ABSTRACT
We present an oriented holistic feature, namely weighted oriented pixel change history (WOPCH), to describe reciprocating motions and differentiate actions similar in appearance for human action recognition. To construct the oriented representation, we incorporate motion information into pixel change history (PCH) image, through splitting the PCH image into several oriented channels according to the corresponding motion direction. Moreover, relative velocities of body parts are also considered and therefore the accumulating is adapted with the weight of each pixel’s relative speed. Afterwards, invariant features are extracted from WOPCH images and a naive Bayes model is used for recognition. Experimental results show that our method outperforms traditional holistic approaches without requiring lots of training samples.

Index Terms— Action recognition, Pixel change history, Naive Bayes classifier

1. INTRODUCTION

Recently, human action recognition in videos has attracted more and more attention due to its wide applications such as intelligent surveillance, video annotation, augmented reality, and human-computer interaction.

Feature extraction plays an important role in human action recognition. Existing feature extraction methods can be broadly categorized as local representation based approaches and global representation based approaches. Local features [1–3], usually referring to appearance representations such as gesture, silhouette and interest point, or motion representations such as optical flow and spatial-temporal interest point, can be extracted from some significant frames and applied for sequential pattern classification task. Needs of salient detection and tracking, make local approaches usually meet limitations in cases of occlusions, illumination changes, camera movements and viewpoint changes in real-life application. On the contrary, global features (or holistic features), often generated in the whole sequence, ignore the temporal information. Compared to local approaches, global approaches usually produce more robust features at lower computational cost, and the obtained features can be employed for simple distance based classifiers. Bobick and Davis [4] defined the motion energy image (MEI), a binary image in which the foreground indicates where motion has occurred, to accumulate all motion regions, and the motion history image (MHI), in which a pixel with higher intensity indicates a more recent movement, to describe the temporal history of motion. The seven Hu moments [5], describing objects independent of position, size, and orientation, are then generated from MEIs and MHIs and used for pattern identification. More recently, Tao Xiang and Shaogang Gong [6] proposed the pixel change history (PCH) image to capture multi-scale temporal change history at each pixel. The scale information is incorporated into the accumulation stage as well as the delay stage to determine the temporal window of movement history and the importance of recent changes. PCH can be seen as an extension of MHI with an additional accumulation stage and shows better discrimination performance than the latter. Other variants [7, 8] of MHI, are also proposed for movement representation and recognition in various applications. However, MHI representation and its variants may encounter troubles and show poor performance for action recognition in some situations. Firstly, it is difficult to distinguish walking from running using MHI and its variants [9]; secondly, they are incapable to handle with reciprocating motions.

In this paper, an oriented holistic feature, namely weighted oriented pixel change history (WOPCH), is proposed for human action recognition to overcome two main disadvantages of traditional holistic features, unsuitable for describing reciprocating motions and difficult to differentiate actions similar in appearance. In particular we incorporate the oriented motion information into PCH, through subdividing the PCH image into several oriented channels according to the corresponding motion direction. Moreover, relative velocities of body parts are also considered and therefore the accumulating factor in each channel is enhanced with a weight indicating the speed of local events. After that, translation and scale
invariant features are extracted from WOPCH images and then a naive Bayes classifier is employed in the classification stage. Experimental results on benchmark data show that our method outperforms state-of-the-art approaches without requiring numerous training samples.

The rest of the paper is organized as follows: Sect.2 presents the weighted oriented pixel change history image; Sect.3 introduces the technique for invariant feature extraction from WOPCH images; Sect.4 introduces the naive Bayes model for recognition. Experimental setup and results on benchmark dataset are presented in Sect.5, and conclusions are given in Sect. 6.

