Detection and classification of repetitious human motions combining shift variant and invariant features

Ichiro Ide[*], Taku Kuhara[*§], Daisuke Deguchi[†*] Tomokazu Takahashi[‡] and Hiroshi Murase[*]
[*] Graduate School of Information Science, Nagoya University
1 Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
E-mail: ide@is.nagoya-u.ac.jp, tkuhara@murase.m.is.nagoya-u.ac.jp, murase@is.nagoya-u.ac.jp
[†] Information and Communication Headquarters, Nagoya University
1 Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
E-mail: ddeguchi@nagoya-u.jp
[‡] Department of Economics and Information, Gifu Shotoku Gakuen University
1-38 Naka-Uzura, Gifu 500-8288, Japan
E-mail: ttakahashi@gifu.shotoku.ac.jp
[§] Currently at TOYOTA COMMUNICATION SYSTEMS Co., Ltd.

***Abstract*— Detection and classification of significant human motions are important tasks when analyzing a video that records human activities. Among various human motions, we consider that repetitious motions are especially important since they are usually results of activities with clear intentions. In this paper, we propose and evaluate a method that detects video segments that contain repetitious motions, which is robust to motion shift. Experimental results showed the effectiveness of the proposed method compared to conventional methods. In addition, we report a preliminary result of an experiment on the classification of the types of the detected repetitious motions.**

I. Introduction

Detection and classification of significant human motions are important tasks when analyzing a video that records human activities. Among various human motions, we consider that repetitious motions are especially important since they are usually results of activities with clear intentions.

In this paper, we first propose a method that detects video segments that contain repetitious motions in Section II, which is robust to motion shift. Next, we report a preliminary result of an experiment on the classification of the types of the detected repetitious motions in Section III. In the end, Section IV concludes the paper.

II. Detection of video segments that contain repetitious motions

A. Overview

In this section, we propose a method that detects repetitious motions caused by human activities. Unlike repetitious motions caused by machines, those caused by human activities are usually not precisely periodic. Thus it is difficult to apply a simple frequency analysis method to detect them.

Considering this, Hamada et al. proposed a method that detects repetitious motions in TV cook shows [1]. The method detected repetitious motions by temporal frequency analysis of local areas in a video. It was then applied to detect important cooking operations in order to summarize a cook show [2]. However, it had a problem that it cannot correctly detect repetitious motions that shift, such as the one shown in Fig. 1.

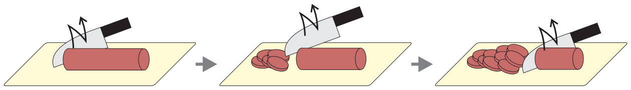


Fig. 1. Example of a repetitious motion that shifts.

Meanwhile, Doman et al. proposed a method that detects repetitious motions by analyzing the temporal change of the global feature in a video frame [3]. This approach could cope with shifting motions to some extent, but it could not detect motions occurring only in a small region.

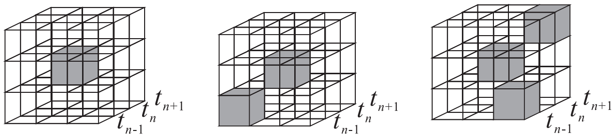
Considering these problems, in this paper, we propose a method that detects repetitious motions even if they shift and/or occur in a small region. This is realized by combining two motion features; binary feature and CHLAC feature, where the former is variant and the latter is invariant to the location where the motion is occurring. By this, both repetitious motions that do and do not shift can be detected in one framework. Details of each feature and the framework follows.

B. Shift variant and invariant features

Two features; shift variant binary feature and shift invariant CHLAC feature, are used in the proposed method. Note that in order to handle the features as one vector per frame, both of them were raster scanned within each frame.

1) Binary feature: The binary feature is simply the raw bit pattern of pixels in a frame. In a given video with a length of N frames, a binary feature vector $\boldsymbol{b}_n$ $(n = 0, ..., N - 1)$

978-0-7695-4791-6/12 $26.00 © 2012 IEEE
DOI 10.1109/EST.2012.7



(a) Mask type 1 (b) Mask type 2 (c) Mask type 3

Fig. 2. Example of spatio-temporal masks for CHLAC feature extraction.

