# Discriminant Appearance Weighting for Action Recognition

Tetsu Matsukawa
The University of Tokyo
Email: te2@iis.u-tokyo.ac.jp

Takio Kurita
Hirosima University
Email: tkurita@hiroshima.ac.jp

***Abstract*—Extending popular histogram representations of local motion patterns, we present a novel weighted integration method based on an assumption that a motion importance should be changed by its appearance to obtain better recognition accuracies. The proposed integration method of motion and appearance patterns can weight information involving "what is moving" by discriminant way. The discriminant weights can be learned efficiently and naturally using two-dimensional fisher discriminant analysis (or, fisher weight maps) of co-occurrence matrices. Original fisher weight maps lose shift invariance of histogram features, while the proposed method preserves it. Experimental results on KTH human action dataset and UT-interaction dataset revealed the effectiveness of the proposed integration compared to naive integration methods of independent motion and appearance features and also other state-of-the-art methods.**

## I. Introduction

Recognizing human actions from video sequences has a wide range of applications such as automatic video searches, human interfaces and video surveillances. To recognize actions in videos, a feature extraction process from spatio-temporal volume plays an important role. We assume that one requirement of basic spatio-tempral features is "shift invariance," i.e. the same feature should be obtained even if the position of the action is changed. Such shift invariance of spatio-temporal features brings a simple action recognition framework that doesn't require segmentation by bounding boxes of person [1]. In this paper, we focus on the problems that how we can improve discriminant abilities of base features without losing the shift invariance.

In recent years, recognition approaches using global representations of local motion patterns have shown impressive performances in action recognitions [1], [2], [3], [4], [6], [7], [8], [9]. In these methods, once the points to calculate feature are determined, local regions (cuboids) around the points are assigned to patterns such as cluster centers of local features [2], [6] or predetermined patterns [1], [9]. Then a histogram of local patterns is created as a global feature representation for recognition. If the histogram is created without dividing regions of video sequences, these feature representation have shift invariance. The patterns of local region can be roughly classified into motion features (e.g, HOF[5], 3DSIFT[6]) and appearance features (e.g, SIFT[2], SURF[5]). Among the motion features, there are features calculated directly on three-dimensional (image plane + time) volume [6], [1]. However, the resolution of image and time may differ and the information about image and motion is less than independent features. Although action recognitions using only appearance or motion information is possible [10], [1], the combination of motion and appearance information produces more reliable recognition than using one type of features. The most commonly used approach is the weighted concatenation of independent feature values of global representations [10], [7], [8]. The effectiveness of these feature combination is thought as true in both the cases that back ground information is crucial cue [7], [8] and even when recognizing different actions in the same background[10].

In this paper, we present a novel discriminant approach to combine local motion and appearance features based on an assumption that motion importance should be changed by its appearance. Consider a problem to classify "boxing" and "walking", boxing is more related to the hand movement and walking is more related to leg movements. Thus we believe changing importance based on its parts is effective for recognition. However, explict object information, such as hand or head, requires human labors for labeling. Instead of using such explict object information, our approach determines the discriminant importance by an automatic learning from only action label and data. More specifically, the weighing is realized by two-dimensional linear discriminant analysis(2DLDA) [11] of co-occurrence matrices of motion and appearance patterns. Because the discriminant weighting is realized based on appearance, the shift invariance of the original features is preserved.

Our proposed approach is inspired by two previous approaches of image classification. The feature weighting based on appearance is inspired from Top-Down Color Attension [12] in which shape descriptor is weighted by color descriptor in the same region. In thier research, the feature weighting is realized by plausibility of classes, not discriminant way. The discriminative weighting is inspired from Fisher Weight Maps (FWM) approaches[13]. FWM is the discriminat weighting of histogram features by its image position. Recently, FWM was applied to region weightings of local image descriptors[14], but weighting by image position loses shift invariance of the histogram features. Although it was not mentioned in previous researches, the formulation of FWM is the same as 2DLDA[11]. There are several variants of 2DLDA. One such variant, a tensor extension is used for gait recognition

978-1-4577-0121-4/11/$26.00 ©2011 IEEE

using gabor filter[15]. However, this method does not have shift invariance. Our technical contribution is to extend FWM to shift invariant version by extending coordinate position weighting to appearance weighting. Further, we shows bi-directional weighting and increaseing number of weights can produce better recognition accuracies, these are not explictly shown in the previous researches in FWM [13], [14].

