types of motion or models. On the basis of the motion features, a sequence can be divided into several short subsequences (segments) for which all motions (restricted to such a subsequence) are identical in a sense and do not include sudden changes. Segmentation is useful for many applications, such as the creation of a summary of a long motion sequence to facilitate browsing, or the content-based compression of motion images. For sports science in particular, segmentation is effective for analyzing detailed motions. The motions of a sports player ordinarily consist of several phases to accomplish a specific task, such as lifting a heavy weight. These phases are required to be automatically segmented, in order to check the form of the players and to analyze any deficiencies. On the other hand, for video sequences, methods for cutting scenes have also been studied [2]. Those detect boundary points at which several kinds of camera break, i.e. video editing, occur. Scene cuts are different from the segmentation studied in this paper which is based on information contained in the motion itself, not on video editing.

Motion segmentation can be approached in either a top-down or bottom-up way [8]. In the top-down approach, human activities are segmented according to human semantics, where boundary points can differ from person to person. In the bottom-up approach, a motion sequence is segmented based only on the motion information itself without reference to top-down semantics. For example, "walking" is a top-down segmented activity consisting of the motion of two legs which can be analyzed in a bottom-up fashion. Top-down activities are composed of bottom-up motion primitives, and the particular way of combining motion primitives to define a top-down activity depends on a user and a task. In this paper, we focus on the bottom-up approach to segmentation.

There are several studies related to segmentation. Rui and Anandan [8] applied the frame difference method[1] or the AR model to motion features based on optical flow. In [11], statistical motion features based on spatio-temporal gradients were extracted and then clustered using the Spectral Clustering method [6]. Janus and Nakamura [3] employed HMM with motion capture data. Though these studies utilized a running window along the time axis, its effect on performance was hardly discussed in those papers. We now consider the toy example of the smooth one-dimensional sequence shown in Fig. 1. The result of frame difference is shown in Fig. 1(b). Since the method of frame difference focuses only on the local (adjacent) data, it can-

[1] Frame difference examines differences between adjacent features as well as frames.

978-1-4244-2154-1/08/$25.00



(a) (b)

Figure 1. Toy example: (a) one-dimensional sequence with segmentation results from the proposed method (1) and Spectral Clustering (2). Both black and white indicate segments (and so forth). (b) The result of frame difference.

not deal with such a continuous sequence which does not include any local sudden changes. Moreover, it is intrinsically difficult to determine a threshold for detecting change points. As shown in Fig. 1(a), the clustering method cannot be successfully applied because it utilizes individual features, disregarding the continuity of the time series and resulting in an identical cluster for the same motions at distinct time which should be different segments (Fig. 1(a-2)). If the method of a running time window with appropriate time length were applied, it might be successful, as in [11]. The appropriate time length, however, is difficult to determine without any prior knowledge about the motion. Besides the approach utilizing a running window would be effective for segmenting cyclic or repeated motions, but it would be likely to fail in the case of transient motion, since it blurs discrimination at boundary points.

In this paper, we propose methods for bottom-up segmentation of motion images. Under weak assumptions and not requiring specific models of motion, dynamic programming is employed to naturally incorporate the property of time sequencing, i.e. the continuity of time series. The methods optimize segmentations based on the global consistency of segments; this approach is different from one that employs local decisions, such as the frame difference method. In addition, the proposed methods do not require a window running along the time axis, as is typical in ordinary methods, and thus they are applicable to the segmentation of transient motion.

2. Proposed Method

Standard clustering methods used for motion image segmentation occasionally yield intermittent results, as shown in Fig. 1(a-2). In motion images, successive data samples are likely to form a segment and thus segment labels assigned to each sample should be smooth (consistent with respect to adjacency) along the time axis. In this paper, we apply dynamic programming (DP) in order to incorporate the continuity of time-series data.

