Human Action Detection by Boosting Efficient Motion Features

Ming Yang, Fengjun Lv, Wei Xu, Kai Yu, Yihong Gong
NEC Laboratories America, Inc.
10080 North Wolfe Road, SW-350, Cupertino, CA 95014
{myang,flv,xw,kyu,ygong}@sv.nec-labs.com

Abstract

Recent years have witnessed significant progress in detection of basic human actions. However, most existing methods rely on assumptions such as known spatial locations and temporal segmentations or employ very computationally expensive approaches such as sliding window search through a spatio-temporal volume. It is difficult for such methods to scale up to handle the challenges in real applications such as video surveillance.

In this paper, we present an efficient and practical approach to detecting basic human actions, such as making cell phone calls, putting down objects, and hand-pointing, which has been extensively tested on the challenging 2008 TRECVID surveillance event detection dataset . We propose a novel action representation scheme using a set of motion edge history images, which not only encodes both shape and motion patterns of actions without relying on precise alignment of human figures, but also facilitates learning of fast tree-structured boosting classifiers. Our approach is robust with respect to cluttered background as well as scale and viewpoint changes. It is also computationally efficient by taking advantage of human detection and tracking to reduce the searching space. We demonstrate promising results on the 50-hour TRECVID development set as well as two other widely-used benchmark datasets of action recognition, i.e. *the KTH dataset and the Weizmann dataset.*

1. Introduction

Detecting human actions from a monocular video is an important task for many emerging applications such as advanced video surveillance and intelligent video content analysis. Recently, we have seen huge advances in action detection or recognition in well-controlled (*e.g.* laboratory or studio-like) environment [26, 4, 27, 32, 2, 10, 22, 9, 6, 25, 30] or in movie or sports videos [5, 11, 13, 19, 24]. Not surprisingly, in order to make this difficult problem more tractable, most of the existing approaches have made assumptions such as known spatial locations and temporal segmentations of actions, no (or very little) scale and viewpoint changes, as well as static and clean background, so that human figures can be reliably extracted and aligned.

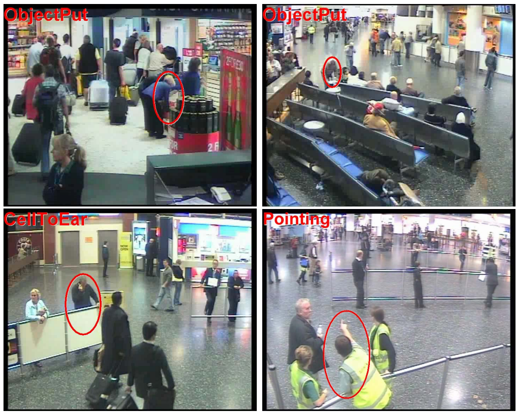


Figure 1. Sample actions of interests in 4 different camera views in the 2008 TRECVID surveillance event detection dataset.

Unfortunately, these assumptions seldom hold in real-world videos such as surveillance videos. In addition, when being recorded in a staged environment, people tend to act very carefully and thus not as naturally as in everyday life. Not to mention that it is very difficult to capture enough variance of the same action performed by various people as in our real life. This indicates that there is a large gap between the existing research efforts and the challenges we are facing in real applications.

To our best knowledge, 2008 TREC Video Retrieval Evaluation (TRECVID 2008) [21] has made the largest effort so far to bridge this gap by providing an extensive 99-hour surveillance video dataset recorded in London Gatwick Airport. As shown in Fig. 1, the highly crowded scenes, the severely cluttered background with noticeable reflections and shadows, the large variance in viewpoints and people subjects (and thus the action execution styles), and the huge amount of data to analyze, combined together, make action detection on this dataset a formidably challenging task. As far as we know, human action detection performance on such a challenging dataset with these practical concerns has been barely evaluated and reported before.

