Fig. 1 Sample frames from our behavior database (25 fps): three frames sampled by every six frames are used to characterize actions (walking). a A normal behavior of walking; b same person but with a diverse appearance; c a different person with a different velocity; d action captured from a different viewpoint; e moving cameras and dynamic backgrounds; f multiple active subjects with occlusions

introduce ÒGentle AdaboostÓ into our framework for histogram-based feature selection, and to learn models on the proposed three kinds of histograms. In turn, we use the standard window scanning technique and apply the learned classiÞers onto the densely sampled rectangular sub-windows (i.e., cubic spatio-temporal volumes) within video sequences for the detection.

The rest of this paper is organized as follows. Section 2 summarizes the related work. Section 3 details feature design and representation for human actions using local motion histograms. Section 4 describes how we learn models on histogram-based features using a boosting framework. Section 5 illustrates our experimental settings and presents experimental results on several video sequences. Section 6 concludes this paper.

2 Related work

Recently, there has been signiÞcant interest in approaches that address human action detection. Many previous approaches for behavior recognition were based on tracking models [17,19,22], which apply tracked motion trajectories of body parts to action recognition. In this case, an accurate segmentation of subjects from backgrounds is assumed beforehand. Consequently, the robustness of the algorithm is highly dependent on the segmenting and tracking system. The authors in [5] have reviewed the previous work on activity recognition, most of which involve tracking body parts segmented from the static background.

Another class of approaches performs recognition by using sparsely detected spatio-temporal features. Schuldt et al. [11,18] devise spatio-temporal feature detectors on Harris corners in 3D case, and Doll‡r et al. [3] design the detectors from local maxima of the response function deÞned by separable linear Þlters. Although these approaches indicate good potentials, they pose a problem toward handling more challenging situations, i.e., it requires analyzing the spatio-temporal conÞguration between different cuboids in some cases, such as multiple actions recognition.

An alternative method is to apply spatio-temporal volumetric feature that efÞciently scan video sequences in space and time domain [8,12]. In substance this approach extends the rectangle features used by Viola and Jones [20] into the spatio-temporal domain for video analysis, which shows promising results in human action recognition. Most closely related to our work is that of [12], who apply similar action classiÞers to test videos in all spaceÐtime windows with variable spatial size and temporal extent. Furthermore, they introduce a technique called ÒKeyframe PrimingÓ that combine subjectÕs appearance with motion into retrieving actions. In spite of that, we argue that subjectÕs appearance is not a stable feature for behavior analysis due to its variance to different views, different persons and even different clothes.

3 Local motion histograms

We focus on spatio-temporal volume analysis for human action detection due to its simplicity and effectiveness. Thus the key problem is to characterize actions within volumes by some kinds of features.

3.1 Feature design

Feature design is a signiÞcant problem because informative features capture the essence of a behavior pattern which facilitates our task of detection. Our idea for feature design is derived from three considerations:

1. It is believed that most of the information salient to recognizing action can be captured from the optical ßow, and the appearance of object is less relevant [8], which includes colors and textures, etc.

Fig. 2 Spatio-temporal unit: basic blocks

2. Using histogram-based image descriptors has achieved a remarkable success in object detection on static images among the vast variety of existing approaches [1,10], and the Histogram of Oriented Gradient (HOG) has been proved effective to describe appearance.
3. To simplify the problem, most existing work assumes that the camera and the background are essentially static [2]. In our case, we make efforts to design a detector that could analyze activities where the camera or the background moves around in the scenes.

Based on the above considerations, we make some attempts as follows:

1. We do not combine any appearance descriptors directly into behavior representation, but focus on motion features and make full use of optical flow vectors as descriptors.
2. Local histograms provide effective means to represent visual information. Besides strong descriptive power, invariant to noise or affine transformation and spatially unordered, they have a good property for densely sampled description that all histograms can be efficiently calculated by using an "integral histograms" technique [15]. We use the similar idea as HOG to define local motion histograms.
3. It is noticeable that local motion histograms can be built up from both differentials of optical flow and absolute optical flow orientation. The former description tends to describe relative motion between different parts against moving backgrounds.

To make behavior representation tractable, we define local motion histograms by means of optical flow for capturing motion independent of appearance. To reduce the computational requirements of detection task, we iterate the calculation of integral histograms on the base of volumetric *basic blocks* rather than pixel points. The term of basic block refers to a spatio-temporal region at a spatio-temporal location (x, y, t) with a small size (dx, dy, dt), which is defined as an aggregate

$$\mathcal{B}=\{p \mid x \le x^{(p)} < x+dx,\ y \le y^{(p)} < y+dy,\ t \le t^{(p)} < t+dt\} \quad (1)$$

where p denotes a pixel point at location $(x^{(p)}, y^{(p)}, t^{(p)})$.