Given motion feature vectors $\boldsymbol{X} = \{\boldsymbol{x}_t \in \boldsymbol{R}^D | t = 1, \cdots, L\}$, segmentation can be formulated in terms of the following optimization problem:

$$\min_{b_0,\cdots,b_N} E(L,N) = \sum_{i=0}^{N-1} G(b_i, b_{i+1})$$
$$\text{s.t. } b_i \in \{0,\cdots,L\}, \quad b_0 = 0, \quad b_i < b_{i+1}, \quad b_N = L, \quad (1)$$

where N is the number of segments, i-th segment ranges over $[b_{i-1}{+}1, b_i]$, $G(a,b)$ is the cost of the segment $[a{+}1, b]$, and $E(l,n)$ is the total cost of n segments over the interval $[1,l]$. Eq.(1) can be solved by dynamic programming (DP) using the following recurrence formula:

$$E(l,n) = \min_{k<l} E(k,n-1) + G(k,l). \quad (2)$$

DP has several advantages for segmenting motion images: Firstly, by defining segment-wise costs, temporal continuity is naturally incorporated. The method can avoid intermittent segmentation and possibly yield a smoothed result. Secondly, the constraint on segment lengths can be easily incorporated into the optimization problem (Eq.(1)) by

$$G(k,l) = \infty \quad \text{if } |l{-}k| < L_{min} \text{ or } |l{-}k| > L_{max}, \quad (3)$$

where L_{min}, L_{max} are minimum and maximum lengths of segments, respectively. This is practically useful and reduces the computational cost of Eq.(1) from $O(NL)$ to $O(N(L_{max} - L_{min}))$. Thirdly, the method does not require a running time window along time axis, unlike other methods [11, 3]. As mentioned before, a running window smoothes out discontinuities at boundary points and besides determination of the window length is intrinsically difficult. Fourthly, in contrast to the frame difference method, segmentation by DP is based on the *global* consistency, which is helpful for robustness of the method with respect to noise.

We propose three versions of the method by defining the segment cost G in three ways: two are based on statistics of data and the third is based on affinities between data pairs.

2.1. Statistics Based Cost

The segment cost is defined on the basis of statistics of motion feature vectors within the segments. It is expected that such vectors have small variance within each segment. From this viewpoint, we can employ the Fisher discriminant criterion: the variance within segments, σ_W^2, is to be minimized, while that between segments, σ_B^2, is to be maximized. Since the total variance σ_T^2 is constant, minimization of σ_W^2 is equivalent to maximization of σ_B^2 ($\sigma_T^2 = \sigma_W^2 + \sigma_B^2 = \text{const}$). The variance between segments is described as follows:

$$\sigma_B^2 = \sum_{i=1}^{N} p_i ||\boldsymbol{\mu}_i - \boldsymbol{\mu}_T||^2 = \sum_{i=1}^{N} p_i ||\boldsymbol{\mu}_i||^2 - ||\boldsymbol{\mu}_T||^2,$$

where $\boldsymbol{\mu}_T, \boldsymbol{\mu}_i$ are the mean vectors for the whole segment and the i-th segment, respectively, and p_i is the probability of the i-th segment. Since $||\boldsymbol{\mu}_T||^2$ is given and constant, the segment cost is defined as

$$G(a,b) = -p_{[a+1,b]} \Big|\Big| \sum_{j=a+1}^{b} p_{[a+1,b]}(\boldsymbol{x}_j)\boldsymbol{x}_j \Big|\Big|^2, \quad (4)$$

where $p_{[a+1,b]}, p_{[a+1,b]}(\boldsymbol{x})$ are the probability of the segment $[a+1,b]$ and the probability that the sample $\boldsymbol{x}$ is in the segment, respectively. We call this **Fisher DP (F-DP)**, a natural extension of [7] to motion feature vectors.

On the other hand, in the case that the sample vectors consistently, e.g. linearly, vary throughout the segment, simply minimizing the variance within segments is not sufficient to deal with such the variation. Therefore, the segment cost is considered to be defined by the least square error in the linear approximation of the sample vectors. The linear approximation and segment cost are given by

$$\boldsymbol{x}_t = \boldsymbol{m}t + \boldsymbol{c} = [\boldsymbol{m}, \boldsymbol{c}][t,1]' = \boldsymbol{A}[t,1]', t \in [a+1,b]$$
$$\boldsymbol{A} = \Big(\sum_{t=a+1}^{b} p(\boldsymbol{x}_t)\boldsymbol{x}_t[t,1] \Big) \Big(\sum_{t=a+1}^{b} p(\boldsymbol{x}_t)[t,1]'[t,1] \Big)^{-1}$$
$$G(a,b) = \sum_{t=a+1}^{b} p(\boldsymbol{x}_t)||\boldsymbol{x}_t - \boldsymbol{A}[t,1]'||^2, \quad (5)$$

where $p(\boldsymbol{x})$ is the sample probability. We call this **Linear DP (L-DP)**. This has more degrees of freedom than F-DP which is the special case of L-DP ($\boldsymbol{m} = 0$). We simply apply a linear approximation in order to avoid over fitting by a polynomial approximation.