2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops
978-1-4244-4441-0/09/$25.00 ©2009 IEEE

All these challenges have to be addressed jointly in a practical action detection approach in which action representations and search scheme shall be able to tolerate enormous variance and be computationally feasible as well. In this paper, we propose a novel action representation based on a set of motion edge history images (MEHIs), which not only effectively encodes both shape and motion patterns of actions but also facilitates learning of fast tree-structured boosting classifiers. Specifically, after detecting edges from frame difference images, the motion edges on objects' boundaries are accumulated with different forgetting factors incrementally to build a set of 2D motion edge history images. These MEHIs only preserve shapes of moving objects and record their non-local temporal orders, thus, they are inherently insensitive to the versatility of individual's appearances. Furthermore, it is very efficient to calculate 2D Haar features on the MEHIs and learn a variant of probabilistic boosting tree classifiers [29, 28] for various actions. This is tantamount to analyze multiple space-time volumes with different lengths simultaneously, thus we obtain a time invariant action model in some sense. Exhaustive search throughout a space-time volume to evaluate all hypotheses [27, 2, 11] is computational prohibitive for analyzing long-duration videos with high resolution, *e.g.* 50 hours of 720 × 576 videos. Therefore, we take advantage of human detection and tracking to reduce the searching space so as to build a complete action detection system.

The main merit of the proposed method is the balance between the need for discriminative power and computational efficiency. The motion edge history images are inherently not sensitive to appearance variations or clutters. Thus, by boosting 2D Haar feature responses from multiple MEHIs, the classifiers can largely learn the motion patterns of different action categories, which relieves the demands for precise temporal or spatial alignment of human figures. This set of MEHIs support analyzing different actions at multiple locations and scales simultaneously. Further, leveraging human detection and tracking to identify candidate regions makes the system computationally feasible. The feature extraction and classification modules run at 20 fps for videos with resolution 720 × 576 and the entire system runs at 2 fps including human detection and tracking.

Actions of interest are generally application dependent [25]. In this paper, we consider *basic actions* that are articulated motions of a single human body which cannot be easily decomposed to simpler actions. In particular, we focus on 3 required actions in the 2008 TRECVID Surveillance Event Detection Evaluation: *CellToEar*, *ObjectPut*, and *Pointing* [1]. On average, the proposed method achieves detection rate 9.73% *vs.* false positive rate 1%. The system incorporating the proposed method as one of the major components is among the top performers in the TRECVID evaluation. To compare with previous algorithms and show the generalization ability, we also evaluate our method on two widely-used benchmark datasets, *i.e.* the KTH [26] and the Weizmann [2] datasets, and demonstrate competitive performance.

2. Related Work

The key of action detection is how to represent actions. Explicit inference of human poses or articulated body motion are supposed to be very helpful, however, they are hard problems themselves, if not even harder. So, we restrict to review related work where actions are regarded as patterns in 3D space-time volumes. Apparently, actions are fundamentally different from rigid 3D objects in that variations in the time domain are typically much larger than that in the spatial domain. In terms of different strategies to organize features in space-time volumes, the action representations in existing approaches can be mainly summarized into 4 categories: 1) a graphical model of key poses or examplars; 2) a holistic space-time template; 3) a bag-of-words model of sparse space-time interest points; and 4) a vast pool of spatio-temporal features. Their respective strengths and concerns are discussed as follows.

As the execution speed of the same type of actions may vary, it is natural to model actions using a hidden Markov model (HMM) [8, 18, 31] or a Conditional Random Field (CRF) model [20] of key poses and recognize actions by inferring the hidden pose sequences. Since direct inference of poses is difficult, these approaches resort to 3D human models to synthesize 2D projections and estimate the key poses by matching with silhouettes extracted from input videos. By employing this analysis-by-synthesis idea, these action representations are view-invariant to some extent.

By representing actions as holistic space-time templates based on their global spatio-temporal characteristics, action detection can be formulated as querying the nearest neighbors of template actions. Given a figure-centric sequence, the action templates can be motion-energy and motion-history images [3], half-wave rectified optical flow fields [5], or space-time gradients of many small space-time patches [27]. Alternatively, actions can be delineated by geometrical properties of 3D templates, *e.g.* stacked silhouettes of human figures, analyzed by the Poisson equation [2] or differential geometric surfaces [32]. As a retrieval task, a single action template may be sufficient to detect similar actions. However, the computation is intensive in these approaches even assuming one fixed scale of all actions.