As illustrated in Fig. 2, the consecutive optical flow fields are equally divided into grids of basic blocks, and marginal areas are truncated if existing.

To explore the descriptive power of optical flow from each pair of successive video frames, we introduce three kinds of local motion histograms on basic blocks into discovering complementary information. These histograms are based upon magnitude function $M(u, v)$ and discrete eight-orientation function $O(u, v)$:

$$M(u, v) = |(u, v)| \quad (2)$$

$$O(u, v) = \arg\max_{i \in [0,7]} \left(u \cos\left(\frac{i\pi}{4}\right) + v \sin\left(\frac{i\pi}{4}\right)\right) \quad (3)$$

where (u, v) stands for a 2D vector, and $O(u, v)$ optimizes the projection: $\cos\left(\arctan(u, v) - \frac{i\pi}{4}\right) \to \max$.

Intra-block absolute histogram (IAH) When cameras and backgrounds are largely stationary, the original optical flow is sufficient to describe those absolute motions caused by actions of subjects. Even when there exist some camera operations (pan, tilt, or zoom), the original optical flow is still a good clue for discovering salient motions.

In this case, we employ the magnitude and orientation of optical flow vector (U, V) at each sample pixel in a basic block to represent its inside behavior. If magnitudes are taken as weights and orientations are used for angular voting, orientation histograms are created as a descriptor to characterize distributions of optical flow. As shown in Fig. 3a, each histogram is arranged in eight discrete directions, whose bins denote the sum of all contributions along the current direction, shown as the length of arrows. Since the eight-bin histogram accumulates all absolute motions inside a block, we call the resulting descriptor as intra-block absolute histogram (IAH). IAH for a given basic block $\mathcal{B}$ is formulated as

$$\mathrm{IAH}_{\mathcal{B}}(i) = C \sum_{p \in \mathcal{B}} M(U^{(p)}, V^{(p)})\,\delta[i - O(U^{(p)}, V^{(p)})] \quad (4)$$

with

$$C = \frac{1}{\sum_{p \in \mathcal{B}} M(U^{(p)}, V^{(p)})} \quad (5)$$

where $i = 0, 1, \ldots, 7$ is the bin index and δ is the Kronecker delta function. The normalization constant C is imposing the condition $\sum_{i=0}^{7} \mathrm{IAH}_{\mathcal{B}}(i) = 1$.

Fig. 3 Local motion histograms of a plane inside basic blocks: a IAH is based on (U, V); b NDH is based on (U_x, U_y) or (V_x, V_y); c IDH is based on (U_x, V_x) or (U_y, V_y)

Neighbor-point differential histogram (NDH). Obviously, local differential of optical flow cancels out most effects of camera motion, such as pan and tilt, which can even reduce the effects incurred by zoom and rotation. Usually, the differentials are maximal at motion boundaries between a stationary region and a motional one, or two motional regions, which coincide with limb and body edges for human subjects with a behavior of walking or running. In some sense, flow differentials possess some character like edge-based descriptors.

We encode a type of local motion histogram to capture the local orientations of motion boundaries in a simple way. The two flow components U, V are regarded as independent features. Taking local gradients (U_x, U_y) and (V_x, V_y) separately and calculating the corresponding gradient magnitudes and orientations (see Fig. 3b), we obtain weighted votes into local orientation histograms, which results two channels of eight-bin histograms. This resulting descriptors are called as neighbor-point differential histogram (NDH). Similar to the above defined IAH, they are formulated as

$$NDH^U(i) = C^U \sum_{p \in \mathcal{B}} M(U_x^{(p)}, U_y^{(p)})\delta[i - O(U_x^{(p)}, U_y^{(p)})] \quad (6)$$

$$NDH^V(i) = C^V \sum_{p \in \mathcal{B}} M(V_x^{(p)}, V_y^{(p)})\delta[i - O(V_x^{(p)}, V_y^{(p)})] \quad (7)$$

where C^U and C^V are both normalization constants.

Inter-block differential histogram (IDH). The differentials mentioned in the above are calculated from neighboring points, which usually outstands motion boundaries. In order to capture the relative motion of body parts, we exploit inter-block flow differentials. This means we compute flow differentials across neighboring basic blocks. This type of differentials is naturally capable to compensate most effects of global motions caused by camera operations as well. Figure 3c depicts how the differentials are computed.