## II. Related Work

There has been a large amount of successes by bag-of-features like approaches in action recognition. Here we describe differences to other researches for clarification.

Recently, the methods that use co-oocurance of motion feature became popular [3], [21]. These methods used co-occurence of single type of feature in different points, that requires large combinations of relative position of features. While the complexity of the proposed method is simple; the co-occurence of different features on the same position. The most recently, Ikizker-Cinbis et.al proposed person and object centric motion features[8]. They distinguished motion features by objects. In this point, this is similar concept to us. However, they relies on the person detector or object detector by tracking, that increase computational costs. Compared to their method, the proposed method is simpler and thus more computationally efficient.

Another difference to previous methods is the method for discriminant weighting. Popular ways to weight discriminant features are AdaBoost[7] or Multiple Kernel Learning (MKL)[8]. Liu. et. al. used AdaBoost to select combination of static and motion features[7]. Ikizker-Cinbis. al. used MKL to learn weights of feature channels[8]. However, AdaBoost requires several rounds to train classifier with weighted samples, and MKL is based on kernel method. Thus, both MKL and AdaBoost require large computational times. Proposed weighting method in this paper is based on eigen problems of discriminant analysis, thus can be learned quite fast.

## III. Base Features

In this section, we explatin the base motion and appearance features that are used in the proposed method. We use Cubic Higher-order Local Auto-Correlation(CHLAC) [1] as motion patterns and Higher-order Local Auto-Correlation (HLAC) [16] as appearance patterns. Originally, these are both global histogram representations of auto-correlation pattterns of local regions. These features can be calculated fast, yet produce good classification accuracies. These are both primitive features, thus the use of these features is suitable for the investigation of our method essentially. However, the proposed integration method is general and thus we believe it can be applied to any other local patterns e.g., k-means clustering of 3D-SIFT [6] and 2D-SIFT [2]. Nevertheless, pre-determined motion patterns such as CHLAC/HLAC are more praticital because one doesn't require the learning process of codebook per dataset.

At first, the video sequences are converted to binary sequences in which each point indicates a moved point or a static point. We thereshold the SSD of the 3x3 pixels of successive frames on the same point to obtain binary sequcences. If the camera is static[1], such binary sequences represent silhouettes of moving regions. Let f($\boldsymbol{r}$) $\in \{0,1\}$ be the binary video data defined on $D : X \times Y \times T$ with $\boldsymbol{r}= (x,y,t)^t$, where $X$ and $Y$ are width and height of the image frame and $T$ is the time length of the sequences. Here f($\boldsymbol{r}$) = 1 means a moved point and f($\boldsymbol{r}$) = 0 means a static point.

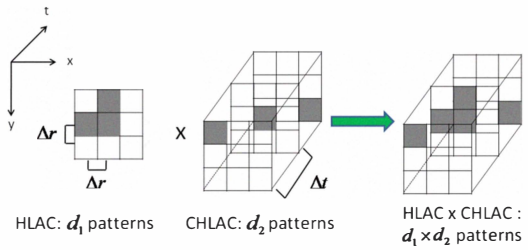


Fig. 1. Combination of appearance and motion patterns.

We calculate local features on all points that satisfy $\{\boldsymbol{r}|f(\boldsymbol{r}) = 1\}$ by following auto-correlation functions of HLAC [16] and CHLAC [1]. The auto-correlation function of HLAC is defined by,

$$v(\boldsymbol{r}) = f(\boldsymbol{r})f(\boldsymbol{r}+\boldsymbol{a}_1)\cdots f(\boldsymbol{r}+\boldsymbol{a}_N), \tag{1}$$

where $\boldsymbol{a}_n= (a_{nx}, a_{ny}, 0)^t, n = 1,...,N$ are displacement vectors in an image plane. These parameters are restricted to $a_{nx}, a_{ny} \in \{\pm\Delta r, 0\}$ and $N \in \{0,1,2\}$. The auto-correlation function $v(\boldsymbol{r})$ corresponding to one configuration of $(\boldsymbol{a}_1,...,\boldsymbol{a}_N)$ is one dimension of HLAC.