The bag-of-words paradigm is quite successful in object and scene categorization. When it is applied to action detection, SIFT-like descriptors [17] are extracted around space-time interest points, then actions are abstracted by histograms of a vocabulary of space-time visual words. For

[1] The detailed definitions of these actions are elaborated in www.nist.gov/speech/tests/trecvid/2008/doc/TRECVid08_Guidelines_v1.6.pdf

instance, the space-time interest points include 3D Harris corners [12, 26] or space-time structures that show strong responses to Gabor-like spatio-temporal filtering [4]. Along this line of research, many bag-of-words approaches [23, 22, 16, 19] demonstrate superb performance in recognizing actions in clean background. In practice the interest point detection may be sensitive to clutters and scale variations.

Actions can also be represented by discriminative models, *e.g.* SVM or Adaboost classifiers, based on a vast pool of spatio-temporal features. [10] extracts 3D Haar features from optical flow fields in a fixed-size cube, which is extended to extract histograms of orientations of spatial gradients and optical flow in [13]. Motivated by biological vision systems, [9] computes optical flow and Gabor filtering over video segments, then the maximal responses of correlation with thousands of small templates are used to learn SVM classifiers. [6] first learns Adaboost classifiers from many small cuboids based on pixel-level optical flow responses [5], then utilizes these classifiers as mid-level features to train a higher level Adaboost classifier. Generally, these approaches require a large number of training samples to obtain good generalization performance.

Our method can be categorized to the 4th category. We learn discriminative boosting tree classifiers [28] from 2D Haar feature responses on a set of motion edge history images. The motion features are inspired by motion history images [3], however, we maintain multiple MEHIs with various forgetting factors and extract sparse responses rather than using them as templates. Therefore, our motion features are time invariant to some extent and do not rely on accurate spatial alignment. Our work is also related to [10, 13]. Besides using new motion features, we evaluate 2D Haar features so that the method is less memory consuming than using integral videos [10] or integral video histograms [13].

3. Our Approach

3.1. Overview of our approach

Our approach intends to learn discriminant models using efficient motion features for individual actions. Given the grey-level input video sequences, we incrementally update a set of motion edge history images, meanwhile, locate the candidate regions to analyze by human detection and tracking. For each detected human, an enlarged region around the tracked head is cropped from all MEHIs. Then, a large number of 2D Haar features are extracted to train a one-against-all boosting tree classifier [28] for each action category of interest. During testing, only the Haar features in the learned classifier need to be evaluated. Note though human detection and tracking results are available, we do not align individual human figures accordingly.

Specifically, for consecutive frames, we first calculate the frame difference image which retains motion information only, then perform edge detection on it to extract the approximate shapes of moving objects. Afterwards, the motion edges are accumulated to a set of history images with various forgetting factors. This kind of motion features is a tradeoff between retaining all relevant information and efficient processing. On one hand, a set of MEHIs preserves more motion information than pure frame-based features, on the other hand, to analyze multiple 2D images is computationally more affordable than analyzing spatio-temporal volumes. Further, we extract 2D Haar features from these MEHIs in enlarged neighborhoods of tracked human heads to train a boosting tree classifier. The block diagram is shown in Fig. 2. Throughout the paper, the grey-level MEHIs are drawn with pseudo colors.