Since the spatial displacements (dx, dy) between basic blocks are relatively large and dx is not essentially equal to dy, new histograms are encoded on (U_x, V_x) and (U_y, V_y). They are respectively, taken as pairs for orientation calculation and angular voting. The resulting descriptors consisting of two channels of eight-bin histograms are called as inter-block differential histogram (IDH). Similarly, they are formulated as

$$IDH^x(i) = C^x \sum_{p \in \mathcal{B}} M(U_x^{(p)}, V_x^{(p)})\delta[i - O(U_x^{(p)}, V_x^{(p)})] \quad (8)$$

$$IDH^y(i) = C^y \sum_{p \in \mathcal{B}} M(U_y^{(p)}, V_y^{(p)})\delta[i - O(U_y^{(p)}, V_y^{(p)})] \quad (9)$$

where C^x and C^y are both normalization constants. It is worth noting that the gradients U_x, U_y, V_x, and V_y for IDH are computed in a different way from those for NDH although the same symbols are used.

Our proposed local motion histograms bear some similarity to Dalal's motion descriptors [2], where differentials of optical flow are employed. However, there exists some substantial distinct between them. Firstly, our proposed descriptors are spatio-temporal volumetric features and the histograms are accumulated in the normalized action cuboids. Secondly, the defined IDH is simpler than Dalal's internal motion histograms (IMH), which are calculated over basic blocks and only two neighboring blocks are involved rather than eight outer cells. Finally, both absolute values and differentials are used for behavior representation, and NDH and IDH are organized together as complementary features for IAH, not alternative features for each other.

3.2 Behavior representation

Using basic blocks as the elementary unit, we denote any behavior region by grouping adjacent basic blocks into a large spatio-temporal volume. To depict its inside action, the volume is divided into sub-regions with different sizes at different positions. Moreover, to reserve some spatial or

Fig. 4 Four different arrangements of basic blocks

temporal information within a sub-region, the histograms are calculated with different arrangements of basic blocks (see Fig. 4), and are respectively, concatenated into a single feature vector.

The Þve-channel histograms (IAH × 1, NDH × 2, IDH × 2) under four arrangements with different regional sizes at different positions are calculated separately, then a dense representation for an action is built up. This representation leads to a histogram set with a huge size, which makes it impossible to be used for detecting directly. It is essential to Þlter the resulting descriptors and Þnd those representative and discriminative features to characterize an action. Sect. 4 will discuss this in details.

There are two pending problems. One is parameter tuning for calculation of histograms. Basic blocks are given the sizes of $dx, dy = (4, 5, 6)$, $dt = (3, 4, 5)$, and the normalized action volumes have different numbers of units dependent on action types. The other is optical ßow estimation, which is important for building up descriptors. Since optical ßow depends on the temporal and spatial resolution, it is computed on the original resolution and then scaled.

However, dense optical ßow between two consecutive frames is known as a coarse feature, which is noisy and unreliable. Moreover, some methods are going together with local aperture effects. The local gradients (U_x, U_y, V_x, V_y) are used to accumulate the histograms, thus smooth and accurate ßow are preferable for NDH and IDH. Based on these considerations, our initial efforts concentrated on some ÒglobalÓ methods, especially multi-scale non-linear diffusion based algorithm [16], and high-quality results are desired.

Based on the later boosting framework, we compared the performance of this optical ßow estimation with different methods on a dataset described later in this paper. Results are shown in Fig. 5. Almost a same conclusion is drawn as [2] that ProesmansÕs ßow does not provide improved performance while it is computationally expensive. It is over-smoothed which tends to blur the motion boundaries, and in turn, reduces the descriptive power of NDH and IDH.

In contrast, HornÕs ßow [7] is fast to be calculated and preserves the motion boundaries well, which keeps more number of selected features for NDH and IDH, but the Þnal performance does not be promoted apparently. Our explanation is that HornÕs ßow is noisy on some pixels (e.g., spatio-temporal corners), whose U and V get unexpected big values and make histograms unstable.

GennertÕs ßow [6] relaxes the brightness constant assumptions, which sounds appropriate to our framework. In fact,

Fig. 5 Comparison of detecting performance by histograms from various optical ßow estimations on KTH dataset. a PrecisionÐrecall curves for detecting ÒwalkingÓ action (real mode); b number of selected features of a; c precisionÐrecall curves for detecting ÒboxingÓ action (real mode); d number of selected features of c. In b and d the total number of features are both 50 it obtains a stable ßow Þeld, but it naturally outstands the moving boundary and despises inner ßow. As a result, it selects relatively less number of NDH and IDH. This method working with our framework leads to the worst performance.