The auto-correlation function of CHLAC, the natural extension of HLAC, is defined by,

$$h(\boldsymbol{r}) = f(\boldsymbol{r})f(\boldsymbol{r}+\boldsymbol{a}'_1)\cdots f(\boldsymbol{r}+\boldsymbol{a}'_N), \tag{2}$$

where $\boldsymbol{a}'_n= (a_{nx}, a_{ny}, a_{nt})^t, n = 1,...,N$ are displacement vectors in image planes and time. These parameters are restricted to $a_{nx}, a_{ny} \in \{\pm\Delta r, 0\}$, $a_{nt} \in \{\pm\Delta t, 0\}$ and $N \in \{0,1,2\}$. As in HLAC, $h(\boldsymbol{r})$ corresponding to one configuration of $(\boldsymbol{a}'_1,...,\boldsymbol{a}'_N)$ is one dimension of CHLAC[2]. Note that CHLAC also contains image plane patterns. However, the information about the image plane of CHLAC is less than that of HLAC. Thus, the combination of CHLAC and HLAC has more detailed information about the image plane (Fig. 1).

[1]We assume camera is static to focus on the effect of proposed weighted integration. If camera is moving, a compensation method of camera motion is required to extract only moving regions of human.

[2]Here we describe the detailed parameter of CHLAC and HLAC we will report in Sec.V. We don't remove duplicate configurations caused by shift of reference points [16]. Because after CHLAC and HLAC are combined, such duplicate configurations become independent. Then by using complete combinations of $\boldsymbol{a}$ up to N= 2, the number of configuration patterns of HLAC becomes 37. To obtain richer information in image plane, we concatenate these HLAC patterns calculated by 5 scales ($\Delta r = 1,2,4,8,12$). The value of $N = 0$ (v($\boldsymbol{r}$) = f($\boldsymbol{r}$)) are common in 5 scales. Therefore we get responses of (36×5 + 1) = 181 dimensional appearance patterns from a position $\boldsymbol{r}$, i.e. the dimension of HLAC is set to $d_1 = 181$. As in HLAC, we don't remove duplicate configurations of CHLAC[1]. Then, there are 352 configurations of $\boldsymbol{a}'$. In this paper, the spatial and time interval of CHLAC are set to $\Delta r = 4$ and $\Delta t = 1$, respectively, i.e. the dimension of CHLAC is set to $d_2 = 352$.

## IV. Weighted Integration Method

In this section, we explain the proposed weighted integration of motion and appearance features. The outline of the proposed method is described in subsection A. Subsection B describes how to learn the appearance weights. In subsection C, we learn motion weights to increase classification accuracies further.

### A. Weighted integration of co-occurrence matrix

At first, we extract motion patterns and appearance patterns in the same position. Let $\boldsymbol{V}(\boldsymbol{r}) = (v_1(\boldsymbol{r}), ..., v_{d_1}(\boldsymbol{r}))^t$ be responses of $d_1$ appearance patterns and $\boldsymbol{H}(\boldsymbol{r}) = (h_1(\boldsymbol{r}), ..., h_{d_2}(\boldsymbol{r}))^t$ be responses of $d_2$ motion patterns of the reference point $\boldsymbol{r}$. Then, we combine the $\boldsymbol{V}(\boldsymbol{r})$ and $\boldsymbol{H}(\boldsymbol{r})$ in following co-occurrence matrix,

$$\boldsymbol{X}(\boldsymbol{r}) = \boldsymbol{V}(\boldsymbol{r})\boldsymbol{H}(\boldsymbol{r})^T, \tag{3}$$

where $\boldsymbol{X}(\boldsymbol{r})$ is a $d_1 \times d_2$ matrix, its $(i, j)$ component has product value of $v_i(\boldsymbol{r})$ and $h_j(\boldsymbol{r})$. We integrate the $\boldsymbol{X}(\boldsymbol{r})$ in spatio-tempral volume $D$. i.e.,

$$\boldsymbol{X} = \sum_{\boldsymbol{r} \in D} \boldsymbol{X}(\boldsymbol{r}). \tag{4}$$

The integral volume $D$ could be a subregion of video sequences. However, to obtain shift invariant features (in both position of space and time), we set $D$ to all regions of one video sequences. Then we create a $d_2$ dimensional feature vector $\boldsymbol{y}$, using a weight coefficient vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{d_1})^t$ and $\boldsymbol{X}$ as following equations.