2. WEIGHTED ORIENTED PIXEL CHANGE HISTORY IMAGE

Given an action video, its optical flows are firstly calculated between adjacent frames using Lucas-Kanade algorithm. Then a median filtering procedure is performed by subtracting the flow vectors from the median of flow field to remove the background motion component. These obtained compensated flows are robust to camera movement. We then extract the compensated flows $[v_x, v_y]$ from regions of interest in all frames and subdivide them into $k$ channels according to their primary angles from the horizontal axis. The orientation channels are spread over 0 to 360 degrees, and for each channel, the oriented pixel change history(OPCH) image can be computed as follows:

$$p_i(x, y, t) = \begin{cases} \min(p_i(x, y, t - 1) + \frac{255}{\tau}, 255) & \text{if } s_i(x, y, t) = 1 \\ \max(p_i(x, y, t - 1) - \frac{255}{\tau}, 0) & \text{if } s_i(x, y, t) = 0 \end{cases} \quad (1)$$

where $(x, y)$ is the pixel position and $t$ is the time. $\tau$ and $\xi$ is the accumulation and decay parameter respectively. The subscript $i$ indicates the $i$th channel, with $1 \leq i \leq k$. $s_i(x, y, t)$ presents the motion of pixel $(x, y, t)$ in the $i$th channel, which can be obtained by:

$$s_i(x, y, t) = \begin{cases} 1 & \text{if } v_i(x, y, t) > T \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $T$ is a selected threshold and $v_i$ is the motion velocity in the $i$th channel, which is equipped to the compensated flow’s magnitude $\sqrt{v_x^2 + v_y^2}$ if the primary angle $\theta = \tan^{-1}(v_y/v_x)$ is within the range of this channel, and 0 otherwise.

Motivation of the introduced OPCH is to resolve a problem of traditional holistic features such as MHI and its invariants, difficult to represent an action being composed of reciprocating motions, just as "walking left and turning back", "bending down and getting up", "waving hand down and up". An example of representations using MHI,PCH and a 4-channels-OPCH on the action "bending down and getting up" is given in Fig.1. For OPCH, the directions of optical flows are assigned to four bins, representing moving left, down, right and up respectively. It can be observed that the OPCH images are able to capture information of movement direction for representing the reciprocating motions, whereas in MHI and PCH image, no information about directional indication has been presented.

![Comparison of MHI, PCH and OPCH on action "bending down and getting up"](image)

Furthermore, we introduce the weighted oriented pixel change history(WOPCH) image to overcome another problem of traditional holistic features, that is, difficult to separate similar actions in appearance. The WOPCH image can be computed as follows:

$$p_i(x, y, t) = \begin{cases} \min(p_i(x, y, t - 1) + \frac{255}{\tau} - r \cdot \frac{v_i(x, y, t)}{\max(v_i(x, y, t))}, 255) & \text{if } s_i(x, y, t) = 1 \\ \max(p_i(x, y, t - 1) - \frac{255}{\xi}, 0) & \text{if } s_i(x, y, t) = 0 \end{cases} \quad (3)$$

Compared to that of OPCH, an weight is added in the accumulating step of WOPCH to ensure that local parts with a fast movement in one direction channel have more augmentation in the accumulation step than that of slow movement at every time. The parameter $r$ ensures that the rate of accumulation at each time is large enough. This weighting operation is beneficial for constructing a more discriminative representation for distinguishing some actions such as walking and running, which are always ambiguous in representation and difficult to be separated for MHI induced approaches. An example of OPCH and WOPCH with four channels describing actions "walking", "running" and "skipping" is illustrated in Fig.2. It is clear that the action "walking" has a uniform mov-
ing speed among all body parts in the first directional channel. For action “running”, the prime movements in this channel are concentrated in the breast and legs of body, whereas for “skipping”, they are concentrated in the torso and lower part of body. Compared to OPCH, WOPCH has enhanced such discriminative information in a natural way.