A saturated HornÕs ßow (S-Horn) is introduced into computing motion descriptors, which originates from very simple idea: the magnitude of HornÕs motion vector is saturated by its neighbors, if $M(u, v) > \frac{\gamma}{8}\sum_{i=0}^{7} M_i(u, v)$, where γ is a experiential factor ($\gamma = 3$ is recommended). It is amazing that such a small modiÞcation enhances the performance apparently. This ßow will be used to compute histograms in all the later experiments.

4 Learning models for histogram-based features

Given a video database involving human actions with labeled positive and negative samples, we are required to classify or detect behaviors in a novel sequence. As outlined above, we Þrst construct local motion histograms for behavior representation.

A naive, nevertheless simple method for classiÞcation is to Þnd the best match to the querying motion descriptor, which can be performed in a KNN framework. In such a case, we must provide an appropriate similarity measure for our proposed motion descriptors. This is really difÞcult problem for our method because our descriptor set is composed of several types of histograms, and it is hard to give the weights for them. In addition, the dense representation leads to a large number of features, which makes the computational cost expensive.

Learning discriminative action models for boosting classiÞers does not need any form of similarity measure, but does well in good feature selection. It is also a promising approach to handle the variation within an action category. However, most existing boosting frameworks have been employed to select one-dimensional features, such as Haar-like features, where an efÞcient classiÞer can be found by selecting an optimal decision threshold. Our histogram-based descriptors are multi-dimensional features with different types and different arrangements, it is difÞcult to Þnd a proper decision threshold.

In this paper, linear projection technique is utilized to deal with histogram-based features. The classiÞcation task is not directly converted from a multi-dimensional problem to a one-dimensional one, but weak classiÞers for boosting are learned on unprocessed histograms via this technique. During the learning process, Gentle AdaBoost will be employed as a strong learner to select the position, the size, the type, and the arrangement of our local descriptors in action volumes. The only criterion is to minimize the training error for the samples.

4.1 Boosting framework

Generally, boosting provides a simple way to approximate additive models of the form

$$H(x) = \sum_{m=1}^{M} \beta_m h_m(x) \qquad (10)$$

where $H(x)$ is called as a strong learner, x is the input feature vector with a class label $y \in (+1, -1)$, and M is the number of boosting rounds. The functions $h_m(x)$, also written as $b(x; \gamma_m)$, are base classiÞers which are usually simple functions of x. The expansion coefÞcients β_m and the parameters γ_m are jointly Þt to training data in a forward Òstage-wiseÓ manner.

All AdaBoost-based techniques can be considered as a greedy optimization method for minimizing exponential error function

$$J[H(x)] = E(e^{-y \cdot H(x)}) \qquad (11)$$

where the term $yH(x)$ is related to the generalization error (out-of-sample error rate), and called as ÒmarginÓ.

For binary classiÞcation problems, there are two versions of the most commonly used AdaBoost procedures: One is ÒDiscrete AdaboostÓ, where each $h_m(x)$ is a classiÞer producing values $(+1, -1)$ and β_m are constants, the corresponding prediction is $\mathrm{sign}(H(x))$. The other is ÒReal AdaboostÓ, where real-valued predictions $h_m(x)$ are combined with β_m and absorb β_m in Eq. (10) with a simpler form. The sign of each $h_m(x)$ gives the classiÞcation, and the value of each $|h_m(x)|$ is a measure of the ÒconÞdenceÓ in the prediction.

Gentle AdaBoost (GAB) is a more robust and stable version of real AdaBoost [4], which has ever been the most practically efÞcient boosting algorithm used in object detector [13]. In this paper Gentle AdaBoost is used to select our histogram-based features.

Gentle AdaBoost takes adaptive Newton steps to minimize error $J[H(x) + h_m(x)]$ by

$$H(x) \quad H(x) + \frac{E[e^{-yH(x)} y|x]}{E[e^{-yH(x)}|x]} = H(x) + E_w(y|x). \qquad (12)$$

Equivalently, the weak hypothesis is written as

$$h_m(x) = E_w(y|x). \qquad (13)$$

Here $E_w(.|x)$ refers to a weighted conditional expectation. The weight $w(x, y) = e^{-yH(x)}$ is updated by

$$w(x, y) \quad w(x, y)e^{-yh_m(x)}. \qquad (14)$$

To get optimized $h_m(x)$, we expand $J[H(x) + h_m(x)]$ to the second order about $h_m(x) = 0$. Minimizing pointwise with respect to $h_m(x)$, there is

$$\widehat{h}_m(x) = \arg\min_{h_m}(E_w[(y - h_m(x))^2|x]) \qquad (15)$$

which gives the way to select a trained base classifier $h_m(x)$ and produce a weak classification rule.