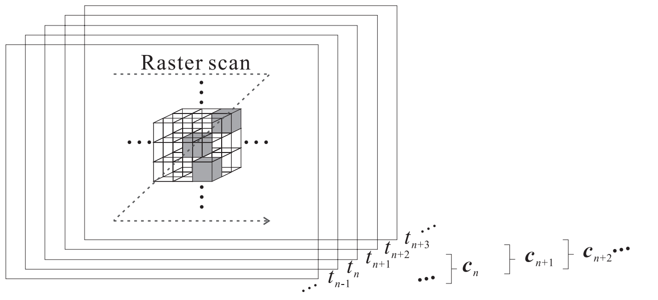


Fig. 3. CHLAC feature extraction from a video clip.

is extracted from each frame, and scanned frame-by-frame through the video clip.

2) CHLAC feature: The CHLAC (Cubic Higher-order Local Auto Correlation) feature is a spatio-temporal feature that could compactly represent the shape and the motion of regions that appear in a video, proposed by Otsu [4]. Since it is invariant to the location of the motion region, we considered that it could be used to analyze the periodicity of a shifting motion. As shown in Fig. 2, the CHLAC feature focuses on the bit patterns of three consecutive video frames, where the number of possible patterns is 251 (Type 1: 1 + Type 2: 13 + Type 3: 237), considering symmetry. Thus, for each frame, a 251-bin histogram is obtained which represents the spatio-temporal bit pattern at the moment.

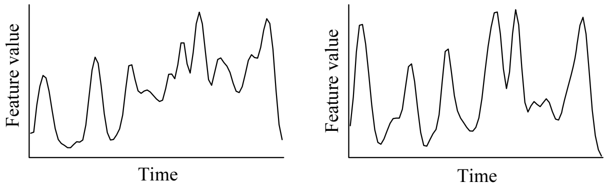
As shown in Fig. 3, for each three consecutive frames t_{n-1}, t_n, t_{n+1} $(n = 1, ..., N - 2)$ in a given video with a length of N frames, a CHLAC feature vector $\boldsymbol{c}_n$ is extracted, and scanned frame-by-frame through the video clip.

Figure 4 shows feature values of an actual shifting motion by the two features. We can see that the CHLAC feature shows a clearer cyclic pattern than the binary feature.

C. Detection process

Figure 5 shows the process flow of the proposed framework. An input video is parsed by a temporal window with a length of N. The following steps are applied to these video clips.

1) Feature extraction and analysis of periodicity: As a pre-process, for each frame in a given video, inter-frame differential from the previous frame is measured in order to extract pixels with motion, and then binarized according to a threshold θ_d in order to reduce noise. Both binary and CHLAC features are obtained from each binarized inter-frame differential patterns.



(a) Binary feature (b) CHLAC feature

Fig. 4. Example of feature values for a shifting repetitious motion.

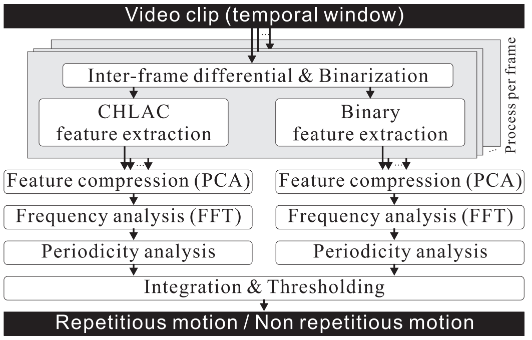

Video clip (temporal window)
Inter-frame differential & Binarization
CHLAC feature extraction
Binary feature extraction
Process per frame
Feature compression (PCA)
Feature compression (PCA)
Frequency analysis (FFT)
Frequency analysis (FFT)
Periodicity analysis
Periodicity analysis
Integration & Thresholding
Repetitious motion / Non repetitious motion

Fig. 5. Process flow of the repetitious motion detection method.

First, in order to reduce noise, PCA (Principal Component Analysis) is applied to all the feature vectors obtained in a video clip. Hereafter, the features are represented in a reduced dimension feature space by using only the top d_b and d_c principal components as bases for binary and CHLAC features, respectively.