3.2. Feature extraction

As appearances of individuals performing actions may vary dramatically, we intend to extract features only related to the shape and motion patterns and ignore appearance information as much as possible. Denote the grey-level intensity of a pixel at (x, y) in the t-th input frame by $I_t(x, y)$. The frame difference $D_t(x, y)$ of the consecutive frames is calculated by thresholding $d_t(x, y) = |I_t(x, y) - I_{t-1}(x, y)|$ with a conservative threshold T_d ($T_d = 10$ in all our experiments), as

$$D_t(x,y) = \begin{cases} 0 & \text{if } d_t(x,y) \le T_d \\ d_t(x,y) & \text{if } d_t(x,y) > T_d \end{cases} . \quad (1)$$

Then, a binary motion edge image $M_t(x, y)$ is obtained by performing Canny edge detection on $D_t(x, y)$. $M_t(x, y)$ preserves the approximate shapes of moving objects. In order to be insensitive to varying execution speeds of the same type of actions, we maintain a set of N motion edge history images $H_t^i(x, y)$ with different forgetting factors $\alpha_i < 1$ ($N = 4$ and $\alpha = \{0.6, 0.7, 0.8, 0.9\}$ in our experiments), as

$$H_t^i(x,y) = \begin{cases} 1 & \text{if } M_t(x,y) > 0 \\ \alpha_i H_{t-1}^i(x,y) & \text{if } M_t(x,y) = 0 \end{cases} . \quad (2)$$

Thus, $H_t^i(x, y)$ is a grey-level image where larger intensity indicates some motion occurs more recently. These MEHIs are incrementally updated on-the-fly, as shown in Fig. 2.

MEHIs are efficient to calculate and less sensitive to noise and clutters than optical flow. In addition, conventional optical flow only preserves local temporal information between a pair of frames, but MEHIs retain motion information within a variable of relative long periods. Sample MEHIs of 3 different actions are shown in Fig. 3.

3.3. Classification

For action detection, the space of negative samples is extremely large including both persons not performing actions of interests and all non-human background regions, which

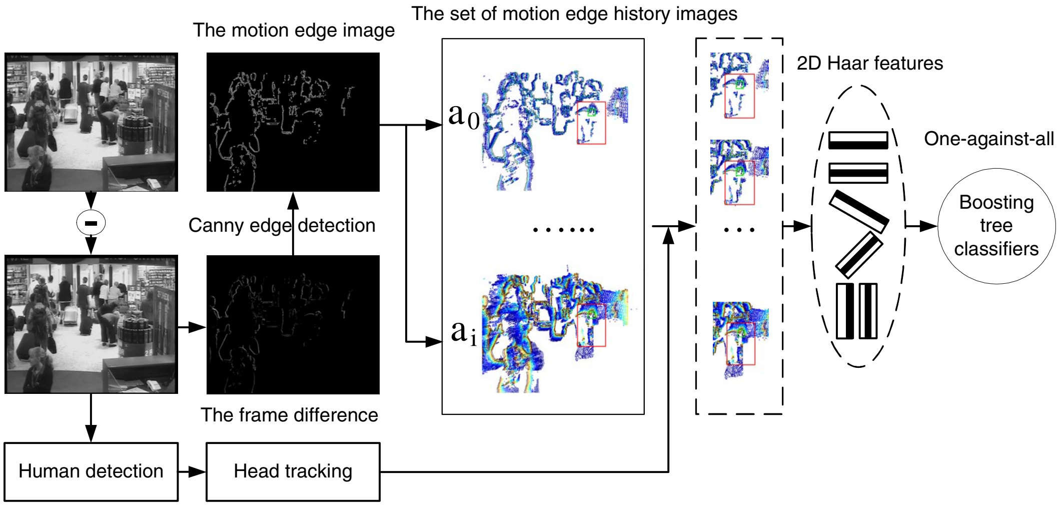


Figure 2. The block diagram of the proposed human action detection approach.

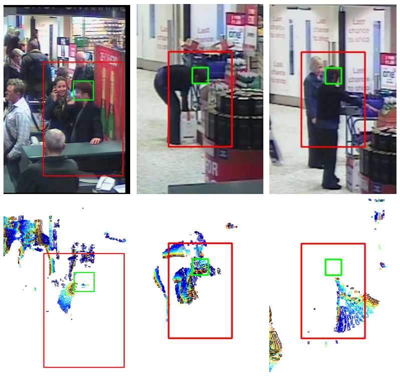


Figure 3. Sample MEHIs for the actions: *CellToEar*, *ObjectPut*, and *Pointing* from left to right ($\alpha = 0.8$). The green rectangle indicates the tracked human head and the red rectangle indicates the region used to analyze.

is an issue for learning discriminative classifiers and even collecting negative training samples. In fact, distinguishing human from non-human regions in images has been investigated intensively for decades. So, it is a natural choice to limit the candidate regions by taking advantage of human detection and tracking. By this means, not only the number of candidate regions is significantly reduced but also the tough requirements of collecting sufficient negative samples are mitigated to make this learning task tractable.