$$\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{\alpha}. \tag{5}$$

This can be rewritten as,

$$\begin{aligned} \boldsymbol{y} = \sum_{\boldsymbol{r} \in D} \boldsymbol{X}(\boldsymbol{r})^T\boldsymbol{\alpha} &= \sum_{\boldsymbol{r} \in D} \boldsymbol{H}(\boldsymbol{r})\boldsymbol{V}(\boldsymbol{r})^T\boldsymbol{\alpha} \\ &= \sum_{\boldsymbol{r} \in D} w(\boldsymbol{V}(\boldsymbol{r}))\boldsymbol{H}(\boldsymbol{r}). \end{aligned} \tag{6}$$

where $w(\boldsymbol{V}(\boldsymbol{r})) = \boldsymbol{V}(\boldsymbol{r})^T\boldsymbol{\alpha}$ is a scalar and we call this value as an appearance weight. Eq.(6) means the responses of motion patterns are integrated using the weight based on the appearance of the point. Thus, one can weight the each motion feature on $\boldsymbol{r}$ without losing shift invariance unlike the position weight $w(\boldsymbol{r})$ of FWM [13].

### B. Appearance weight learning with 2DLDA

Next, we explain how to determine the weight coefficient vector $\boldsymbol{\alpha}$. From the discussion of previous subsection, one can understand if we determine $\boldsymbol{\alpha}$ so that maximize discriminant ability of $\boldsymbol{X}^T\boldsymbol{\alpha}$, then we can get appearance weight $w(\boldsymbol{V}(\boldsymbol{r}))$ which improve discriminant power of $\boldsymbol{y}$. To learn the weight coefficient vector $\boldsymbol{\alpha}$, we use two-dimensional discriminant analysis [11], [13], [18][3], which is proposed to slove this type problemes.

[3]The 2DLDA method is firsty proposed by Liu et.al[11]. There are several variations of 2DLDA. Among them, our formulation is the same as [13]. See references, e.g. introduction of [18] for more information about 2DLDA.

Suppose there are $M$ co-occurrence matrices $\{\boldsymbol{X}_i\}, i = 1, ..., M$ for training. We define the generalized within class covariance matrix $\boldsymbol{S}_W \in R^{d_1 \times d_1}$ and between class covariance matrix $\boldsymbol{S}_B \in R^{d_1 \times d_1}$ as following equations,

$$\boldsymbol{S}_W = \frac{1}{M}\sum_{j=1}^{C}\sum_{i \in c_j}(\boldsymbol{X}_i - \overline{\boldsymbol{X}_j})(\boldsymbol{X}_i - \overline{\boldsymbol{X}_j})^T, \tag{7}$$

$$\boldsymbol{S}_B = \frac{1}{M}\sum_{j=1}^{C} M_j(\overline{\boldsymbol{X}_j} - \overline{\boldsymbol{X}})(\overline{\boldsymbol{X}_j} - \overline{\boldsymbol{X}})^T. \tag{8}$$

where $M_j$ is the number of the samples in class $c_j$, $C$ is the number of classes, $\overline{\boldsymbol{X}_j} = \frac{1}{M_j}\sum_{i \in c_j}\boldsymbol{X}_i$ is the mean of $\boldsymbol{X}_i$ in the class $c_j$, $\overline{\boldsymbol{X}} = \frac{1}{M}\sum_{i}^{M}\boldsymbol{X}_i$ is the mean of the all training samples. Then, the extended fisher discriminant criterion to two-dimensional data is defined by,

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^T\boldsymbol{S}_B\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\boldsymbol{S}_W\boldsymbol{\alpha}}. \tag{9}$$

The largest $s$ weight coefficients $\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_s$, that maximize this criterion under the condition $\boldsymbol{\alpha}^T\boldsymbol{S}_W\boldsymbol{\alpha} = 1$ is obtained as the largest $s$ eigen vectors of following generalized eigen problem.

$$\boldsymbol{S}_B\boldsymbol{\alpha} = \lambda\boldsymbol{S}_W\boldsymbol{\alpha}. \tag{10}$$

To solve this generalized eigen problem, we transformed e.q.(10) to an ordinary eigen problem by applying eigenvector decomposition of $\boldsymbol{S}_W$ as described in [20]. We form a $d_2 \times s$ dimensional feature vector for classification by unfolding $\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{A}$, where $\boldsymbol{A} = [\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_s] \in R^{d_1 \times s}$. From e.q.(3) and (4), one can see $\boldsymbol{Y}$ is a concatenation of $s$ weighted motion features as, $\boldsymbol{Y} = \sum_{r \in D}\boldsymbol{X}(\boldsymbol{r})^T\boldsymbol{A} = \sum_{r \in D}\boldsymbol{H}(\boldsymbol{r})\boldsymbol{V}(\boldsymbol{r})^T\boldsymbol{A} = \sum_{r \in D}\boldsymbol{H}(\boldsymbol{r})[w_1(\boldsymbol{V}(\boldsymbol{r})), ..., w_s(\boldsymbol{V}(\boldsymbol{r}))]$.