![OPCH-walking](a)  ![WOPCH-walking](b) ![OPCH-running](c)  ![WOPCH-running](d) ![OPCH-skipping](e)  ![WOPCH-skipping](f)

**Fig. 2.** Comparison of OPCH and WOPCH on actions “walking”, “running” and “skipping” in the first channel

### 3. IN Variant FEATURE EXTRACTION

We extract a kind of discriminative and representative feature from the WOPCH images obtained in above. Firstly, scale and translation invariance is achieved using the approach proposed in [10]. Given a WOPCH image \( f(x, y) \) of one channel, it is transformed to a scale and translation invariant space by:

\[
g(x, y) = f\left(\frac{x}{\sqrt{\beta/m_{00}}} + m_{10}, \frac{y}{\sqrt{\beta/m_{00}}} + m_{01}\right) \tag{4}
\]

where \( g(x, y) \) is transformed image invariant to translation and normalized to a unified scale \( \beta \) predetermined by user. \( m_{pq} \) is the raw moment of image \( f(x, y) \):

\[
m_{pq} = \int \int x^p y^q f(x, y) \, dx \, dy \tag{5}
\]

where \( p, q = 0, 1, \ldots \). Specially, the zeroth order moment \( m_{00} \) is the sum of grey level of image \( f(x, y) \), and the first order moments \((m_{10}, m_{01})\) is the localization of center of mass. Through the transformation of Eq.(4), all WOPCH images are normalized to scale \( \beta \) and centered at location \((m_{10}, m_{01})\). The nearest interpolation is employed to obtain the grid location for function \( g(x, y) \). The rotational invariance is not considered here due to its negligible value for distinguishing different actions.

Afterwards, we apply the principal component analysis (PCA) dimensional reduction on the transformed images. PCA reduces the feature vectors to \( n \) principal components by preserving eigenvectors corresponding to \( n \) largest eigenvalues after performing eigenvector vector decomposition on the pooled covariance matrix of data points. In our experiment, the percentage of preserved principal components is set as 90%. Then, for that WOPCH images whose zeroth order \( m_{00} \) are larger than a predefined threshold, the fuzzy C-means (FCM) clustering is performed in the reduced subspace of the \( i \)th channel to obtain \( C_i \) clusters. We denote the assigned cluster labels of an action video in the \( i \)th channel as \( h_i \), whose values belong to set \( \{0, 1, 2, \ldots, C_i\} \). WOPCH images with \( m_{00} \) smaller than the threshold are directly assigned to the cluster “0”, indicating the background, clutter or small movements with limited help for recognition.

### 4. NAIVE BAYES CLASSIFYING

Now we have cluster labels of all channels \( \{h_1, h_2, \ldots, h_k\} \) for every action video. Our aim is to estimate the action label \( y \) given cluster labels \( \{h_1, h_2, \ldots, h_k\} \) of each sample. The naive Bayes classifier model the posterior:

\[
P(y|h_1, \ldots, h_k) = \frac{P(y)\prod_{i=1}^{C} P(h_i|y)\prod_{i=1}^{C} P(h_i)}{Z} \tag{6}
\]

where \( P(y) \) is the prior probability and \( P(h_1, \ldots, h_k|y) \) is the likelihood function. Given the assumption that each channel is independent of any other ones, we have:

\[
P(y|h_1, \ldots, h_k) = \frac{P(y)\prod_{i=1}^{C} P(h_i|y)}{\prod_{i=1}^{C} P(h_i)} \tag{7}
\]

where \( P(h_i|y) \) is the probability for one sample with label \( y \) to be assigned to cluster \( h_i \), and \( Z = \prod_{i=1}^{C} P(h_i) \) is a constant.

