# Action recognition using three-way cross-correlations feature of local motion attributes

Tetsu Matsukawa[1] and Takio Kurita[2]
[1]*University of Tsukuba,* [2]*Hiroshima University, Japan*
*t.matsukawa@aist.go.jp, tkurita@hiroshima-u.ac.jp*

## Abstract

*This paper proposes a spatio-temporal feature using three-way cross-correlations of local motion attributes for action recognition. Recently, the cubic higher-order local auto-correlations (CHLAC) feature has been shown high classification performances for action recognition. In previous researches, CHLAC feature was applied to binary motion image sequences that indicates moving or static points. However, each binary motion image lost informations about the type of motion such as timing of change or motion direction. Therefore, we can improve the classification accuracy further by extending CHLAC to multivalued motion image sequences that considered several types of local motion attributes. The proposed method is also viewed as an extension of popular bag-of-features approach. Experimental results using two datasets shows proposed method outperformed CHLAC features and bag-of-features approach.*

## 1. Introduction

Action recognition is an important research area in computer vision because it has wide ranges of applications such as video surveillance, video annotation, retrieval and human-computer interfaces. To recognize action in video, the feature extraction process from spatio-temporal volume plays an important role. To date, many spatio-temporal features were proposed [1, 4, 6, 7, 8]. Among them, we especially focus on Cubic Higher-order Local Auto-Correlations(CHLAC)[1] feature because of its desirable properties.

CHLAC can be obtained with a little computation and produces high classification performances in action recognition. CHLAC computes global histogram of local auto-correlations. Therefore, CHLAC has sift-invariance and thus segmentation free. In previous studies, CHLAC was applied to binary motion image sequences that indicates moved(+1) or static(0). However, each binary motion image lost informations about the type of motion such as motion direction or timing of larger motion happened in near frame. Therefore we can improve the classification accuracy further by extending CHLAC to multivalued motion image sequences that considered the several types of local motion attributes.

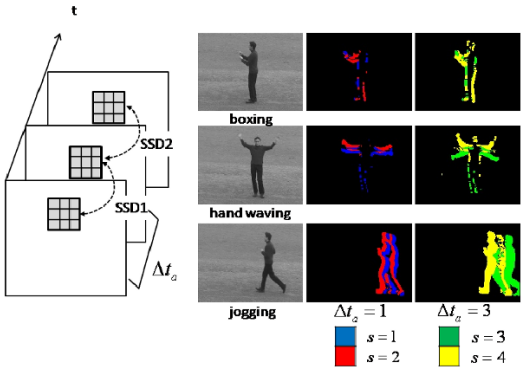


**Figure 1. Multivalued motion images: encoding process (left) and examples of KTH action dataset[5](right).**

As in object recognition, the bag-of-features approach becomes popular method for action recognition. This method uses an orderless collection of local motion attributes created by local features and a vector quantization algorithm. Previous bag-of-features methods use spatial-temporal layout information to improve the description power of global integration of local motions [6][7]. However, we think dividing spatial and temporal grid is not desirable, because the position of action may be different in each data. Thus, the sift invariance feature like CHLAC is desirable. Instead of using spatial temporal grid information, we propose a method that uses relative positions of local motion attributes in the frame work of CHLAC. We call this method as Motion Index Cubic Higher-order Local Auto-Correlations (MICHLAC). By regarding each local motion attribute

as a codebook in bag-of-features approach, this method is viewed as an extension of bag-of-features.

There are several related works to the proposed method. Instead of the binary motion image of CHLAC, there are HLAC fetures on motion history images[2]. However, it is kown thet HLAC is more efficient in quantized images. The idea of correlation of multivalued images is similar to GLAC [3] that correlates quantized gradient in image domain. The encoding part of local motion attributes in the proposed method is similar to Local Trinary Patterns (LTP)[4]. However, LTP uses histogram of local motion attributes as in bag-of-features method.

