Action recognition using weighted integration of co-occurrence patterns with discriminant weights

Tetsu Matsukawa
te2@iis.u-tokyo.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.

Key words: Action recognition, Bag-of-features, CHLAC features, Co-occurrence, Discriminant Analysis

1. 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-temporal 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 (Kobayashi and Otsu, 2009). In this paper, we focus on the problems that how we can improve discriminant abilities of base features without losing the shift invariance.

Classical action recognition methods use template matchings of spatial temporal templates. For example, Bobick and Davis (2001) used motion-history image and motion-energy image as temporal templates and matched by 7 Hu moments of temporal templates. Efros et al. (2003) matched spatio-temporal volume centered on a person by the optical flow vectors. Rodriguez et al. (2008) used MACH filter to create a template of a given action class. Methods in this category should perform segmentation of person regions or match template in several positions in videos, expect for the method of (Bobick and Davis, 2001) that uses shift invariant features.

In recent years, recognition approaches using global representations of local motion patterns have shown impressive performances in action recognitions (Cinbis and Sclaroff, 2010; Dollar et al., 2005; Kobayashi and Otsu, 2009; Ryoo et al., 2009; Liu et al., 2009; Schuldt et al., 2004; Scovanner et al., 2007; Yeffet and Wolf, 2009). In these methods, once the points to calculate feature are determined, local regions (cuboids) around the points are assigned to patterns. 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.

By the creating process of local patterns, the approaches are categorized into two classes; cluster centers of local features (Dollar et al., 2005; Scovanner et al., 2007) or predetermined patterns (Kobayashi and Otsu, 2009; Yeffet and Wolf, 2009). The use of cluster center was inspired from the bag-of-visualwords method of image classification(Csurka et al., 2004). The later class, predetermined mask pattern is extensions of mask pattern features (e.g., HLAC(Otsu and Kurita, 1988) or LBP(Ojala et al., 1996)) to spatio-temporal features. The merit of this predetermined mask pattern is simple and practical because one doesn't require the learning process of codebook.

By the information of local regions, the patterns also can be roughly classified into two classes; motion features (e.g, HOF(Wang et al., 2009), 3DSIFT(Scovanner et al., 2007)) and appearance features (e.g, SIFT(Dollar et al., 2005), SURF(Wang et al., 2009)). Among the motion features, there are features calculated directly on three-dimensional (image plane + time) volume (Scovanner et al., 2007; Kobayashi and Otsu, 2009). 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 (Schindler and Gool, 2008; Kobayashi and Otsu, 2009), 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 inde-

pendent feature values of global representations (Cinbis and Sclaroff, 2010; Schindler and Gool, 2008; Liu et al., 2009). The effectiveness of these feature combination is thought as true in both the cases that back ground information is crucial cue (Cinbis and Sclaroff, 2010; Liu et al., 2009) and even when recognizing different actions in the same background(Schindler and Gool, 2008).

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, explicit object information, such as hand or head, requires human labors for labeling. Instead of using such explicit 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) (Liu et al., 1993) 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 Attention (Khan et al., 2009) in which shape descriptor is weighted by color descriptor in the same region. In their research, the feature weighting is realized by plausibility of classes, not discriminant way. The discriminative weighting is inspired from Fisher Weight Maps (FWM) approaches(Shinohara and Otsu, 2004). FWM is the discriminat weighting of histogram features by its image position. Recently, FWM was applied to region weightings of local image descriptors(Harada et al., 2010), 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(Liu et al., 1993). There are several variants of 2DLDA. One such variant, a tensor extension is used for gait recognition using gabor filter(Tao et.al., 2007). 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 show bi-directional weighting and increasing number of weights can produce better recognition accuracies, these are not explicitly shown in the previous researches in FWM (Shinohara and Otsu, 2004), (Harada et al., 2010).

2. Related Studies

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-occurrence of motion feature became popular (Gilbert et al., 2009; Ryoo et al., 2009). These methods used co-occurrence 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-occurrence of different features on the same position. The most recently, Cinbis and Sclaroff (2010) proposed person and object centric motion features. 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(Liu et al., 2009) or Multiple Kernel Learning (MKL)(Cinbis and Sclaroff, 2010). Liu et al. (2009) used AdaBoost to select combination of static and motion features. Cinbis and Sclaroff (2010) used MKL to learn weights of feature channels. 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 for training. Proposed weighting method in this paper is based on eigen problems of discriminant analysis, thus can be learned quite fast.