In the two proposed methods above, sample probabilities are taken into account. These are interpreted as sample weights $w_t (\sum_t w_t = 1)$ in the time series:

$$p(\boldsymbol{x}_t) = w_t, p_{[a+1,b]} = \sum_{t=a+1}^{b} w_t, p_{[a+1,b]}(\boldsymbol{x}_t) = \frac{w_t}{p_{[a+1,b]}}.$$

Usually the samples are equally weighted ($w_t = 1/L$). In the case that we have prior knowledge about the segmentation, such as *key frames*, the weights at those frames could be more highly assigned.

The only parameter in DP is the number of segments, N. Through the process of calculating Eq.(2), the results of $\forall n \leq N$ segments are also obtained. Thus, it is sufficient to obtain results for a large number N and then automatically select the optimal one from those results. We propose the measure for evaluating segmentation results by focusing on differences between adjacent pair-wise segments. In F-DP, adjacent segments are evaluated by the pair-wise Fisher criterion:

$$\eta_F(n) = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{\sigma_B^{i,i+1}}{\sigma_W^{i,i+1}}.$$

L-DP utilizes the differences in the approximated linear forms between adjacent segments:

$$\eta_L(n) = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{||\boldsymbol{\mu}_{i+1} - \boldsymbol{\mu}_i||^2 + ||\boldsymbol{m}_{i+1} - \boldsymbol{m}_i||^2}{\epsilon_i^2 + \epsilon_{i+1}^2},$$

where ϵ_i^2 is the least square error in the segment i. These measures penalize the cuts between flat segments. The segmentation result of the highest score is considered to be optimal ($n^* = \arg\max_{2 \leq n \leq N} \eta(n)$).

2.2. Affinity Based Cost

Affinities (or similarities) between samples have been utilized in Spectral Clustering [6] and Affinity Propagation [1]. The segment cost is defined based on these affinities. Similarly to minimization of σ_W^2 in F-DP, the samples in each segment are expected to have strong affinities each other. The segment cost is defined as

$$G(a,b) = -\sum_{t=a+1}^{b} s_{i^*t}, \quad i^* = \arg\max_{a<i\leq b} \sum_{t=a+1}^{b} s_{it}, \quad (6)$$

where s_{it} is the affinity between samples i and t. G is the summation of affinities between all samples t and the representative sample i^* in the segment. We call this **Affinity DP (A-DP)**. It is the same formulation as Affinity Propagation (AP) [1]. AP optimizes the cost by Belief Propagation and tends to be trapped in local minima, whereas a global minimum is obtained by DP in this paper. The evaluation measure for determining the optimal number of segments is directly given by the (negative) total cost ($\eta_A(n) = -E(L,n)$). Note that the computational cost for calculating all G defined in Eq.(6) is $O(L^3)$ while that of F/L-DP in Eq.(4,5) is $O(L^2)$.

Here, self-affinities s_{tt} play different roles from the other affinities between samples. The bigger the self-affinity is, the more likely the sample becomes the representative. This can be interpreted as the sample weight (importance), which could be larger at key frames. The total cost can then be decomposed as follows:

$$E(L,n) = -\sum_{i=1}^{n} \sum_{t=b_i+1}^{b_i+1} s_{i^*t} = -\sum_{i=1}^{n} \sum_{t=b_i+1, t\neq i^*}^{b_i+1} s_{i^*t} - \sum_{i=1}^{n} s_{i^*i^*}.$$

The second term on the right hand side is the sum of self-affinities at the representative samples. Thus, the sum and/or the scale of self-affinities may affect the total cost (or η_A), i.e. the optimal number of segments. The self-similarities are usually set to 0.

3. Experiment

The proposed methods are applied to segmentation of motion image sequences for sports motion analysis. The

performances are compared with those of the other segmentation methods by using several motion features.