## 2. Index of local motion attribute

To encode local motion attributes, we can use several methods such as motion direction or timing of larger motion happened. In this paper, we use similar encoding method to [4] because of simplicity and low computational cost. However, we encodes timing of larger motion by considering only the same positions among near frames, while [4] encodes the pattern of motion directions.

Consider 3×3 patches of the same positions in time t - $\Delta\, t_a$, t, and t + $\Delta\, t_a$. Then the objective of the encoding process is to assign a motion index that indicates the timing of largest motion (change) happens among these patches. We use sum of square differences (SSD) for the distance between patches. If the change happens in t - $\Delta t_a \sim$ t, the SSD between t - $\Delta\, t_a$ and t (SSD1) will be high. If the change happens in time t $\sim$ t + $\Delta\, t_a$, the SSD between t and t + $\Delta t_a$ (SSD2) will be high. The larger motion timing $s$ is encoded as follows : $s = 1$ if SSD1 - SSD2 $>$ TH, $s = 0$ if |SSD1 - SSD2| $<$ TH, or $s = 2$ if SSD1 - SSD2 $<$ - TH. Where the value $s = 1$ indicates a larger motion happens in t - $\Delta_a \sim$ t, and the value $s = 2$ indicates a larger motion happens in t $\sim$ t + $\Delta t_a$. The value $s = 0$ indicates there are no motion in three frames and thus can be neglect. Namely, by using a single parameter of $\Delta t_a$, we can obtain 2 motion indexes $s \in \{1, 2\}$. We can obtain several motion indexes by using multiple $\Delta t_a$ and regard as each motion index as different motion index. For example, we get 4 motion indexes $s \in \{1, 2, 3, 4\}$ by using $\Delta t_a \in \{1, 3\}$. Example of multivalued motion images obtained by different time parameter $\Delta t_a$ are shown in Figure 1.

We use gray value between 0 and 255 for each frame image, and prepare two different thresholds TH. These are low threshold (TH = 1000) and Otsu's adaptive threshold. In adaptive threshold, we get TH by considering binarization problems of | SSD1- SSD2 |. Adaptive threshold searches the optimal threshold that maximize discriminant criterion $\sigma_B/\sigma_W$ per frame by 300 intervals from 1000 to max(| SSD1 - SSD2|). Here, $\sigma_W$ and $\sigma_B$ indicates within and between class variance of binarization, respectively.

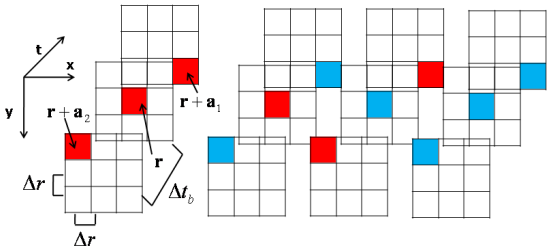


**Figure 2. Example of mask patterns of auto/cross-correlations: N=2, S=2.**

## 3. Three-way cross-correlations features

### 3.1 Definition of MICHLAC

Let f($s$,$\boldsymbol{r}$) be a three-way (cubic) data for the motion attribute $s \in \{1, .., S\}$ defined on the region $D$:$X \times Y \times T$ with $\boldsymbol{r} = (x, y, t)^T$, where $X$ and $Y$ are the width and height of the image frame and $T$ is the time length of the time window. First, the method creates f($s$, $\boldsymbol{r}$) for all indexes $s$ so that f($s$,$\boldsymbol{r}$) = 1 if $\boldsymbol{r}$ contains the motion index $s$, and $\boldsymbol{r}$ = 0 in other cases. Then the Nth order MICHLAC feature is defined as follows.