Next, in order to obtain a power spectrum, FFT (Fast Fourier Transform) is applied to the feature vectors. Then, the power spectrum is normalized so that it should be invariant to the size of the motion region.

Based on the features proposed by Hamada et al. [1], the following five features that represent the periodicity of the motion are calculated from the power spectrum. The features are defined based on the assumption that in the case a periodic motion exists, there should be a clear and strong peak at $f_0 \leq f_p < f_1$, where $[f_0, f_1]$ is a frequency range that repetitious motions by human beings are expected.

- **Overall power** (F_{Power})
  The hatched area shown in Fig. 6; $F_{\text{Power}} = \sum_{f=f_0}^{f_1} P(f)$.
- **Prominence of the peak** (F_{Prom})
  The ratio of the peak power to the average powers in the hatched area excluding the peak itself; $F_{\text{Prom}} = (f_1 - f_0 - 1)P(f_p) / \sum_{f=f_0, f \neq f_p}^{f_1 - 1} P(f)$.
- **Sharpness of the peak** (F_{Sharp})
  The ratio of the peak power against the powers of its neighboring frequencies; $P_{\text{Sharp}} = 4P(f_p) / \sum_{f=f_p-2, f \neq f_p}^{f_p+2} P(f)$.
- **Relative powers of the peak** (F_{R1} and F_{R2})

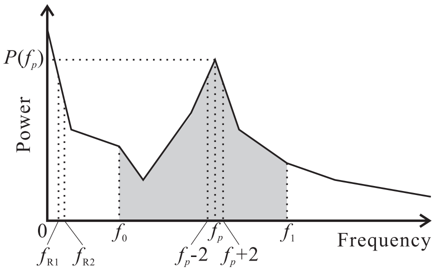


Fig. 6. Feature extraction from the power spectrum.

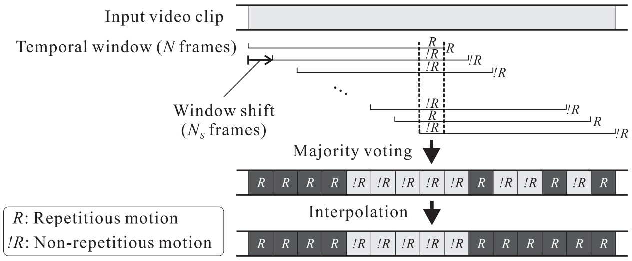

Input video clip
Temporal window (N frames)
Window shift
(N_s frames)
R
!R
Majority voting
R R R R !R !R !R !R !R R !R !R R !R R
Interpolation
R R R R !R !R !R !R !R R R R R R R
R: Repetitious motion
$!R$: Non-repetitious motion

Fig. 7. Detection of video segments that contain repetitious motions.

Ratios of the peak power to powers in very low frequencies; $F_{\mathrm{R1}} = F(f_p)/F(f_{\mathrm{R1}})$ and $F_{\mathrm{R2}} = F(f_p)/F(f_{\mathrm{R2}})$.

The overall periodicity is defined as

$$\begin{aligned} F &= \alpha_{\mathrm{Power}} F_{\mathrm{Power}} + \alpha_{\mathrm{Prom}} F_{\mathrm{Prom}} + \alpha_{\mathrm{Sharp}} F_{\mathrm{Sharp}} \\ &\quad + \alpha_{\mathrm{R1}} F_{\mathrm{R1}} + \alpha_{\mathrm{R2}} F_{\mathrm{R2}}. \end{aligned}$$

This is calculated from the power spectrum obtained from both the binary and the CHLAC features. Thresholding is applied to the sum of the overall periodicity for each of the features to detect if the temporal window contains repetitious motions or not.

2) Detection of video segments that contain repetitious motions: The temporal windows where the above process are applied, are shifted by N_S frames. Hence, as shown in Fig. 7, each segment with a length of N_S frames will have multiple results. Here, we take a majority vote approach to decide the final result for each such segment. In order to ignore short halts or noises, gaps of non-repetitious motions shorter than N_I frames are interpolated as repetitious motions.