3.1. Motion Features

We employ two kinds of motion features: Cubic Higher-order Local Auto-Correlation (CHLAC) [5] and the gradient histogram (GH) [11]. These view-based motion features are derived from the statistics of object motions without any prior knowledge. CHLAC is a histogram of local configuration patterns (auto-correlations) of moving points found by frame difference. It extracts features not only of velocity but also of acceleration, and it has been successfully applied to motion recognition [5]. GH is a concatenated histogram of absolute component values of normalized spatio-temporal gradients. It captures only the velocity without respect to its directional sign. Thus, CHLAC may extract more detailed motion features than GH.

For comparison, the minimum time length T of a running window is set to 4 frames in both CHLAC and GH. In order to calculate affinities in A-DP and Spectral Clustering, Euclidean distance is used in CHLAC whereas the χ^2 distance [11] is used in GH.

3.2. Segmentation methods

We apply two other segmentation methods for comparison. One is a linear filter (LF) method including frame difference and the AR model. Suppose we have time windows of length T. Then the linear filter $\boldsymbol{a} \in \boldsymbol{R}^T$ is calculated by

$$\min_{\boldsymbol{a}} \sum_t ||\boldsymbol{X}_t \boldsymbol{a}||^2, \quad \text{s.t.} ||\boldsymbol{a}|| = 1 \therefore \sum_t \boldsymbol{X}_t' \boldsymbol{X}_t \boldsymbol{a} = \lambda_{min} \boldsymbol{a},$$

where $\boldsymbol{X}_t = [\boldsymbol{x}_t, \cdots, \boldsymbol{x}_{t+T-1}]$. The filter $\boldsymbol{a}$ is determined so as to possibly decrease the norm of output vectors $\boldsymbol{X}_t \boldsymbol{a}$. It is the eigenvector associated with the minimum eigenvalue. The boundary point is detected by thresholding the filter response values $||\boldsymbol{X}_t \boldsymbol{a}||$. The appropriate threshold value is determined by hand in this experiment.

The other is the Spectral Clustering (SC) method [6], successfully applied with GH motion features in [11] for clustering of cyclic motions. The motion features are calculated from the frames within the running time windows. Segments are detected by clustering the motion features into k clusters, say $k = 10$ in the experiment.

The proposed methods are applied with the maximum number of segments, $N = 20$. Then, the segmentation results with the highest evaluation score are shown.

3.3. Motion Images

The subject for this experiment is weightlifting. It is a sport involving dynamic motion using full body kinematics, and athletes' motions transit through several phases to finally lift up the heavy barbell [9]. Since these phases are difficult to segment even by hand, automatic segmentation of transient motion is a challenging task but quite useful for analyzing the motion to improve athletes' performances. On the other hand, the segmentation of motion sequences has been traditionally applied to cyclic motions, such as repeated gestures, and the boundary points have been easily determined. We captured motion images of two kinds of weightlifting, the "clean and jerk" and the "snatch," as performed by members of the Japanese National Women Weightlifting team. It should be noted that the environment for capturing images is an actual practice scene at a training camp, not a carefully controlled environment designed for this study, and thus background noise derived from other moving persons is found in the images.

3.4. Segmentation Result

Since we do not have definitive prior knowledge about the boundary points, the segmentation results are evaluated *a posteriori* in terms of their appearances. For the "clean and jerk" motion sequence, the combination of F-DP and CHLAC yields quite a favorable result, as shown in Fig. 2. The important phases, *clean*, *jerk* and *hold*, are detected. In the *clean* phase, it is noteworthy that the *first and second pulls* (2,3)[2] are separately detected, even though the boundary point between these pulls, i.e. the moment of maximum acceleration of the pull motion, is not apparent. In the *jerk* phase, the making ready motion (*crouch* (12)) is detected before the *split jerk* (13) segment. The resulting segments may appear to be too detailed; for example, *first and second stand up* (6,7) and *jump up and down* (8,9) might be merged into *stand up* and *jump*, respectively, by human observers. The proposed method, however, detects bottom-up segments that are motion primitives. How to merge them depends on the application tasks or users, which is out of the scope of this paper. By comparing to this most favorable result, which is shown with the caption of "optimal" in all figures, all segmentation methods and motion features are comparatively studied below.