Given action label \( y \), the probability of \( h_i \) can be estimated simply by:
\[
P(h_i = p | y = q) = \frac{P(h_i = p, y = q)}{\sum_{p=1}^{C} P(h_i = p, y = q)} \]
\[
= \frac{\text{Card}(C^{-1}(p) \cap L^{-1}(q))}{\text{Card}(L^{-1}(q)) - \text{Card}(C^{-1}(0) \cap L^{-1}(q))}
\]

where \(p \in \{1, 2, \ldots, C\}\) is a cluster label; \(q \in Y\) is an action label; \(C\) is the mapping function between feature \(X_i\) and cluster label \(h_i\); \(L\) is the mapping function between cluster label \(h_i\) and action label \(y\). The cardinal function \(\text{Card}(A)\) indicates the cardinal number of set \(A\). Specially, for samples whose cluster label \(p = 0\), we directly let \(p(h_i = 0 | y = 0) = 0\) (in actual application a very small constant instead), since small movements have limited benefit for improving the classification whereas the background and clutter always present as a common component of all channels so as to depress the discriminative ability, especially for a classifier based on the assumption of independence between channels.

Finally, the action label can be estimated through the maximum posterior probability criteria as follows:

\[
\hat{y} = \arg \max_{y \in Y} P(y) \prod_{i=1}^{k} P(h_i | y)
\]

5. EXPERIMENTAL RESULTS

A challenging human action recognition database, Weizmann dataset [11], is used as benchmark. It contains ten types of human action (running, walking, skipping, jumping, jack, jumping-forward-on-two-legs, jumping-in-place-on-two-legs, galloping-sideways, waving-two-hands, waving-one-hand and bending) performed several times by nine different people. This database contains 93 video sequences, each of which has a resolution of 180 × 144 and a fps of 50. For each sequence, binary silhouette of foreground region has been provided by background subtraction.

We fix \(\tau = 10, \xi = 255, k = 4, r = 3, T = 0.7, \beta = 10^5\). A total of 33 sequences are allocated to the training set, in which each action contains about three training samples, and the remind 60 sequences are used for testing. For each sequence, a period of action is extracted for generating features. An accuracy of 90.0% on Weizmann dataset is achieved, and the confusion matrix is given in Fig.3. A large confusion exists between running and skipping for their shape similarity in representation. Our oriented feature is compared with two popular holistic features, MHI and PCH. The Hu moments, commonly used for invariant feature extraction, are employed to extract features from MHI and PCH images. The k-nearest neighbor(KNN) classifier is then employed for recognition; leave-one-out cross-validation is performed to determine the optimal choice of parameter \(K\) and count the classification accuracy. The comparison result is shown in Table 1. We can see that our WOPCH feature significantly outperform both of MHI and PCH. We also compare our method to other state-of-the-art methods. Only here our method is evaluated on the strategy of splitting dataset to the training and testing set. Compared to other methods using leave-one-out cross-validation, our method achieves a considerable accuracy at the price of fewer training samples.

### Table 1. Comparison of our method with other methods on Weizmann dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our method</td>
<td>90.00</td>
</tr>
<tr>
<td>MHI+HU+KNN</td>
<td>48.39</td>
</tr>
<tr>
<td>PCH+HU+KNN</td>
<td>37.63</td>
</tr>
<tr>
<td>Niebles et al.</td>
<td>90.00</td>
</tr>
<tr>
<td>Scovanner et al.</td>
<td>82.60</td>
</tr>
<tr>
<td>Liu et al. [12]</td>
<td>90.40</td>
</tr>
<tr>
<td>Klaser et al. [13]</td>
<td>84.30</td>
</tr>
</tbody>
</table>

### Fig. 3. Confusion matrix of our method on Weizmann dataset

6. CONCLUSIONS

This paper presented a novel method for human action recognition using an oriented holistic representation, weighted oriented pixel change history(WOPCH), with naive Bayes model. The oriented holistic feature provided a natural and discriminative representation for reciprocating motions using information of motion direction. Moreover, by adapting the accumulating with weight of each pixel’s relative speed, the problem of ambiguous to represent some actions similar in appearance such as walking and running was also refrained from to some extent. The approach was validated on Weizmann dataset, showing a considerable performance without requiring plenty of training samples. Our future work will include improvement of the classifier and evaluation on more human action videos.

7. REFERENCES