D. Experiment

In order to evaluate the ability of the proposed method, we evaluated the use of each feature (Binary, CHLAC, and Binary & CHLAC) and also compared them with the conventional methods by Hamada et al. [1] and Doman et al. [3], through an experiment on actual repetitious motions that appeared in cook shows.

1) Conditions: The specification of the data set is shown in Table I, and the parameters used in the experiment are shown in Table II. Most of the parameters were decided intuitively according to preliminary experiments. From the data set, 323 shots with close-ups of hand motions, namely “hand shots”, were manually extracted. Their lengths were from 2 to 70 seconds, and on average, 17 seconds long.

TABLE I
SPECIFICATION OF THE DATA SET USED IN THE EXPERIMENT.

Cook show	NHK “Today’s cooking”
Period	November – December, 2009
Frame rate	29.97 frames/second
Resolution	1,400 × 1,080 pixels
	(Resized to 40 × 30 pixels)
Video clips	10 shows
Duration	250 minutes

TABLE II
PARAMETERS USED IN THE EXPERIMENT.

Parameter	Value	Parameter	Value
N	80 frames	α_{Power}	10.0
θ_d	30	α_{Prom}	1.0
d_b	7 dimensions	α_{Sharp}	1.0
d_c	4 dimensions	α_{R1}	1.0
f_0	3.00 Hz	α_{R2}	1.0
f_1	7.50 Hz	N_S	10 frames
f_{R1}	0.375 Hz	N_I	40 frames
f_{R2}	0.750 Hz		

We evaluated the accuracy of the repetitious motion detection by considering that a result was correct when both the starting and ending frames were within 30 frames from the ground-truth segment given manually.

2) Result: Table III shows the result of the detection accuracy for each method. We can see that the proposed method (Binary & CHLAC) showed a significantly higher accuracy than the conventional methods. We observed that the false-detection by the proposed method occurred mostly by the effect of camera-works.

Next, comparing the features used in the proposed framework, each feature used alone did not show a significant difference. However, based on our observation, the longer a repetitious motion shifts, the better the CHLAC feature performed. In total, the proposed method using both the binary and the CHLAC features showed higher accuracy than those using each feature individually. This shows the effect of combining shift variant and invariant features.

III. CLASSIFICATION OF REPETITIOUS MOTIONS TYPES

A. Overview

It is often the case that after the detection of repetitious motions, we would like to classify them into multiple types. The classes are task dependent. Since the data set used in Section II was from a cook show, here, we assume a cooking task, and try to detect repetitious motions peculiar to cooking activities; cut, mix, add, knead, and others.

B. Classification process

Figure 8 shows the flow of the classification process; It is composed of the training and the classification stages.

TABLE III

EVALUATION OF THE DETECTION ACCURACY OF VIDEO SEGMENTS THAT CONTAIN REPETITIVE MOTIONS.

Method (Features)	Recall	Precision	F-measure
Proposed (Binary & CHLAC)	0.77	0.79	0.78
Proposed (CHLAC)	0.73	0.63	0.68
Proposed (Binary)	0.73	0.60	0.65
Conventional (Local) [1]	0.52	0.42	0.47
Conventional (Global) [3]	0.59	0.54	0.56

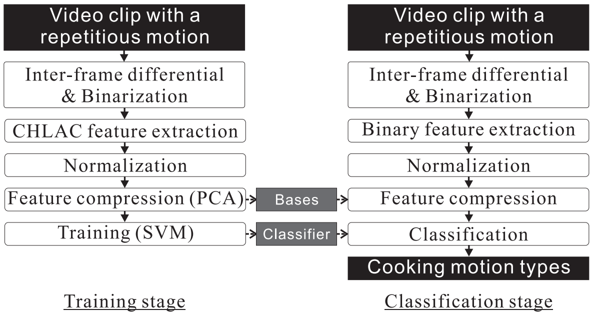

Video clip with a repetitious motion
Inter-frame differential & Binarization
CHLAC feature extraction
Normalization
Feature compression (PCA)
Training (SVM)
Bases
Classifier
Video clip with a repetitious motion
Inter-frame differential & Binarization
Binary feature extraction
Normalization
Feature compression
Classification
Cooking motion types
Training stage
Classification stage

Fig. 8. Process flow of the repetitious motion types classification method.