3. Base Features

In this section, we explain the base motion and appearance features that are used in the proposed method. We use Cubic Higher-order Local Auto-Correlation(CHLAC) (Kobayashi and Otsu, 2009) as motion patterns and Higher-order Local Auto-Correlation (HLAC) (Otsu and Kurita, 1988) as appearance patterns. Originally, these are both global histogram representations of auto-correlation patterns of local regions. These features can be calculated fast, yet produce good classification accuracies. Several successes of HLAC/CHAC in action recognition are found in (Kobayashi and Otsu, 2009; Kurita and Hayamizu, 1988; Matsukawa et al., 2010). 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 (Scovanner et al., 2007) and 2D-SIFT (Dollar et al., 2005). Nevertheless, pre-determined motion patterns such as CHLAC/HLAC are more practical 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 (Fig. 1). We threshold the sum of squared distance (SSD) of the 3×3 pixels of successive frames on the same point to obtain binary sequences. If the camera is static[1], such binary sequences represent silhouettes of moving regions. Let f($\boldsymbol{r}$)

[0]This technical report is the extended version of Matsukawa et al. (2011). The extension includes additional survey and experimental results.

[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.

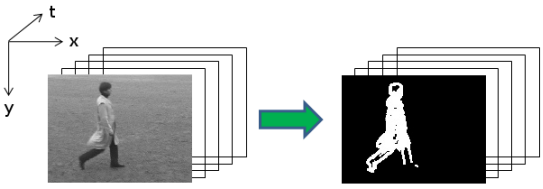

Figure 1: Preprocessing for HLAC and CHLAC. Left; original image sequences, Right: binarization of frame difference.

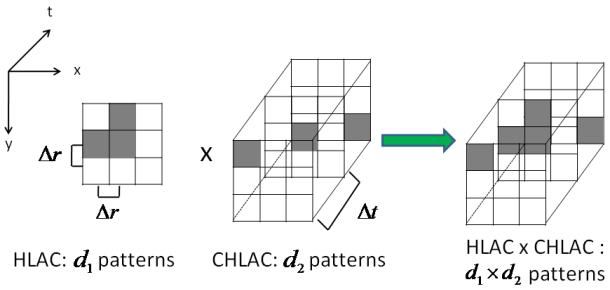

Figure 2: Combination of appearance and motion patterns.

$\in \{0, 1\}$ 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.

We calculate local features on all points that satisfy $\{\boldsymbol{r}|f(\boldsymbol{r}) = 1\}$ by following auto-correlation functions of HLAC (Otsu and Kurita, 1988) and CHLAC (Kobayashi and Otsu, 2009). 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.

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. 2).

4. 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 4.1. Subsection 4.2 describes how to learn the appearance weights. In subsection 4.3, we learn motion weights to increase classification accuracies further.

4.1. 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 (Shinohara and Otsu, 2004).

4.2. 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 (Liu et al., 1993; Shinohara and Otsu, 2004; Yan et al., 2007)[2], which is proposed to solve this type problemes. The 2DLDA is the extended version of linear discriminant analysis (LDA) so that can apply to matrix without vectorize the data.

[2]The 2DLDA method is firsty proposed in (Liu et al., 1993). There are several variations of 2DLDA. Among them, our formulation is the same as (Shinohara and Otsu, 2004). See references, e.g. introduction of (Yan et al., 2007) for more information about 2DLDA.

This can gives same weight to each column (in our case, appearance patterns) of matrix. The 2DLDA shows better recognition abilities when the high dimensional and small sample problems. The co-occurrence matrix is high dimensional, thus we believe these weightings are suitable for co-occurrence feature.

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,

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

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

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,

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

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.

$$\boldsymbol{S}_B\boldsymbol{\alpha} = \lambda\boldsymbol{S}_W\boldsymbol{\alpha}. \quad (10)$$

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 (He et al., 2002). 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 $C \times d$ (Yan et al., 2007). Another advantage of 2DLDA against standard LDA is the ability to handle large dimensional features in more natural way. 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 coefficients without diving feature vector because the size of within/between covariance matrix is $d_1 \times d_1$.