Table 1 Accuracy (%) of different feature combinations for detecting "walking" (lite mode) on the KTH dataset

Number of features	1	5	10	20	35	55	80
IAH	27	45	51	50	54	57	57
NDH	42	53	57	63	65	69	69
IDH	40	52	59	67	69	71	71
IAH + NDH	35	57	80	89	91	93	94
NDH + IDH	44	65	76	87	90	95	95
IAH + IDH	40	63	84	86	91	92	91
IAH + NDH + IDH	51	69	88	90	92	96	97

Table 2 Accuracy (%) of different feature combinations for detecting "boxing" (lite mode) on the KTH dataset

Number of features	1	5	10	20	35	55	80
IAH	18	35	42	39	44	44	43
NDH	31	52	56	56	55	56	56
IDH	37	54	65	67	67	67	67
IAH + NDH	31	42	54	70	79	79	79
NDH + IDH	39	62	72	85	91	91	92
IAH + IDH	38	57	68	82	89	87	88
IAH + NDH + IDH	39	65	77	87	91	95	95

4.2 Weak learner

Although Adaboost doesn't have strict requirement for the choice of weak learners, effective weak learners tends to enhance the performance of the final classifier. Available weak learners are usually Classification and Regression Trees (CART). Whereas, motivated by Laptev's work [10], we use Weighted Fischer Linear Discriminant (WFLD) as a weak learner for multi-valued histogram features, where multi-dimensional features are projected onto a pre-defined set of one-dimensional manifolds using a fixed set of functions. The weak learner is defined as

$$h(x) = w^T x + b, \quad \text{with } w = (S_1 + S_2)^{-1}(\mu_1 - \mu_2) \qquad (16)$$

where μ_1, μ_2 stand for the weighted class means and S_1, S_2 for the weighted class covariance matrices, and b is the threshold obtained by projecting total means with a negative w. Given the weights $\{w_i\}$ corresponding to samples $\{x_i\}$, the matrices have the form

$$\mu = \frac{1}{n \cdot \sum w_i} \sum_i^n w_i \, f(z_i) \qquad (17)$$

$$S = \frac{1}{(n - 1) \cdot \sum w_i^2} \sum_i^n w_i^2 (f(z_i) - \mu) \quad (f(z_i) - \mu)^T. \qquad (18)$$

WFLD seeks for finding the optimized projecting directions which are efficient for discrimination. The resulting linear projection transformation yields the maximum ratio of between-class scatter to within-class scatter for weighted samples. In comparison with CART, each trained WFLD produces a more compact classification rule, which leads to a more efficient boosting classifier.

4.3 Feature selection with GAB + WLFD

As mentioned before, the proposed five-channel histograms (IAH × 1, NDH × 2, IDH × 2) build up a dense representation for an action with a huge histogram set. Intuitively, each kind of feature delivers different data semantics for an "atomic" action. However, it keeps unknown whether they indeed carry a different amount of discriminative information for action classification. Using the GAB+WLFD algorithm, feature selection is used to explore the useful histograms for our application.

To understand the added value of these introduced descriptors, results are reported separately for different features and in combination with each other. Tables 1 and 2 list the performance independently. From the two tables, the best accuracies of detection are obtained by about 50 features. For each individual feature, it reaches the best accuracy within 30 features.

The two actions involving in these comparisons hold different moving nature, "walking" action has a relatively stable moving region, and bears relative motions between body parts, while most "boxing" actions have a largely stationary region, the actions focus on the movement from hands. The results in the two tables coincide with our intuition that IDH shows great importance on catching motions between parts, and in most cases, NDH working with IDH achieves a satisfactory performance without need of IAH, especially when the moving region has a large size or with simple motions on the 2D plane.

5 Experiments

5.1 Datasets

To evaluate the performance of the proposed method, we conducted experiments on three different datasets: Facial Expression dataset (FE) [3], KTH human activity dataset (KTH) [18], and Drinking and Smoking dataset (D&S) [12]. They are all publicly available online, and some sample images from them are shown in Fig. 6.

The face data in the FE dataset involves two subjects, each expressing six emotions under two lighting setups. Under

Fig. 6 Sample images from some public datasets

each lighting setup, each individual was asked to repeat six different expressions eight times. The expressions are anger, disgust, fear, joy, sadness and surprise. The subject always start with a neutral express, expresses an emotion, and returns to neutral, all in about 60 frames.

The human action data from the KTH dataset contains 25 individuals, each engaged in the following activities: walking, jogging, running, boxing, hand clapping, and hand waving. Each person repeats the six actions in each of four scenarios: outdoor, outdoor with camera zooming, outdoor wearing different clothes, indoor. And their durations are about 20 s.