The positive samples of an action could appear quite different from diverse view angles, as shown in Fig. 6. To account for these intra-class variations, we follow the divide-and-conquer strategy to train a variant of probabilistic boosting-tree classifier [28]. Specifically, given the human head of a positive sample $\mathbf{X}$, we crop an enlarged region, *e.g.* 4×6 size of the head for the TRECVID dataset or 4×8 size of the head for the KTH and Weizmann datasets, and normalize it to 40×40 conceptually. A large number of 2D Haar features $h_j(H_t^i)$ are extracted from all MEHIs using integral image technique to train an Adaboost classifier C, *i.e.* a weighted sum of decision stumps $f_j(\cdot)$ on $h_j(H_t^i)$,

$$C(\mathbf{X}; H_t^1, \cdots, H_t^N) = \sum_{j=1}^{M} \beta_j f_j(h_j(H_t^i)), \tag{3}$$

where M is the number of features selected with the weight β_j. Then, $\mathbf{X}$ is divided to build a binary tree according to its probability $p(+1|\mathbf{X})$ at the classifier C,

$$p(+1|\mathbf{X}) = \frac{\exp\{2C(\mathbf{X}; H_t^1, \cdots, H_t^N)\}}{1 + \exp\{2C(\mathbf{X}; H_t^1, \cdots, H_t^N)\}}. \tag{4}$$

If $p(+1|\mathbf{X}) > 0.5 - e$, $\mathbf{X}$ is divided to the left sub-tree, or if $p(+1|\mathbf{X}) < 0.5 + e$, it is put to the right sub-tree. Note those $\mathbf{X}$ falling in the range of $[0.5 - e, 0.5 + e]$ are put to both sub-trees. A cascaded Adaboost classifier [29] is trained at each leaf node to make the final classification. By this means, positive samples with large intra-class variations are handled seperately. The depth of this binary tree is set to 2 empirically and $e = 0.1$. The entire classification procedure is illustrated in Fig. 4.

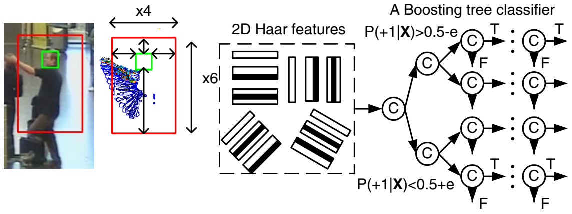


Figure 4. The feature extraction and classification procedures.

In implementation, the Haar feature responses are normalized *w.r.t* the area of the candidate region, instead of actual image normalization or down-sampling, which leads to some kind of scale and translation invariance of the Haar feature responses. The Haar feature pool includes 12 types uniformly sampled on the candidate regions. In particular, line-like Haar features are preferred since the MEHIs mainly contain edge segments, so we only keep those features whose aspect ratios are larger than $3:1$. The total number of features is 35220 for one candidate from all 4 MEHIs. Since the numbers of positive and negative samples are quite unbalanced, we generate more positive samples by randomly shifting and zooming the annotated ground truths, as well as mirroring their Haar responses. For the TRECVID dataset, we train a one-against-all boosting tree classifier for each action where the leaf node consists of up to 10 stages. At each stage, target detection and false positive rates are set to 0.95 and 0.5. For the KTH and Weizmann datasets, we train multiple one-against-all Adaboost classifiers with the detection rate 0.995 and false positive rate 0.001, which is sufficient to yield good performance.