Note that $s$ can be taken larger dimension than $C$-1 dimension unlike standard LDA. Previous FWM papers did not noticed this merit but actually it is easily confirmed that e.q.(10) is the standared LDA formulation that the class number becomes $C \times d$ [18]. Another advantage of 2DLDA against standard LDA is the ability to handle large dimensional features in more natural way[4].

### C. Bi-linear weight of motion and appearance

The learned weights from previous subsections are only with regard to appearance patterns. However, like CHLAC was compressed by LDA [1], weighting against motion patterns could improve the clssification abilities. Thus, better classification performances can be obtained by weighting the motion patterns in $\boldsymbol{X}$ not only the appearance patterns. This is also naturally realized with 2DLDA by applying the 2DLDA to two-directional of the matrix as to firstly proposed in [19].

[4]If we weight each element of $\boldsymbol{X} \in R^{d_1 \times d_2}$ with LDA by vectorizing $\boldsymbol{X}$, the weight coefficient dimension becomes $d_1 \times d_2$. For calculation of weight coefficients, $d_1d_2 \times d_1d_2$ dimensional within/between class covariance matrices are required. When $d_1$ or $d_2$ is large (e.g, $d_1 = 300$,$d_2 = 200$), LDA suffers from the memory lack or small sample problem. Thus un-natural process such as division of feature vector is required. 2DLDA can naturally learn weight coefficents without diving feature vector because the size of within/between covariance matrix is $d_1 \times d_1$.

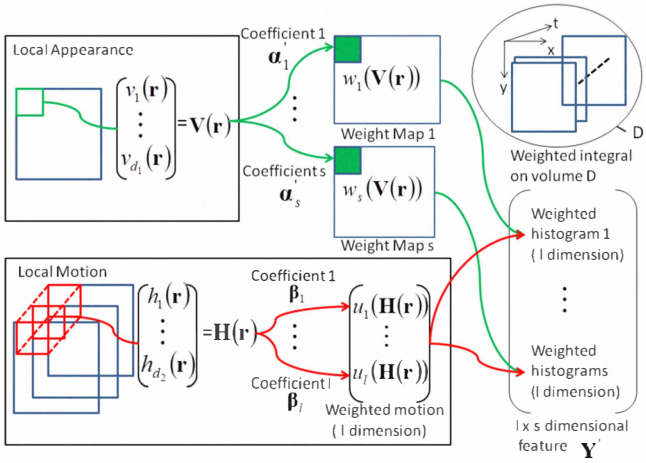


Fig. 2. Bi-linear weighted integral of local motion and appearance.

First, we weight the matrix using motion weight coefficient vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_{d_2})^t$ as follows,

$$\boldsymbol{X'} = (\boldsymbol{X}^T)^T \boldsymbol{B} = \boldsymbol{X}\boldsymbol{B}, \tag{11}$$

where $\boldsymbol{B} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, .., \boldsymbol{\beta}_l] \in R^{d_2 \times l}$. The learning method of the $\boldsymbol{\beta}$ is the same as the previous subsection. Namely, we calculate the generalized within class covariance matrix $\boldsymbol{S'}_W \in R^{d_2 \times d_2}$ and between class covariance matrix $\boldsymbol{S'}_B \in R^{d_2 \times d_2}$ with regard to $\boldsymbol{X}^T$ by replacing the $\boldsymbol{X}$ of e.q. (7) and (8) to $\boldsymbol{X}^T$. Then the $l$ weight coefficient vectors $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_l$ are obtained as the largest $l$ eigen vectors of following generalized eigen problem.

$$\boldsymbol{S'}_B \boldsymbol{\beta} = \lambda \boldsymbol{S'}_W \boldsymbol{\beta}. \tag{12}$$

Next, we apply the weight learning method with regard to $\boldsymbol{X'} \in R^{d_1 \times l}$ and get a weight coefficient matrix $\boldsymbol{A'} = [\boldsymbol{\alpha'}_1, ..., \boldsymbol{\alpha'}_s] \in R^{d_1 \times s}$. Finally, we get the following bi-linear weighted feature[5],

$$\boldsymbol{Y'} = (\boldsymbol{X'})^T \boldsymbol{A'} = \boldsymbol{B}^T \boldsymbol{X}^T \boldsymbol{A'}. \tag{13}$$