4.3. 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 (Kobayashi and Otsu, 2009), weighting against motion patterns could improve the classification 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 (Yang et al., 2005). Although the iterate algorithm of this weighting was proposed (Ye et al., 2004), our method is non-iterate algorithm as in (Yang et al., 2005) for simplicity.

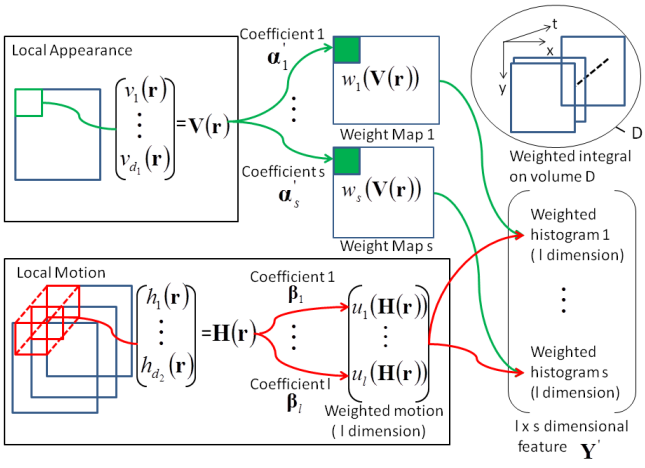

Figure 3: Bi-linear weighted integral of local motion and appearance.

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

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

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.

$$\boldsymbol{S}'_B\boldsymbol{\beta} = \lambda\boldsymbol{S}'_W\boldsymbol{\beta}. \quad (12)$$

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,

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

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) element 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 clearly shown in Fig.3.

5. Experiment

We evaluated the proposed feature integration method using KTH human action dataset (Schuldt et al., 2004) and UT-interaction dataset (Ryoo and Aggarwal, 2010). KTH dataset contains six classes of actions performed by 25 subjects in four different scenarios. Following previous researches(Ryoo et al., 2009; Liu et al., 2009; Matsukawa et al., 2010), we carried out a leave-one-out cross validation evaluation, i.e. for each run the weights and classifiers were trained using the videos of 24 subjects and remaining subjects were used for test samples. UT-interaction dataset contains six classes of human-human interactions. Some challenging factors of this dataset include moving background, cluttered scenes, camera jitters/zooms and different clothes. 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(Ryoo and Aggarwal, 2010), we performed 10-fold leave-one-out cross validation per set. , i.e., for each set, we leave one among 10 sequences for the testing and use the other 9 for the training. 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. A five-fold cross validation was carried out on training set to tune the parameters of SVM. Because the number of training samples of UT dataset is small, we used K-NN classifier with $k = 5$ for this dataset.

5.1. Detailed setup of base features

Here we describe the detailed parameter of CHLAC and HLAC we will report.

We didn't remove duplicate configurations caused by shift of reference points (Otsu and Kurita, 1988). Because after combining CHLAC and HLAC, such duplicate configurations become independent. Then by using complete combination 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 effectiveness of such multi-resolution of mask patterns is reported in (Toyoda and Hasegawa, 2007). The value of $N = 0$ ($\mathrm{v}(\boldsymbol{r}) = \mathrm{f}(\boldsymbol{r})$) are common in 5 scales. Therefore we get responses of $(36\times5 + 1) = 181$ dimensional appearance patterns from a position $\boldsymbol{r}$, i.e. $d1 = 181$, As in HLAC, we didn't ignore duplicate mask patterns of CHLAC. Then, there are 352 configurations of $(\boldsymbol{a}'_1, ..., \boldsymbol{a}'_N)$. The time interval of CHLAC is set to $\Delta t = 1$, and spatial interval of CHLAC is one of the $\Delta r = \{1, 2, 4, 8\}$, i.e. $d2 = 352$.