3.4. Computational complexity

One of the prominent merits of our method is the computational efficiency which is critical for detecting actions from hours of real surveillance videos within reasonable time. In our method, at the testing stage, the feature extraction module only involves calculating the MEHIs and their integral images, and the classifiers merely evaluate up to several thousands of 2D Haar features.

Human detection and tracking are computational expensive but still affordable. For a Core2Duo 3.16GHz desktop, if the image resolution is 160×120 (as in the KTH dataset), the entire system including human detection and tracking runs at over 25 fps. In contrast, recent work [6] reports 0.75 sec per frame, and 2.4 sec per frame in [9] for the same image resolution. Note most of existing approaches only report the classification time while the processing time for tracking and stabilizing human figures is unknown. For videos at resolution 720×576 in the TRECVID dataset, the feature extraction and classification modules run at 20 fps and the system runs at 2 fps including human detection and tracking. Consequently, it is computationally feasible to process 50 hours videos in parallel in about one day.

4. Experiments

The action detection task can be evaluated in terms of the accuracy either frame-based or video-based. The video-based results are typically obtained by majority voting of the frame-based results if temporal segmentations of actions are known. We evaluate the proposed method on the TRECVID dataset in terms of the frame-based performance and compare with a bag-of-words approach and an optical flow based approach. To further show the effectiveness of this new action representation and the generalization ability of the system, we also present both frame-based and video-based performance on the KTH and the Weizmann datasets.

4.1. Human detection and tracking

In our system, we employ a human detector based on Convolutional Neural Networks (CNN) [15] and a multiple hypotheses based tracker [1, 7] to locate human heads. For the TRECVID dataset, we annotated all human heads every 750 frames to test the performance. For 4 different camera views, the average number of true heads per frame (*avg. #*), the recall rate (*rec.*) and precision rate (*pre.*) of detection plus tracking are summarized in Tab. 1 with sample frames in Fig. 5. Certainly, wrong human detections may degrade the action detection performance later on. However, in practice, applications have to cope with such imperfect detection and tracking results especially in complex and crowded scenes and do not expect accurate annotations.

	CAM1	CAM2	CAM3	CAM4	Average
avg. #	5.94	25.36	11.93	7.71	**12.72**
rec.	61.58%	31.39%	60.46%	70.90%	**49.25%**
pre.	77.31%	60.56%	78.09%	55.68%	**66.13%**

Table 1. Performance of human detection plus tracking on the TRECVID dataset.

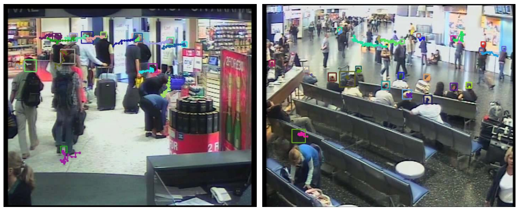


Figure 5. Sample human detection and head tracking results.

4.2. The TRECVID surveillance event dataset

The 2008 TRECVID surveillance event detection dataset [21] consists of 50-hour (5 days × 2 hours/day × 5 cameras) videos in the development set and 49-hour videos in the evaluation set. There are about 190K frames per 2-hour video with image resolution 720×576. The ground truths of occurrences of actions in the development set were provided by NIST [21]. We further labeled the locations of the persons performing actions every 3 frames for training. The actions of interests are 3 required events in the evaluation: *CellToEar*, *ObjectPut*, and *Pointing*. Sample positive training data are shown in Fig. 6, where we observe large intra-class variations due to different view angles and diverse ways people performing the same type of actions.