When applying GH features with running windows of various time lengths T, the segmentation results are shown in Fig. 3. Linear Filter (LF) hardly distinguishes the *clean* phase in which the important motions of *first, second pull* are performed (Fig. 3(a)). For Spectral Clustering (SC) in Fig. 3(b), the sequence is evenly over-segmented and the boundary points seem less meaningful even when a larger time interval T is taken. As shown in Fig. 3(c), with F-DP, the *clean* phase is detected but the other segments are not so meaningful. The result of L-DP is similar to that of F-DP except that the *clean* phase is not detected. On the other hand, A-DP produces a somewhat similar segmentation result to the most favorable one. This is due to a difference

[2] For the Italic name and the numbers in parentheses, refer to the caption of Fig. 2.

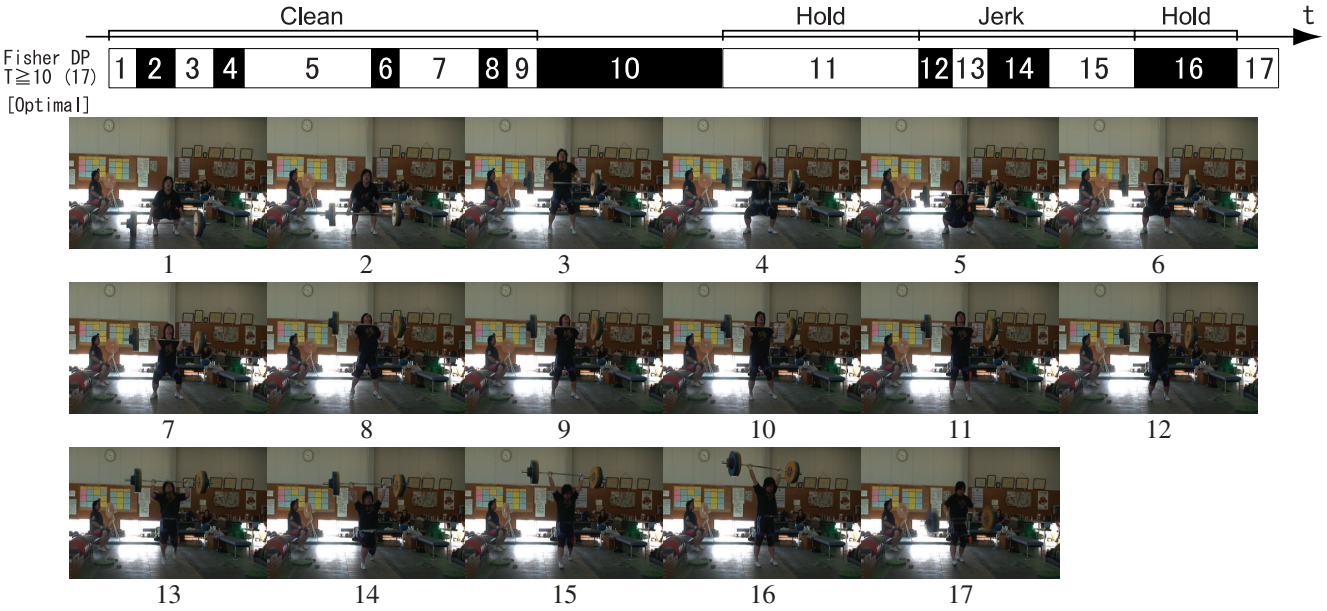
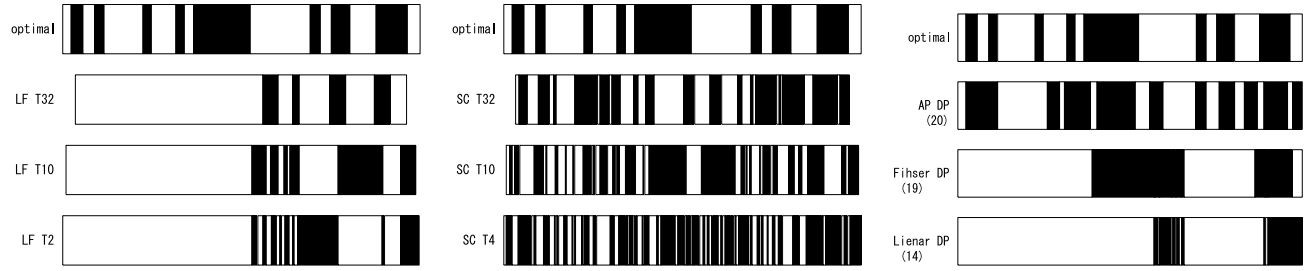