The D&S dataset consists of annotations for two action classes ÒDrinkingÓ and ÒsmokingÓ in the movies ÒCoffee and CigarettesÓ (2003), ÒSea of loveÓ (1989) and in the ÒdrinkingÓ sequence recorded by INRIA/Vista. In contrast to most existing datasets, the actions in this dataset are not recorded in controlled or simpliÞed setting with simple backgrounds, but in comprehensive and realistic scenes with different subjects and from different view points. Although the two kinds of actions are ÒatomicÓ actions with a reasonable well-deÞned structure in time, the annotated samples possess large variability of scales, locations, and surroundings.

5.2 Training and testing settings

When annotating all training samples for detection it is extremely important to align them in both spatial and temporal domains. Similar to the well-annotated D&S dataset, we discard those original coarse frame-based annotations for the KTH dataset, and manually segment their action instances to a Þne degree for training. In the spatial domain, actions are labeled by rectangles or squares which keep the subject in the middle-center (boxing, hand clapping, etc.) or cover the active regions (walking, running, etc.). In the temporal domain, all sorts of actions are deÞned to start at the same phrase, and end at the same phrase as well (see Fig. 7), although acting in various ways with different durations. Even more, we split periodic actions in the KTH dataset into non-periodic actions (i.e., ÒatomicÓ actions), which considerably simpliÞes the annotation with temporal alignment.

Besides alignment, normalization on all training samples is another key technique when training our detectors. In the spatial domain, each frame from a given action volume is resized into a preset size by interpolation on pixels. In most cases, redundant regions or blanks are added in order to maintain aspect ratios. Furthermore, temporal interpolation is not utilized in our method because behavioral durations are good features for action description. In this way, the action volumes are normalized with same spatial size but different temporal lengths, and our detectors will learn to generalize the noisy durations of actions.

The Þrst set of experiments examines our methodÕs ability to classify actions in short segmented video clips. We experiment on the FE dataset whose samples for both training and testing are well labeled. Specially, training is done on a single subject under only one lighting setup. Finally we test the model on three testing sets: (1) the same subject under different illuminations; (2) different subjects under the same illumination; (3) different subjects under different illuminations.

The second set of experiments evaluates our methodÕs performance in classifying actions in unsegmented testing video sequences by taking a detection measure. For the KTH dataset different actions tend to have distinct durations, such as ÒrunningÓ and ÒwalkingÓ, and testing our models on segmented samples will greatly simplify the problem and make our results unable to be compared directly. Hence for a given testing sequence we run our resulting six detectors, respectively, and summarize the detecting likelihoods to categorize the entire sequence with the highest likelihood. Particularly when the highest likelihood is below our threshold, we label the sequences as Òunknown actionsÓ.

In our experiments, training and testing are done on extracted about 1/4 length of original sequences in the KTH dataset with two modes. (1) Lite mode: Under two scenarios (outdoor, outdoor wearing different clothes), we use a light-weighted subset of the dataset which includes six subjects for training and six different subjects for testing; (2) Real Mode: We use the same training and testing sequences as in previous works [8,18]. Namely, under each of four scenarios eight subjects in the training set and nine subjects in the testing set.

The third set of experiments evaluates our methodÕs performance in detecting actions in long video sequences. We train a model for the class ÒdrinkingÓ on the D&S dataset. For comparison we use the same training set and testing set as in LaptevÕs work [12]. Namely, for training we use 106

Fig. 7 Frames from two Òhand clappingÓ actions after spatial and temporal alignment: top frames from #0 to #24 (dt = 24) constitute an action volume; bottom frames from #54 to #90 (dt = 36) constitute an action volume

Fig. 8 Results on the FE dataset. a compare with the best existing results; b confusion matrix for the same subject under different illuminations; c confusion matrix for different subjects under the same illumination; d confusion matrix for different subjects under different illuminations

drinking samples selected from all the video clips as positive samples, and for testing we scan three episodes containing 38 drinking actions. The training and the testing sets have no overlap in subjects or scenes.

For the first set of experiments, we learn a multi-class model for the classifying problem on the FE dataset, and each action class has the same number of samples. On the other hand, we train binary classifiers for the detecting tasks in the later two sets of experiments. In addition, around the annotated positive samples we choose random durations of clips at random positions with random scales to generate negative samples for training binary classifiers.

5.3 Results

For the FE dataset, we compare the overall classification error rates with the best results of Dollár's work [3] in Fig. 8a. It is obvious that our method performs better under changes of illuminations (case 2 and case 3). Furthermore, the confusion matrices on three testing sets are presented in Fig. 8b–d, which show large confusion between similar actions “anger” and “disgust”.