Since the videos were recorded on 5 different days, we perform 5-fold cross-validation accordingly, which guarantees the same person hardly appears in both training and testing sets. The positive training samples of an action are the frame-based labeled instances and the negative samples

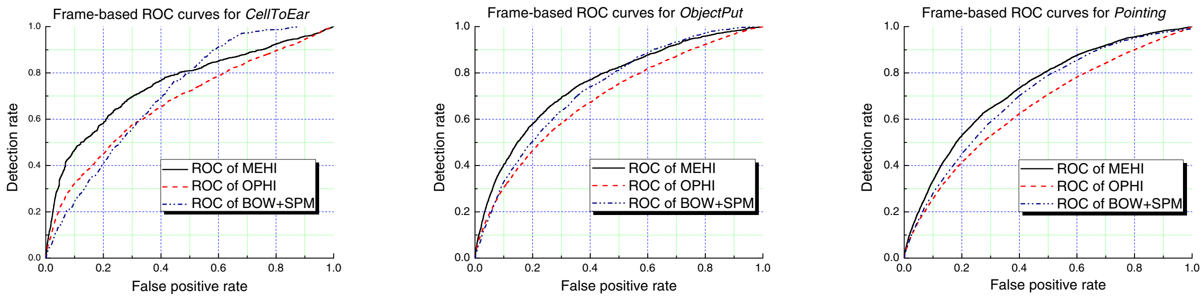


Figure 7. Comparison of the average frame-based ROC curves for detecting *CellToEar*, *ObjectPut*, and *Pointing*.

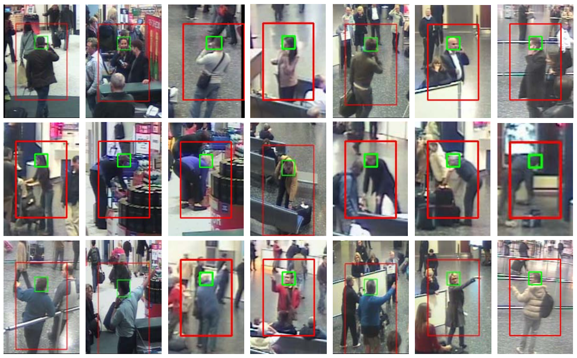


Figure 6. Examples of positive training samples for *CellToEar* (1st row), *ObjectPut*(2nd row), and *Pointing* (3rd row).

FPR	*CellToEar*	*ObjectPut*	*Pointing*
1%	12.00%	9.65%	7.54%
2%	17.34%	13.41%	11.06%
5%	35.88%	26.70%	20.36%
10%	47.28%	39.21%	38.29%

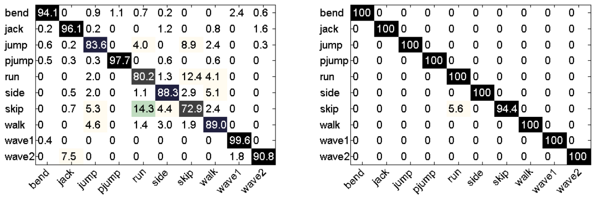
Table 2. The detection rates at different false positive rates for the 3 actions on the TRECVID dataset.

are the human detection and tracking outputs including false positives. We train one-against-all boosting-tree classifiers for each action with up to 2200 Haar features. There are 2114, 2172, and 8725 positive samples of *CellToEar*, *ObjectPut*, and *Pointing*, respectively, and about 188K negative training samples. As the actions are not segmented, we evaluate the performance in terms of the frame-based classification results. The average detection rates of 5-fold cross-validation at different positive rates (*FPR*) are present in Table 2. On average our method achieves about detection rate 9.73% *vs.* false positive rate 1% for these 3 actions. This performance is promising considering the huge variations in this dataset and the efficiency of the method.

We implement two baseline methods for comparison. One is a bag-of-words approach where dense SIFT features [17] are quantized with a 256-word codebook in a candidate region. Then, up to 4×4 spatial pyramid matching [14] of the histograms is used to incorporate the spatial layout information of these local features to train SVM classifiers. This method is denoted by *BOW+SPM*. We also substitute motion edge images in our method by Lucas-Kanade optical flow fields and accumulate them to a set of optical flow history images (*OPHIs*), while, the 2D Haar feature pool and the classification module remain the same. Our method outperforms these two methods in terms of the average ROC curves of 5-fold cross-validation shown in Fig. 7. These comparisons show that the MEHI-based representations using 2D Haar features are more effective to extract relevant information to motion patterns of actions than local features. Moreover, the MEHIs are far less sensitive to clutters than optical flow responses.