For both stages, the feature extraction process from each frame is different than that described in II-C.1; Instead of obtaining a histogram of spatio-temporal bit-patterns from each pair of three consecutive frames, here, one histogram is obtained that represents the overall spatio-temporal bit-patterns of a repetitious motion segment, by raster-scanning all frames in the segment.

In the training stage, first, PCA is applied to the feature vectors from training data, in order to reduce noise. Hereafter, the features are represented in a reduced dimension feature space by using only the top d_{class} principal components as bases. Next, the feature vectors are trained using a Support Vector Machine (SVM)[1] with manually tagged classes.

In the classification stage, first the feature vectors are compressed according to the bases defined in the training stage, and then classified according to the SVM constructed also in the training stage.

C. Experiment

1) Conditions: For the training data, 60 video clips manually extracted from hand-shots with repetitious motions were used. For the classification, 199 video clips manually extracted from the 323 hand-shots with repetitious motions used in II-D were used. The parameter d_{class} was set to 15 so that the accumulative contribution of the top d_{class} bases should be 99% in the training stage.

2) Result: Table IV shows the result of the classification in the form of a confusion matrix. In total, the classification accuracy was relatively good. However, the accuracy varied according to the motion types. In the cases of "cut" and "mix", since their motions were characteristic, they could be classified well. In the case of "add", since most of the motions tended to appear in the center of the frame, they were often confused with "others", and resulted in a relatively low accuracy. In the case of "others", some motions that did not appear in the training data could not be classified correctly.

TABLE IV

CONFUSION MATRIX OF THE CLASSIFICATION RESULT OF THE REPETITIVE MOTIONS.

	Ground truth						
Result	Cut	Mix	Add	Knead	Others	Total	Accuracy
Cut	**22**	1	1	0	3	27	81%
Mix	0	**85**	1	2	2	90	94%
Add	1	4	**23**	1	8	37	62%
Knead	0	2	1	**10**	3	16	63%
Others	4	5	4	2	**14**	29	48%
Total	27	97	30	15	30	**199**	77.3%

IV. CONCLUSION

In this paper, we proposed a method that detects and extracts video segments that contain repetitious motions, which is robust to motion shift. An experiment on actual repetitious motions showed the effectiveness of the proposed method compared to conventional methods. In addition, we reported a preliminary result of an experiment on the classification of the types of the detected repetitious motions.

Although the method was targeted for the detection and the classification of cooking activities, we consider that they can be applied to other kinds of real-world applications such as human activity detection and recognition in surveillance or life-log videos.

Future work includes dealing with the effect of camera-works, and the improvement of the integration of the two features in the detection process.

ACKNOWLEDGMENT

Parts of this work were supported by the Grant-in-Aid for Scientific Research from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

REFERENCES

[1] R. Hamada, S. Satoh, S. Sakai, and H. Tanaka, "Detection of important segments in cooking videos," in *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries 2001*, Dec. 2001, pp. 118–123.

[2] K. Miura, R. Hamada, I. Ide, S. Sakai, and H. Tanaka, "Motion based automatic abstraction of cooking video," in *Proceedings of the ACM Multimedia 2002 Workshop on Multimedia Information Retrieval*, Dec. 2002.

[3] K. Doman, C.-Y. Kuai, T. Takahashi, I. Ide, and H. Murase, "Video CooKing: Towards the synthesis of multimedia cooking recipes," in *Advances in Multimedia Modeling —The 17th International Multimedia Modeling Conference, MMM2011, Taipei, Taiwan, January 5–7, 2011 Proceedings, Part II*, ser. Lecture Notes in Computer Science, K.-T. Lee, W.-H. Tsai, H.-Y. M. Liao, T. Chen, J.-W. Hsieh, and C.-C. Tseng, Eds., vol. 6524, Jan. 2011, pp. 135–145.

[4] N. Otsu, "Towards flexible and intelligent vision systems —from thresholding to CHLAC," in *Proceedings of the 9th IAPR Conference on Machine Vision Applications*, May 2005, pp. 430–439.

[1] LIBSVM (`http://www.csie.ntu.edu.tw/cjlin/libsvm/`) with Radial Basis Function as a kernel was used.