For the KTH dataset we summarize the recognition rate through those detecting results, and give the confusion matrices for the two modes in Fig. 9c, d. It shows large confusion between “hand clapping” and “boxing”, as well as “running” and “jogging”. This is consistent with our intuition that those actions involve similar hand motions or similar leg motions. The results for the two modes depict that outdoor with camera zooming and indoor scenarios markedly degenerate our method's performance. A reasonable explanation is that various resolutions and changes of illuminations cause a considerable side effect in calculating optical flow.

Table 3 compares classifying accuracy with previous studies on the same dataset KTH. Since the experimental setting of most existing studies are not the same as ours, the results cannot be compared directly and the listing accuracies do not

Fig. 9 Results on the KTH dataset: a classiÞcation for ÒwalkingÓ actions, where the lighted boxes mark correct decisions by the ÒwalkingÓ detector and the darkened boxes mark false decision by ÒjoggingÓ detector; b several detections obtained by the Òhand wavingÓ detector, where the last two darkened boxes mean false detections for Òhand clappingÓ; c, d confusion matrices for two testing modes

mean too much. For example, KimÕs method [9] has achieved an impressive accuracy at 95%, but spaceÐtime alignment of actions is manually done. In fact, a more challenging work is done in this paper, where the classifying problem is treated as an automatic detecting problem of multiple actions from unsegmented testing sequences. Although our detecting method does not speciÞcally aim for whole sequence classiÞcation, the accuracy obtained by our method is competitive.

For the D&S dataset we detect human actions in realistic scenarios [1] with variation in subjects and scenes, etc. In Fig. 10b precisionÐrecall curves and average precision

[1] Demo video: http://ccs.sjtu.edu.cn/673/ßd3mh

Table 3 Comparison of classifying accuracy on the KTH dataset

Related studies	Accuracy (%)
Our method (IAH + NDH + IDH)	85.1
Ke et al. [8]	63.0
Schuldt et al. [18]	71.7
Doll‡r et al. [3]	81.2
Niebles et al. [14]	81.5
Wong et al. [21] (pLSA-ISM)	83.9
Wong et al. [21] (WX-SVM)	91.6
Kim et al. [9]	95.3

Fig. 10 Results on the D&S dataset: a the Þrst nine detections on the testing set obtained by our ÒdrinkingÓ detector, where the second box in the Þrst row and the Þrst box in the second row are false detections; b comparison of precisionÐrecall curves in ÒdrinkingÓ action detection task

(AP) values illustrate the detecting performance on ÒdrinkingÓ action. Our method outperforms LaptevÕs best result [12] (OF5 with Keyframe priming) with a better average precision, and our method tends to perform better in rejecting similar non-drinking actions.

From all above experiments, we observe that the combination of different motion descriptors plays an important role in action recognition. It is found that without any explicit appearance of shape information, human actions under clutter background or moving background can also be well characterized by the local motion histograms. At the same time, Gentle AdaBoost is proved to be powerful enough to select parameters for classiÞers.

6 Conclusions

In this paper we addressed human action detection in realistic scenarios. Our method shows great potentials in action representation within a spatio-temporal volume. The extracted histogram-based descriptors act as complementary information for each other. The Gentle Adaboost framework working with WFLD is proved to be able to select discriminative histogram-based features and learn robust and efÞcient detectors.

Our method is tested in a number of experiments against well established algorithms, and all experimental tests show the satisfying results. Despite of different appearance, scale changes, clutter background or moving background, our results on classifying and detecting problems are comparable with the previous studies, or outperform the previous results.

The experimental results not only validate the effectiveness of our method, but also prove that without the support of appearance and shape information, only motion information is capable of describing human actions well. However, since we have not combined any appearance or shape descriptors with local motion histograms, it is unknown whether the combination will improve the performance. This is part of our future work.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. Comput. Vis. Pattern Recognit. 2, 886Ð893 (2005)
2. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of ßow and appearance. Eur. Conf. Comput. Vis. 2, 428Ð441 (2006)
3. Doll‡r, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS, pp. 65Ð72 (2005)
4. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: A statistical view of boosting. Ann. Stat. 38(2), 337Ð374 (2000)
5. Gavrila, D.M.: The visual analysis of human movement: A survey. Comput. Vis. Image Underst. 73(1), 82Ð98 (1999)
6. Gennert, M.A., Negahdaripour, S.: Relaxing the brightness constancy assumption in computing optical ßow. A.I. Memo, p. 975. MIT Press, Cambridge (1987)