5.2. Results

5.2.1. Recognition rates

First, we compared the proposed method to naive integration methods of CHLAC and HLAC. Experimental results are shown in Table 1-3. In the table, Δr means the parameter of CHLAC, while parameter of HLAC was fixed as described in previous subsection. The results $\Delta r = (1, 2, 4, 8)$ means the concatenated feature by weighting results of each parameters. The symbol + means the concatenated features of two independent feature vectors. As the dimension of LDA (Linear Discriminant Analysis), we used 5 (C-1) dimensions. The Weighted_2DLDA means the method of Sec.4.2. The results show better performances than LDA(CHLAC+HLAC) and LDA(CHLAC)+LDA(HLAC) could be achieved by Weighted_2DLDA in many cases of KTH and some cases of UT. The result of Bi-Weighted_2DLDA means when the both direction of co-occurrence matrix is weighted (the method of Sec.4.3). Bi-Weighted_2DLDA outperforms the Weighted_2DLDA 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 of Bi-Weighted_2DLDA are shown in Fig. 4. In most cases, 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 especially in UT-Set2 led to reduce the accuracies. This may be because the number of training sample for classification was small.

In Table 1-3, the result of Weighted_2DPCA and Bi-Weighted_2DPCA mean the results by using weighting two-dimensional principal component analysis (Yang, et al., 2004) instead of weighting by 2DLDA. The 2DPCA is the unsupervised algorithm and thus the weighting is not discriminative. It is observed that the weighting by 2DLDA is better than weighting by 2DPCA in most cases, this shows the effectiveness of discriminant weightings than non discriminant weightings.

5.2.2. Learned weights

Next, we show the examples of learned weights by the proposed method in Fig. 5 - 6. 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 difference of the sign is meaningful. 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 of KTH and first weights of UT, the same sign of the appearance weights are 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.

5.2.3. Comparison to other methods

Finally, we compared the accuracy of the proposed method to other state-of-the art methods in KTH. The best result of the proposed weighted integration method is 94.54 % as shown in Table 1. This result is much higher than early papers (e.g,

Table 1: Recognition accuracy (%) of different integration methods.

Method	Dim.	Δ r = 1	Δ r = 2	Δ r = 4	Δ r = 8	Δ r = (1,2,4,8)
Bi-Weighted_2DLDA	50×50	**92.40**	**90.70**	**94.54**	**90.79**	**93.38**
Bi-Weighted_2DLDA	50×5	90.19	90.07	88.06	90.58	91.99
Bi-Weighted_2DPCA	50×50	92.28	92.58	92.67	91.90	92.54
Bi-Weighted_2DPCA	50×5	87.33	87.63	88.46	86.23	89.76
Weighted_2DLDA	352×5	88.58	90.17	90.02	89.78	91.23
Weighted_2DPCA	352×5	82,11	87.96	90.22	87.70	90.09
LDA(CHLAC+HLAC)	5	88.24	90.25	89.74	89.41	90.49
LDA(CHAC)+LDA(HLAC)	5+5	86.61	89.82	88.65	89.03	91.37
LDA(CHLAC)	5	85.32	88.32	87.70	88.39	88.95
LDA(HLAC)	5	79.38	79.38	79.38	79.38	79.38
CHLAC+HLAC	352+181	79.39	83.22	87.70	88.39	88.95
CHLAC	352	72.34	81.78	87.56	87.97	88.49
HLAC	181	73.61	73.61	73.61	73.61	73.61

Table 2: Recognition accuracy (%) of different integration methods UT-Set1.

Method	Dim.	Δ r = 1	Δ r = 2	Δ r = 4	Δ r = 8	Δ r = (1,2,4,8)
Bi-Weighted_2DLDA	50×20	73.33	**76.66**	**76.66**	**73.33**	**63.33**
Bi-Weighted_2DLDA	50×5	**78.33**	73.33	71.66	63.33	58.33
Bi-Weighted_2DPCA	50×20	36.66	36.66	38.33	38.33	36.66
Bi-Weighted_2DPCA	50×5	36.66	35.00	33.33	45.00	36.66
Weighted_2DLDA	352×5	51.66	55.00	56.66	58.33	58.33
Weighted_2DPCA	352×5	36.66	36.66	40.00	40.00	35.00
LDA(CHLAC+HLAC)	5	58.33	58.33	60.00	61.66	58.33
LDA(CHAC)+LDA(HLAC)	5+5	61.66	61.66	58.33	56.66	55.33
LDA(CHLAC)	5	56.66	65.00	60.00	51.66	53.33
LDA(HLAC)	5	61.66	61.66	61.66	61,66	61.66
CHLAC+HLAC	352+181	36.66	35.00	36.66	38.33	33.33
CHLAC	352	41.66	41.66	40.00	38.33	38.33
HLAC	181	38.33	38.33	38.33	38.33	38.33

Table 3: Recognition accuracy (%) of different integration methods UT-Set2.