$$R_N(s_0, .., s_N, \boldsymbol{a}_1, ..., \boldsymbol{a}_N) = \int_{D_s} f(s_0, \boldsymbol{r}) f(s_1, \boldsymbol{r} + \boldsymbol{a}_1) \cdots f(s_N, \boldsymbol{r} + \boldsymbol{a}_N) d\boldsymbol{r}$$
$$D_s = \{\boldsymbol{r} | \boldsymbol{r} + \boldsymbol{a}_i \in D \; \forall i\}, \quad (1)$$

where $\boldsymbol{a}_i$ (i=1,..,N) are displacement vectors from the reference point $\boldsymbol{r}$. Equation (1) can take many different forms by varying $\boldsymbol{a}_i$ and N. These parameters are restricted to the following subset: $a_{nx}, a_{ny} \in \{\pm\Delta r, 0\}$, $a_{nt} \in \{\pm\Delta t_b, 0\}$ and N $\in \{0, 1, 2\}$. The parameter $\Delta r$,$\Delta t_b$ denote the spatial and temporal intervals, respectively. The interval along the x-axis is made equal to that along the y-axis because of isotropy in the x-y plane. On the other hand, the temporal interval $\Delta t_b$ may be different from the spatial interval $\Delta r$ because the resolution of space and time may differ. MICHLAC feature becomes the same as CHLAC when $S = 1$. CHLAC feature can be calculated by sliding predetermined mask patterns. By eliminating duplicates that arises from shifts of the reference point, there are 251 mask patterns of CHLAC for binary images [1]. One is for 0th order, and 250 are for 1st and 2nd order. We extended CHLAC mask patterns to MICHAC that

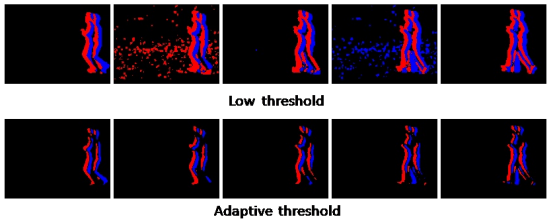


**Figure 3. Example of Robustness.**

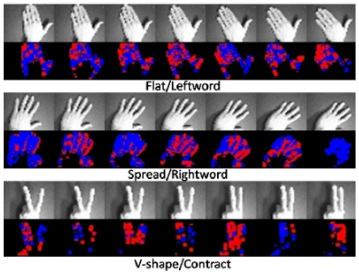


**Figure 4. Example of Hand-Gesture.**

has product of both same and different motion indexes (auto/cross-correlations) of local motion attributes. To avoid the increase of dimensions, we restrict the combination of motion index for 2nd order as follows: ($s_o$, $s_1$, $s_2$) = ($s_0$, $s_1$, $s_1$). Thus, the number of feature dimension becomes $S \times S \times 250 + S$. The example of mask patterns for 2nd order MICHLAC is shown in Figure 2. When using only 0th order, MICHLAC features becomes the same as bag-of-features of each local motion attributes.

### 3.2 Properties of MICHLAC

The desirable properties of CHLAC are its additivity, sift-invariance and robustness against noise. These properties are inherited in MICHLAC.

We considered the robustness against noise. Example of continuous five frames that contains weak camera motions are shown in Figure 3. In low threshold, it is shown one of the motion indexes appears as background noise in each frame because the camera motion affected all pixels consistently. Therefore, local cross-correlations between different attributes are less affected by noise caused by weak camera motions than local auto-correlation of single attributes. Adaptive threshold produce higher thresholds than low threshold, so the motion index images obtained by adaptive threshold are more robuster against weak background movements. However, low threshold produces high classification performances in standard data because it encodes many actual moving points than the case of the adaptive threshold.