7. Horn, B.K., Schunck, B.G.: Determining optical ßow. Artif. Intell. 17, 185Ð203 (1981)
8. Ke, Y., Sukthankar, R., Hebert, M.: EfÞcient visual event detection using volumetric features. IEEE Int. Conf. Comput. Vis. 1, 166Ð173 (2005)
9. Kim, T.K., Wong, S.F., Cipolla, R.: Tensor canonical correlation analysis for action classiÞcation. In: CVPR (2007)
10. Laptev, I.: Improvements of object detection using boosted histograms. In: BMVC, vol. 3, pp. 949Ð958 (2006)
11. Laptev, I., Lindeberg, T.: Space-time interest points. IEEE Int. Conf. Comput. Vis1, 432Ð439 (2003)
12. Laptev, I., PŽrez, P.: Retrieving actions in movies. IEEE Int. Conf. Comput. Vis., pp. 432Ð439 (2007)
13. Lienhart, R., Kuranov, A., Pisarevsky, V.: Empirical analysis of detection cascades of boosted classiÞers for rapid object detection. MRL technical report (2002)
14. Niebles, J.C., Wang, H., Li, F.F.: Unsupervised learning of human action categories using spatial-temporal words. In: BMVC (2006)
15. Porikli, F.: Integral histogram: A fast way to extract histograms in cartesian spaces. Comput. Vis. Pattern Recognit. 1, 829Ð836 (2005)
16. Proesmans, M., Gool, L.J.V., Pauwels, E.J., Oosterlinck, A.: Determination of optical ßow and its discontinuities using non-linear diffusion. In: European Conference on Computer Vision, pp. 295Ð304. Springer, London (1994)
17. Ramanan, D., Forsyth, D.A.: Automatic annotation of everyday movements. In: Advances in Neural Information Processing Systems, vol. 16. MIT Press, Cambridge (2004)
18. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. Int. Conf. Pattern Recognit. 3, 32Ð36 (2004)
19. Shah, M., Jain, R.: Motion-Based Recognition. Computational Imaging and Vision Series. Kluwer, Dordrecht (1997)
20. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. Comput. Vis. Pattern Recognit. 1, 511Ð518 (2001)
21. Wong, S.F., Kim, T.K., Cipolla, R.: Learning motion categories using both semantic and structural information. In: CVPR (2007)
22. Yilmaz, A., Shah, M.: Recognizing human actions in videos acquired by uncalibrated moving cameras. IEEE Int. Conf. Comput. Vis. 1, 150Ð157 (2005)

Author biographies

Q. Luo received the B.S. and M.E. degrees in ship engineering from Wuhan University of Technology, China, in 2000 and 2003, respectively. Currently he is a Ph.D. candidate at the Department of Electronic Engineering in Shanghai Jiaotong University, China. His research interests include computer vision, video retrieval and video understanding. He is working on projects involving video surveillance, activity analysis in intelligent video systems.

X. Kong received the B.S. and M.E. degrees in computer science and automation from Anhui University, China, in 2000 and 2004, respectively. Currently he is undertaking a Ph.D. degree at the Department of Electronic Engineering in Shanghai Jiaotong University, China. His research interests include content-based image retrieval, pattern recognition, image processing and intelligent video surveillance.

G. Zeng received the B.S. and M.E. degrees in electronic engineering from Xidian University, China, in 1988 and 1991, respectively, and Ph.D. degrees from Shanghai optic-machine center, in 1997. From 1997 to 1999, he was a postdoctoral researcher at the State Key Laboratory of ISN in Xidian University, China. From 2001 to 2002, he was a research scientist in the University of Albert-Ludwigs Freiburg, Germany. Now he is a professor at the Department of Electronic Engineering in Shanghai Jiaotong University, China. His research interests include video content computing, indexing and security, quantum communication and quantum identity veriÞcation.

J. Fan received the M.S. degree in theory physics from Northwestern University, Xian, China, in 1994 and the Ph.D. degree in optical storage and computer science from Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, China, in 1997. He was a Researcher at Fudan University, Shanghai, during 1998. From 1998 to 1999, he was a Researcher with Japan Society of Promotion of Science (JSPS), Department of Information System Engineering, Osaka University, Osaka, Japan. From September 1999 to 2001, he was a Researcher in the Department of Computer Science, Purdue University, West Lafayette, IN. In 2001, he joined the Department of Computer Science, University of North Carolina at Charlotte as an Assistant Professor and then became Associate Professor. His research interests include content-based image/video analysis, classiÞcation and retrieval, surveillance videos, and statistical machine learning.