Method	Dim.	Δ r = 1	Δ r = 2	Δ r = 4	Δ r = 8	Δ r = (1,2,4,8)
Bi-Weighted_2DLDA	50×20	60.00	75.00	**71.66**	**65.00**	66.66
Bi-Weighted_2DLDA	50×5	61.66	**81.66**	66.66	61.66	**73.33**
Bi-Weighted_2DPCA	50×20	46.66	48.33	51.66	43.33	51.66
Bi-Weighted_2DPCA	50×5	48.33	46.66	51.66	41.66	51.66
Weighted_2DLDA	352×5	**73.33**	66.66	63.33	53.33	68.33
Weighted_2DPCA	352×5	46.66	48.33	48.33	43.33	51.00
LDA(CHLAC+HLAC)	5	71.66	73.33	70.00	61.66	**73.33**
LDA(CHAC)+LDA(HLAC)	5+5	63.33	63.33	56.66	63.33	68.33
LDA(CHLAC)	5	50.00	46.66	45.00	60.00	63.33
LDA(HLAC)	5	65.500	65.00	65.00	65.00	65.00
CHLAC+HLAC	352+181	53.33	51.66	51.66	53.33	45.00
CHLAC	352	40.00	40.00	41.66	35.00	45.00
HLAC	181	55.00	55.00	55.00	55.00	55.00

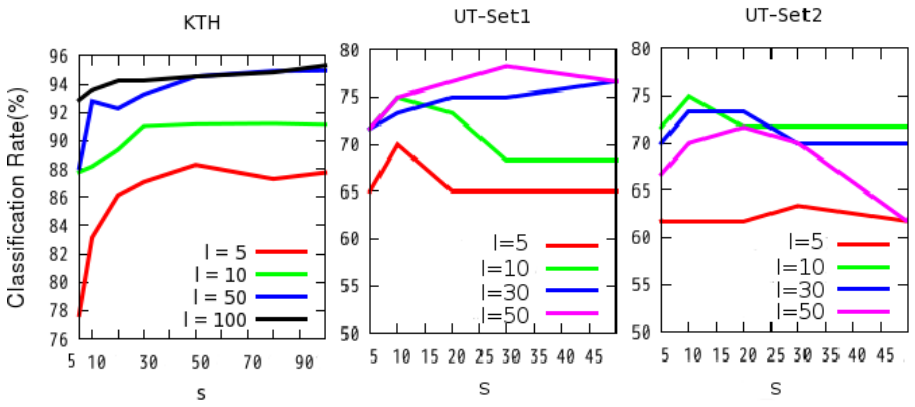

Figure 4: Results of different number of bi-linear weights (the spatial parameter of CHLAC is $\Delta r = 4$).

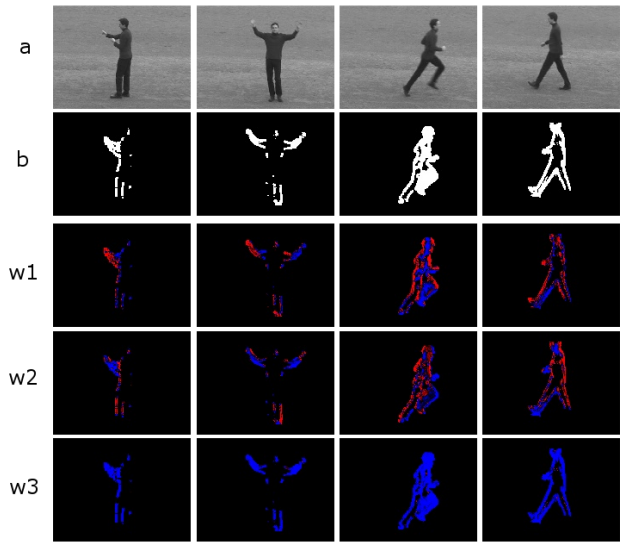

Figure 5: Examples of appearance weights $w(\boldsymbol{V}(\boldsymbol{r}))$ of $\boldsymbol{X}'$ (KTH). a: original images, b: binarization of frame differencing, w1-3 : the weights corresponding to 1st-3rd eigen vectors.