We form a $l \times s$ dimensional feature vector for classification by unfolding $\boldsymbol{Y'} \in R^{l \times s}$. From e.q.(3) and (4), one can see $\boldsymbol{Y'}$ as $\boldsymbol{Y'} = \sum_{r \in D} \boldsymbol{B}^T \boldsymbol{X}(\boldsymbol{r})^T \boldsymbol{A'} = \sum_{r \in D} \boldsymbol{B}^T \boldsymbol{H}(\boldsymbol{r}) \boldsymbol{V}(\boldsymbol{r})^T \boldsymbol{A'} = \sum_{r \in D} [u_1(\boldsymbol{H}(\boldsymbol{r})), ..., u_l(\boldsymbol{H}(\boldsymbol{r}))]^T [w_1(\boldsymbol{V}(\boldsymbol{r})), ..., w_s(\boldsymbol{V}(\boldsymbol{r}))]$, where $u(\boldsymbol{H}(\boldsymbol{r})) = \boldsymbol{\beta}^t \boldsymbol{H}(\boldsymbol{r})$ is a value of weighted motion feature. Thus, the $(i, j)$ elemsnt of $\boldsymbol{Y'}$ is corresponding to following weighted integral feature, $\boldsymbol{y'}_{i,j} = \sum_{\boldsymbol{r}} w_j(\boldsymbol{V}(\boldsymbol{r})) u_i(\boldsymbol{H}(\boldsymbol{r}))$. This means the weighted motion feature $u_i(\boldsymbol{H}(\boldsymbol{r}))$ is integrated with the appearance weight $w_j(\boldsymbol{V}(\boldsymbol{r}))$. This process is cleary shown in Fig.2.

## V. Experiments

We evaluated the proposed feature integration method using KTH human action dataset [4] and UT-interaction dataset [17]. KTH dataset contains six classes of actions performed by 25 subjects in four different scenarios. Following previous researches[3], [7], [22], we carried out a leave-one-out cross validation evaluation. UT-interaction dataset contains six classes of human-human interactions. This dataset is divided to 2 sets by the places where the actions were performed. Each set has 10 sequences per category. Following the dataset protocol[17], we performed 10-fold leave-one-out cross validation per set. We used segmented sequences of this dataset. Because the training samples of UT-interaction set is small, we increase the number of training samples twice by adding the horizontally flipped versions of original training sequences.

For classification, a linear SVM was used by one-against-all. Because the number of training samples of UT dataset is small, we used K-NN classifer with $k = 5$ for this dataset.

[5]Some methods that update $\boldsymbol{A}$' and $\boldsymbol{B}$ iteratively were proposed. But we didnot use iteration process as in [19] for simplicity.

TABLE I
Recognition accuracy (%) of different integration methods.

| Method | Dim. | KTH | UT Set1 | UT Set2 |
|---|---|---|---|---|
| Bi-Weighted | a | **95.50** | **76.66** | **71.66** |
| Bi-Weighted | b | 92.70 | 71.66 | 66.66 |
| Weighted CHLAC | 352×5 | 90.02 | 56.66 | 63.33 |
| LDA(CHLAC+HLAC) | 5 | 89.74 | 60.00 | 70.00 |
| LDA(CH.)+LDA(HL.) | 5+5 | 88.65 | 58.33 | 56.66 |
| LDA(CHLAC) | 5 | 88.99 | 60.00 | 45.00 |
| LDA(HLAC) | 5 | 79.38 | 61.66 | 65.00 |
| CHLAC+HLAC | 352+181 | 87.70 | 36.66 | 51.66 |
| CHLAC | 352 | 87.56 | 40.00 | 41.66 |
| HLAC | 181 | 73.61 | 38.33 | 55.00 |

a: $l \times s$ is set to 80×100 for KTH and 50×20 for UT.
b: $l \times s$ is set to 80×5 for KTH and 50×5 for UT.