## 4. Experiment

### 4.1 Experimental setup

We evaluated the proposed feature extraction algorithm on two commonly used dataset for human action recognition: Hand-Gesture dataset[9] and KTH-action dataset[5]. Hand-Gesture dataset contains nine hand gesture classes defined by three primitive hand shapes and three primitive motions. Each class contains 100 image sequences performed by two subjects under five different illuminations. Example images of Hand-Gesture dataset are shown in Figure 4. We resized each image size to 40×30 pixel. We followed the experimental settings of [9], i.e., training was performed on the data acquired in the single plane illumination and testing was done on the data acquired in the remaining settings. KTH human action dataset contains six classes of actions performed by 25 subjects in four different scenarios. We carried out a leave-one-out cross validation evaluation, i.e., for each run classifiers were trained using the videos of 24 subjects and remaining subjects were used for test samples. We did not use spatial-temporal grid informations for MICHLAC, i.e., $D$ is all frames and pixels in one action sequence. We used four scale spatial intervals ($\Delta$r = 1, 2, 4, 8) to extract richer spatial information per frame and concatenated each MICHLAC features. We set the temporal interval $\Delta t_b = 1$. The parameter for motion index $\Delta t_a$ was tuned per dataset.

We compared our methods to three baseline methods: CHLAC using binary images created by multi-valued motion images of low threshold (Baseline-a), CHLAC using binary images used in [1] (Baseline-b), and bag-of-motion indexes in spatial temporal grid (Baseline-c). In CHLAC, the same spatial and temporal parameters to MICHLAC is used. In Baseline-a, binary image sequences were created by regarding moved (+1) if s is not 0 and static(0) if s is 0. In Baseline-c, all frame images are divided to (m,n) spatial grids and time window is divided to k. Histograms of motion indexes are created in each grid. All combinations of (m,n,k) were evaluated from (m,n) = (2,2), (4,4), (8,8), (16,16) and k = (1, 2, 3, 4, 5, 6, 7, 8). We report the best result from them.

For classifications, a linear SVM was used by one-against-all. Before classification, each feature dimension was normalized so that the mean of each dimension (over all training data) is zero, and the standard deviation is one. A five-fold cross validation was carried out on the training set to tune the parameters of SVM.

**Table 1. Results of Hand-Gesture.**

| Methods | set1 | set2 | set3 | set4 | total |
|---|---|---|---|---|---|
| MICHLAC-a | **86.66** | **94.44** | **87.77** | **95.55** | **91.10** |
| MICHLAC-b | 47.22 | 63.88 | 55.55 | 71.66 | 59.58 |
| Baseline-a | 73.88 | 89.44 | **87.77** | 87.22 | 84.58 |
| Baseline-b | 73.33 | 68.88 | 39.44 | 47.77 | 57.36 |
| Baseline-c | 77.22 | 85.55 | 72.77 | 75.55 | 77.77 |
| Kim et. al.[9] | 81 | 81 | 78 | 86 | 82 |
| Niebles et. al.[6] | 70 | 57 | 68 | 71 | 66 |

**Table 2. Results of KTH human action.**

| Methods | rates | Methods | rates |
|---|---|---|---|
| MICHLAC-a | **93.85** | Kim et. al.[9] | **95.33** |
| MICHLAC-b | 89.55 | Bregonzio et. al.[8] | 93.17 |
| Baseline-a | 85.85 | Zhang et. al.[6] | 91.33 |
| Baseline-b | 81.46 | Zhao et. al.[7] | 89.67 |
| Baseline-c | 82.17 | Niebles et.al.[10] | 83.3 |

### 4.2 Recognition rates

Average classification rates of all categories for Hand-Gesture and KTH action dataset are shown in Table 1, 2, respectively. In these tables, the method MICHLAC-a is MICHLAC using low threshold and MICHLAC-b is MICHLAC using adaptive threshold. In Hand-Gesture dataset, we used two motion indexes created by $\Delta t_a = 2$ and the best combination of Baseline-c was (m, n, k) = (8, 8, 4). In KTH action dataset, we used 4 motion indexes created by $\Delta t_a$ = { 1, 3 } and the best combination of Baseline-c was (m, n, k) = (16, 16, 6). It is shown that MICHLAC shows better performances than baseline methods. MICHLAC exhibited higher classification performances than previous methods that reported in the same evaluation expect for [9] in KTH action dataset. Figure 5 shows the confusion matrices of each dataset. Similar actions "jogging" and "running" were most confused.