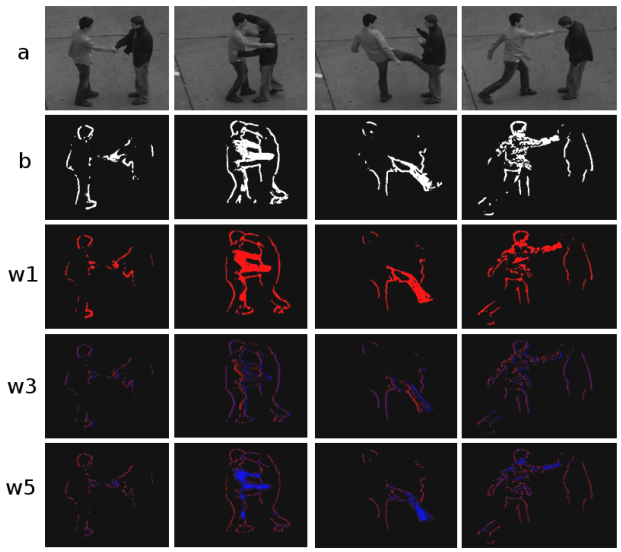

Figure 6: Examples of apperance weights $w(\boldsymbol{V}(\boldsymbol{r}))$ of $\boldsymbol{X}'$ (UT). a: original images, b: binarization of frame differencing, w1,3,5 : the weights corresponding to 1st,3rd,5th eigen vectors.

Neible and FeiFei (2008) 83.33%). Although the better results than the proposed method is also reported (e.g, Gilbert et al. (2009)), 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-occurrence of single feature (Ryoo et al. (2009) 93.8%) and motion and static features (Liu et al. (2009) 93.8%). As predetermined pattern based methods, there are LTP(Yeffet and Wolf, 2009) and MICHLAC (Matsukawa et al. (2010) 93.85%). Because the recognition protocol of paper in (Yeffet and Wolf, 2009) 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 (Yeffet and Wolf, 2009)). Thus, the proposed weighted integration of CHLAC and HLAC achieved the best result among the predetermined features, which is computationally faster than codebook based approach. The weight learning of the proposed method is also fast, once co-occurrence matrix $\boldsymbol{X}$ are extracted, the weights could be learned within 1.5 minutes per each training set with a standard workstation. 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.

6. 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. In addition, the discriminant abilities of base features can be increased without losing the shift invariance of histogram 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 (LDA of base features).

There are some possible extensions of the proposed integration for future work. One is to learn sparse weight coefficients for more understandable appearance weights. Another possible extension is to learn weights of frame importance as another type of appearance weights.

References

Bobick, A.F., Davis, J.W., 2001. The Recognition of Human Movement Using Temporal Templates, IEEE Transaction of Pattern Analysis and Machine Intelligence 23 (3), 257–267.

Cinbis, N.I.-, Sclaroff, S., 2010. Object, Scene and Actions: Combining Multiple Features for Human Action Recognition, in: European Conference on Computer Vison, pp.494–507.

Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints. in: ECCV Workshop on Statistical Learning in Computer Vision, pp.1–22.

Dollar, P., Rabaud, V., Cottrell, G., Belongie, S., 2005. Behavior recognition via sparse spatio-temporal features, in: ICCV workshop VS-PETS.

Efros, A., Berg, A., Mori, G., 2003. Recognizing action at a distance, in: International Conference on Computer Vision, pp.726–733.

Gilbert, A., Illingworth, J., Bowden, R., 2009. Fast Realistic Multi-Action Recognition using Mined Dense Spatio-temporal Features, in: International Conference on Computer Vision.

Harada, T., Nakayama, H., Kuniyoshi,Y., 2010. Improving Local Descriptors by Embedding Global and Local Spatial Information, in: European Conference on Computer Vision, pp.736–749.