Figure 2. The most favorable segmentation result for the "clean and jerk" obtained by F-DP with $L_{min}=10$, $N=20$ and CHLAC features where 17 motion primitives are detected. The numbers indicate segmented motion primitives: (1) *lift off* (2) *first pull* (3) *second pull* (4) *drop under the bar to catch* (5) *small bounce* (6) *first stand up (fast)* (7) *second stand up (slow)* (8) *jump up* (9) *jump down* (10) *recover balance* (11) *hold* (12) *crouch* (13) *split jerk* (14) *catch the barbell over the head* (15) *lift up* (16) *hold* (17) *throw away the barbell*.

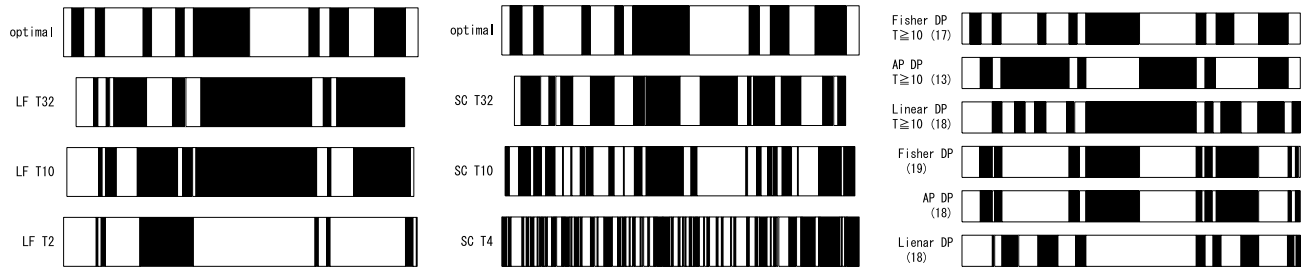


(a) Linear Filter (b) Spectral Clustering (c) Proposed Method

Figure 3. The segmentation results when GH features are applied. In (a), (b), T denotes the time length of the running window. In (c), the numbers in parentheses are optimal numbers of segments which are obtained based on the evaluation criteria (η_F, η_L, η_A).

of the metric for features; namely, A-DP makes use of the χ^2 distance whereas F/L-DP utilizes Euclidean distances. A-DP can be seen to be effective compared to SC which also makes use of the χ^2 distance. Totally, favorable results are not obtained by using GH features. Since GH features describe motion characteristics in a simple way without a directional sign, it is difficult to capture detailed characteristics of weightlifting motion which is based mainly on vertical movements only.

In the case of CHLAC features, the results of all methods are improved (Fig. 4). LF with $T=32$ (Fig. 4(a)) roughly segments the *clean* phase and detects both the end of the *second pull* (3) and the start of *first stand up* (6). The points detected, however, are not quite correct due to smoothing of the sequence arising from running windows. In Fig. 4(b), some of the boundary points from SC with T=32 are somewhat close to those in the most favorable result, though some meaningless boundaries still remain due to over segmentation. As shown in the bottom three results of Fig. 4(c), the proposed methods detect the key phases, *clean*, *jerk*, and *hold*. However, they also detect some segments that are too short, especially within *throw away* (17) which contains extreme motions of the thrown barbell. We limit the segment length to L_{min}=10 in Eq.(3), so as to detect somewhat long and meaningful segments. Then the results are improved and all segments are meaningful as shown in the top three results of Fig. 4(c). The result of A-DP is favorable and quite similar to that of F-DP (the most favorable one) except that a few segments are missed: the *first pull* (2) and the *stand up* (6,7). Note that the segmentation result of A-DP with the fourth score (16 segments) is almost the same as the most favorable result. On the other hand, L-DP yields a slightly inferior result: the *second pull* (3) is missed and the cycle within the *small bounce* (5) is overly segmented.



(a) Linear Filter (b) Spectral Clustering (c) Proposed Method

Figure 4. The results when CHLAC features are applied. In the top three results of (c), the constraint of $L_{min} = 10$ in Eq.(3) is applied.