### 4.3 Robustness to camera motions

To evaluate the robustness to camera motions of MICHLAC, we add artificially camera motions to video sequences of KTH action dataset. The y coordinate centers of all frames were moved by $y_{center} = y_{center} + Asin(2\pi t/60)$, here $t$ is the frame number. The training were done by using the clean data, i.e. A = 0. In this comparison, we used 2 motion indexes created by $\Delta t_a$ = 1, and single spatial and temporal interval ($\Delta r = 1, \Delta t_b$ = 1). The classification results of each methods are shown in Table 3. Only cross and without cross shows only cross correlations feature vectors of MICHLAC and without cross correlations vectors, respectively. It is shown low threshold exhibited higher classification performance than adaptive threshold in A = 0, while adaptive threshold exhibited higher classification performances than low threshold as to increase the magnitude of motion. Only cross-correlations features exhibited the best results in MICHLAC using low threshold when $A = 2, 3$.

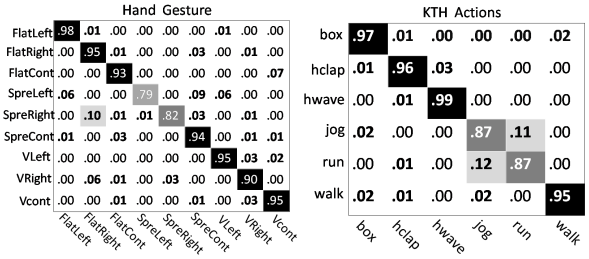


**Figure 5. Confusion matrices.**

**Table 3. Robustness Experiment.**

| threshold | feature | A=0 | A=1 | A=2 | A=3 |
|---|---|---|---|---|---|
| low | all | **87.72** | 80.23 | 65.47 | 57.02 |
| low | only cross | 87.14 | 79.81 | 70.10 | 65.29 |
| low | without cross | 84.51 | 77.59 | 68.17 | 61.00 |
| adaptive | all | 82.05 | **80.53** | **77.84** | **72.38** |

## 5. Conclusions

We proposed a spatio-temporal feature called MICHLAC. MICHLAC is an extension of CHLAC for improvement of the limitation of binary motion image sequences and uses three way cross-correlations of local motion attributes. Experimental results shows the proposed method outperformed the performances of CHLAC features and bag-of-features approach.

## References


[1] T.Kobayashi, and N.Otsu, Three-way auto-correlation approach to motion recognition, Pattern Recognition Letters, Volume 30, Issue 3, pp.212–221, 2009.

[2] K.Watanabe, and T.Kurita, Motion Recognition by Higher Order Local Auto Correlation Features of Motion History Images, In BLISS, 2008.

[3] T.Kobayashi, and N.Otsu, Image Feature Extraction Using Gradient Local Auto-Correlations, ECCV, 2008.

[4] L.Yeffet, and L.Wolf, Local Trinary Patterns for Human Action Recognition, In ICCV, 2009.

[5] C.Schuldt, I.Laptev, and B.Caputo, Recognizing human actions: a local svm approach, In ICPR, 2004.

[6] Z.Zhang, Y.Hu, S.Chan, and L.-T. Chia, Motion Context: A New Representation for Human Action Recognition, In ECCV, 2008.

[7] Z.Zhao, and A.Elgammal, Human Activity Recognition from Frame's Spatiotemporal Representation, In ICPR, 2008.

[8] M.Bregonzio, S.Gong and T.Xiang, Recognition Action as Clouds of Space-Time Interest Points, In CVPR, 2009.

[9] T.K. Kim, S.F.Wong, and R.Cipolla, Tensor Canonical Correlation Analysis for Action Classification, In CVPR, 2007.

[10] J.Niebles, and L.Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, IJCV, 79(3), 2008.