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

Khan, F.S., Wijer, J.v.d., Vanrell, M., 2009. Top-Down Color Attention for Object Recognition, in: International Conference on Computer Vision.

Kobayashi, T., Otsu, N., 2009. Three-way auto-correlation approach to motion recognition, Pattern Recognition Letters 30 (3), 212–221.

Kurita, T., Hayamizu, S., 1988. Gesture recognition using HLAC features of PARCOR images and HMM based recognizer, in: International Conference on Automatic Face and Gesture Recognition, pp.422-427.

Liu, K., Cheng, Y., Yang, J., 1993. Algebraic feature extraction for image recognition based on an optimal discriminant criterion, Pattern Recognition 26 (6), pp.903-911.

Liu, J., Luo, J., Shah, M., 2009. Recognizing Realistic Actions from Videos "in the Wild" , in: IEEE Conference on Computer Vision and Pattern Recognition.

Matsukawa, T., Kurita,T., 2010. Action Recognition Using Three-way Cross Correlation Features of Local Motion Attributes, in: International Conference on Pattern Recognition.

Matsukawa, T., Kurita,T., 2011. Discriminant Apperance Weighting for Action Recognition, in: Asian Conference on Pattern Recognition.

Niebles,J., Fei-Fei,L., 2008. Unsupervised learning of human action categories using spatial-temporal words, International Journal of Computer Vision 79 (3), 299–318.

Ojala, T., Pietikäinen, M., Mäenpää,T., 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary pattern, IEEE Transaction on Pattern Analysis and Machine Intelligence 22 (7), 971–967.

Otsu, N. Kurita, T., 1988. A new scheme for practical flexible and intelligent vision systems, in: IAPR Workshop on Computer Vision, pp.431–435.

Rodriguez, M.D., Ahmed, J., Shah, M., 2008. Action MACH A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition, in: IEEE Conference on Computer Vision and Pattern Recognition.

Ryoo, M.S., Aggarwal.J.K., 2009. Spatio-temporal relationship match: video structure comparison for recognition of complex human activities, in: International Conference on Computer Vision.

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

Schindler, K., Gool, L.V., 2008. Action snippets: how many frames does human action recognition require ?", in: IEEE Conference on Computer Vision and Pattern Recognition.

Schuldt, C., Laptev, I., Caputo, B., 2004. Recognizing human actions: a local svm approach, in: International Conference on Pattern Recognition.

Scovanner, P., Ali, S., Shah, M., 2007. A 3-dimensional SIFT descriptor and its application to action recognition, in: ACM Multimedia.

Shinohara, Y., Otsu, N., 2004, Facial expression recognition using fisher weight maps, in: International Conference on Automatic Face and Gesture Recognition.

Tao, D., Li, X., Wu, X., Maybank, S.J., 2007. General Tensor Discriminant Analysis and Gabor Features for Gait Recognition, IEEE Transaction on Pattern Analysis and Machine Intelligence 29 (10), 1700–1715.

Toyoda, T., Hasegawa, O., 2007. Extension of higher order local autocorrelation features, Pattern Recognition 40, 1466–1473.

Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C., 2009. Evaluation of local spatio-temporal features for action recognition, in: British Machine Vision Conference.

Yan, S., Xu, D., Yang, Q., Zhang, L., Tang, X., Zhang, H.-J., 2007. Multilinear discriminant analysis for face recognition, IEEE Transaction on Image Processing 16 (21), 212–220.

Yang, J., Zhang, D., Yong, X., Yang, J., 2005. Two-dimensional discriminant transform for face recognition, Pattern Recognition 38 (7), 1125–1129.

Yang, J., Zhang, D., Frangi, A.F., Yang, J., 2004. Two-dimensional PCA: a new approach to appearance-based face representation and recognition, IEEE Transaction on Pattern Analysis and Machine Intelligence 26 (1), 131–137.

Ye, J., Janardan, R., Li, Q., Two-dimensional linear discriminant analysis, in: Advances in Neural Information Processing Systems (NIPS), vol17:pp.1569–1576.

Yeffet, L., Wolf, L., 2009, Local Trinary Patterns for Human Action Recognition, in: International Conference on Computer Vision.