### A. Results

First, we compared the proposed method to naive integration methods of CHLAC and HLAC. Experimental results are shown in Table I. In the table, + means the concatenated features of two independent feature vectors. As the dimension of LDA (Linear Discriminant Analysis), we used 5 (C-1) dimension. The Weighted CHLAC means the method of Sec.IV.B. The results show better performances than LDA(CHLAC+HLAC) and LDA(CHLAC)+LDA(HLAC) could be achieved by Weighted CHLAC in the case of KTH. The result of Bi-Weighted means when the both direction of co-occurrence matrix is weighted (the method of Sec.IV.C). Bi-Weighted outperforms the Weighted CHLAC and other integration methods. Furthermore, it is observed that the performances increase as to increase the number of weights by comparing different number of weights in Bi-Weighted. The classification accuracies by varying the number of weighs are shown in Fig. 3. It is observed that the higher performances can be obtained when the number of weights are higher than 5, i.e. $C - 1$, in both $s$ and $l$. The 2DLDA can take larger dimension than $C - 1$ while the dimension of LDA is limited up to $C - 1$. This is one of the merits of weightings by 2DLDA. However, too many weights especialy in UT-Set2 led to reduce the accuracy. This may be because the number of traning sample for classifiaction was small.

Next, we show the examples of learned weights by the proposed method in Fig. 4. These appearance weights are calculated from $\boldsymbol{X'}$. The color shows the sign the weights; red(+) and blue(-), but one can reverse the sign of weight coefficients $\boldsymbol{\alpha}$ as $-\boldsymbol{\alpha}$ because $J(\boldsymbol{\alpha}) = J(-\boldsymbol{\alpha})$, so only the differece of the

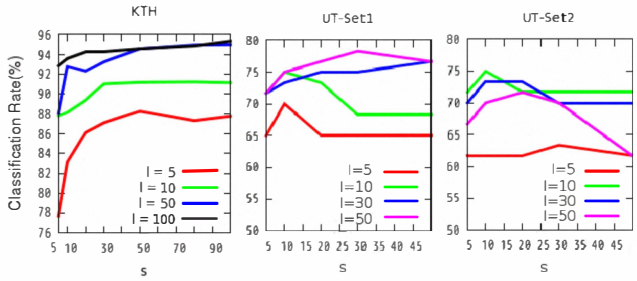


Fig. 3. Results of different number of bi-linear weights.

sign is meaningful. The distribution of the sign of the weights become different in each weights corresponding to different eigen vectors. Although the meaning of the weights are less understandable, it is shown that the weighed motion features by different appearance weights could be obtained. In third weights, the same sign of the appearance weight is learned. This is because the discriminant learning of the motion weight was sufficient to determine the sign of appearance weight. However, the absolute values of each position are different. In summary, the appearance weights have some varieties. We believe these varieties could help recognition.

Finally, we compared the accuracy of the proposed method to other state-of-the art methods in KTH. The best result of the proposed weghted integration method is 95.50 % as in table I. Alghouth the better results than the proposed method is also reported (e.g, Glibert et.al [21]), these methods use the learning of local patterns (i.e., codebook). Nevertheless of using predetermined patterns, our result is better than several codebook based approaches; the co-occurence of single feature (Ryoo at.el. 93.8% [3]) and motion and static features (Liu.et.al, 93.8%[7]). As predetermined pattern based methods, there are LTP[9] and MICHLAC (93.85% [22]). Because the recognition protocol of paper in [9] is different from us, we implemented LTP and classified with the same settings to us. The recognition rates of LTP were 89.41% (without partitioning spatial temporal grid as in us), and 94.66% (with best spatial temporal grid as in [9]). Thus, the proposed weighted integration of CHLAC and HLAC achieved the best result among the predetermined features, which is computationaly faster than codebook based approach. The weight learning of the proposed method is also fast, once co-occurence matrix $\boldsymbol{X}$ are extracted, the weights could be learned within 1.5 minutues per each training set with a standard workstation[6].

## VI. Conclusion

We have proposed a feature integration method of appearance and motion patterns by using two-dimensional fisher discriminant analysis. The proposed method could learn discriminant weights efficiently and naturally from co-occurrence matrices of base features. Experimental results showed the classification accuracies of the proposed integration were 5.76 % (in KTH), 15.00 % (in UT-set1), and 1.66% (in UT-set2) better than the native integration methods.

[6]With one core of Xeon 2.66GHz CPU and c++ implementation, it took only 67.05sec($\boldsymbol{B}$)+8.12sec($\boldsymbol{A'}$) for KTH, 5.3sec($\boldsymbol{B}$)+0.66sec($\boldsymbol{A'}$) for UT.