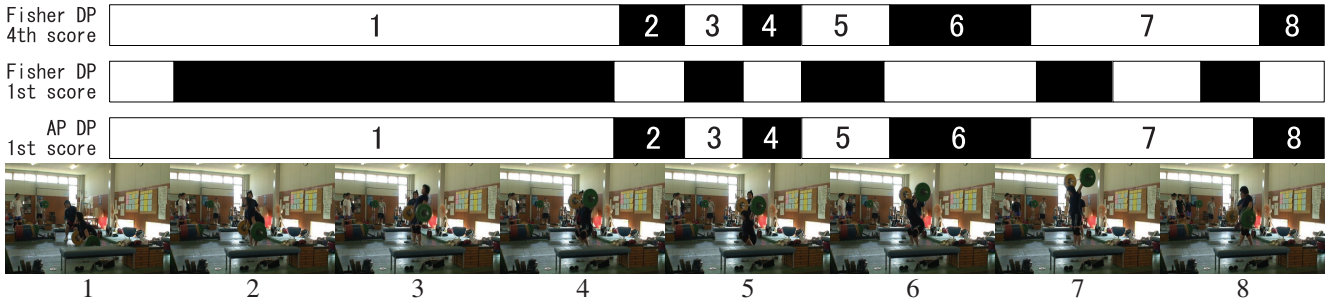


Figure 5. The segmentation results for the "snatch": (1) *lift off* (2) *first pull* (3) *second pull* (4) *drop under the bar to catch* (5) *small bounce* (6) *stand and lift up* (7) *hold* (8) *throw away the barbell*. The motion primitives are similar to those in "clean and jerk" (Fig. 2).

Fig. 5 shows the experimental results for the "snatch" weightlifting motion using CHLAC and the proposed methods. The "snatch" consists of motion phases similar to those for the "clean and jerk" and the segmented motion primitives in Fig. 5 are also similar to those in Fig. 2. The motion images contain more noise caused by background persons moving around. In F-DP, the result of the first score is slightly affected by the background noise, whereas that of the fourth score yields the most favorable result. A-DP of the first score also produces the most favorable result.

In summary, the combination of F/A-DP and CHLAC is the best; in terms of computational cost, F-DP may be better than A-DP. It is found that the *simple* segmentation method, which has few degrees of freedom, is suitable for the *detailed* motion features extracted by CHLAC. F/A-DP imposes the simple assumption, whereas L-DP has more degrees of freedom, producing unfavorable results.

4. Conclusion

We have proposed methods for segmenting motion images by using dynamic programming. The methods can incorporate temporal continuity and optimize the segmentation based on the global consistency of segments. In addition, the methods do not require a running window along the time axis. The proposed methods with CHLAC motion features were applied to two kinds of weightlifting motion sequences, the "clean and jerk" and the "snatch", and produced favorable results compared to other methods.

References

[1] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007. 3

[2] U. Gargi, R. Kasturi, and S. Antani. Performance characterization and comparison of video indexing algorithms. In *CVPR*, 1998. 1

[3] B. Janus and Y. Nakamura. Unsupervised probabilistic segmentation of motion data for mimesis modeling. In *International Conference on Advanced Robotics*, 2005. 1, 2

[4] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, 2007. 1

[5] T. Kobayashi and N. Otsu. A three-way auto-correlation based approach to human identification by gait. In *IEEE Workshop on Visual Surveillance*, 2006. 1, 4

[6] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithms. *Advances in Neural Information Processing Systems*, 14, 2001. 1, 3, 4

[7] N. Otsu. Discriminant and least squares threshold selection. In *ICPR*, 1978. 3

[8] Y. Rui and P. Anandan. Segmenting visual actions based on spatio-temporal motion patterns. In *CVPR*, 2000. 1

[9] B. K. Schilling, M. H. Stone, H. S. O'bryant, A. C. Fry, R. H. Coglianese, and K. C. Pierce. Snatch technique of collegiate national level weightlifters. *Journal of Strength and Conditioning Research*, 16(4):551–555, 2002. 4

[10] A. D. Wilson and A. F. Bobick. Parametric hidden markov models for gesture recognition. *Pattern Analysis and Machine Intelligence*, 21:884–900, 1999. 1

[11] L. Zelnik-Manor and M. Irani. Event-based analysis of video. In *CVPR*, 2001. 1, 2, 4