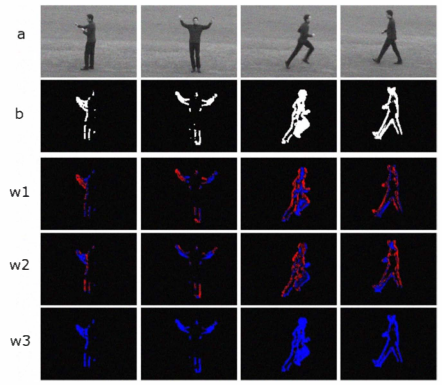


Fig. 4. Examples of appearance weights $w(\boldsymbol{V}(\boldsymbol{r}))$ of $\boldsymbol{X'}$ (KTH). a: original images, b: binarization of frame differencing, w1-3 : the weights corresponding to 1st-3rd eigen vectors.

As a future work, we are planning to examine a learning of sparse weight coefficients to obtain more understandable appearance weights.

## References


[1] T.Kobayashi and N.Otsu, "Three-way auto-correlation approach to motion recognition", *Pattern Recognition Letters*, 2009.

[2] P.Dollar, V.Rabaud, G.Cottrell, and S.Belongie, "Behavior recognition via sparse spatio-temporal features", in *ICCV VS-PETS*, 2005.

[3] M.S.Ryoo and J.K.Aggarwal, "Spatio-temporal relationship match: video structure comparison for recognition of complex human activities", in *ICCV*, 2009.

[4] C.Schuldt, I.Laptev, and B.Caputo, "Recognizing human actions: a local svm approach", in *ICPR*, 2004.

[5] H.Wang, M.M.Ullah, A.Klaser, I.Laptev, C.Schmid, "Evaluation of local spatio-temporal features for action recognition", in *BMVC*, 2009.

[6] P.Scovanner, S.Ali, and M.Shah, "A 3-dimensional SIFT descriptor and its application to action recognition", in *ACM MM*, 2007.

[7] J.Liu, J.Luo, and Mubarak Shah, "Recognizing Realistic Actions from Videos "in the Wild" ", in *CVPR*, 2009.

[8] N.I.-Cinbis, S.Sclaroff, "Object, Scene and Actions: Combining Multiple Features for Human Action Recognition", in *ECCV*, 2010.

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

[10] K.Schindler and L.V.Gool, "Action snippets: how many frames does human action recognition require ?", in *CVPR*, 2008.

[11] K.Liu, Y.Cheng, and J.Yang, "Algebraic feature extraction for image recognition based on an optimal discriminant criterion", *Pattern Recognition*, vol.26, no.6, pp.903-911, 1993.

[12] F.S.Khan. J.v.d.Wijer, M.Vanrell, "Top-Down Color Attenstion for Object Recognition", in *ICCV*, 2009.

[13] Y.Shinohara and N.Otsu, "Facial expression recognition using fisher weight maps", in *FG*, 2004.

[14] T.Harada, H.Nakayama, and Y.Kuniyoshi, "Improving Local Descriptors by Embedding Global and Local Spatial Information", in *ECCV*, 2010.

[15] D.Tao, X.Li, X.Wu, and S.J.Maybank, "General Tensor Discriminant Analysis and Gabor Features for Gait Recognition", *PAMI*, 2007.

[16] N.Otsu and T.Kurita, "A new scheme for practical flexible and intelligent vision systems", in *IAPR Workshop on Computer Vision*, 1988.

[17] M.S.Ryoo and J.K.Aggarwal, "UT-interaction dataset", *ICPR contest on Semantic Description of Human Activities(SDHA)*, 2010.

[18] S.Yan, D.Xu, Q.Yang, L.Zhang, X.Tang and H.-J.Zhang, "Multilinear discriminant analysis for face recognition", *IEEE trans. on Image Processing*, vol.16, no.21, pp.212–220, 2007.

[19] J.Yang, D.Zhang, X.Yong, and J. Yang "Two-dimensional discriminant transform for face recognition", *Pattern Recognition*, 2005.

[20] X.He and P.Niyogi, "Locality preserving projections (LPP)", *Technical Report*, The University of Chicago, 2002.

[21] A.Gilbert, J.Illingworth, R.Bowden, "Fast Realistic Multi-Action Recognition using Mined Dense Spatio-temporal Features", in *ICCV*, 2009.

[22] T.Matsukawa, T.Kurita, "Action Recongnition Using Three-way Cross Correlation Features of Local Motion Attributes", in *ICPR*, 2010.