4.3. The Weizmann dataset

The Weizmann dataset was first used in [2] and consists of 9 subjects performing 10 actions: *bending down*, *jumping jack*, *jumping*, *jumping in space*, *running*, *galloping sideways*, *walking*, *waving one hand* and *waving both hands*. In total, there are 93 sequences with resolution 180×144. Sample frames including the tracked trajectories are shown in the first row of Fig. 8 with the corresponding MEHI representations in the second row.



	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	94.1	0	0.9	1.1	0.7	0.2	0	0	2.4	0.6
jack	0.2	96.1	0.2	0	0	1.2	0	0.8	0	1.6
jump	0.6	0.2	83.6	0	4.0	0	8.9	2.4	0	0.3
pjump	0.5	0.3	0.3	97.7	0	0.6	0	0.6	0	0
run	0	0	2.0	0	80.2	1.3	12.4	4.1	0	0
side	0	0.5	2.0	0	1.1	88.3	2.9	5.1	0	0
skip	0	0.7	5.3	0	14.3	4.4	72.9	2.4	0	0
walk	0	0	4.6	0	1.4	3.0	1.9	89.0	0	0
wave1	0.4	0	0	0	0	0	0	0	99.6	0
wave2	0	7.5	0	0	0	0	0	0	1.8	90.8

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	100	0	0	0	0	0	0	0	0	0
jack	0	100	0	0	0	0	0	0	0	0
jump	0	0	100	0	0	0	0	0	0	0
pjump	0	0	0	100	0	0	0	0	0	0
run	0	0	0	0	100	0	0	0	0	0
side	0	0	0	0	0	100	0	0	0	0
skip	0	0	0	0	5.6	0	94.4	0	0	0
walk	0	0	0	0	0	0	0	100	0	0
wave1	0	0	0	0	0	0	0	0	100	0
wave2	0	0	0	0	0	0	0	0	0	100

Figure 9. The average confusion matrices of action recognition results on the Weizmann dataset: per-frame results (on the left) and per-video results (on the right).

Following [4, 23, 22], we perform leave-one-out cross-validation using 8 persons for training and the other one for testing. Multiple one-against-all Adaboost classifiers are trained for each action category, where the number of Haar features selected in the classifier is up to 150 (this

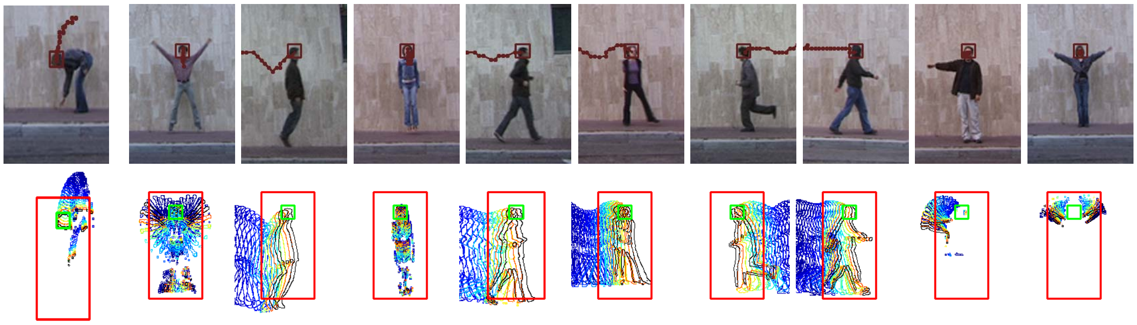


Figure 8. Sample frames of the Weizmann dataset (1st row) and the corresponding MEHIs ($\alpha = 0.8$) (2nd row).

number reveals the difficulty of the task in some sense). The video-based classification is obtained by voting of the frame-based results. The average confusion matrices are present in Fig. 9. The average accuracies compared with other approaches are listed in Table 4 illustratively.

4.4. The KTH dataset

The KTH dataset was first recorded for [12, 26] and includes 6 types of actions: *boxing*, *handclapping*, *handwaving*, *jogging*, *running*, and *walking*. There are 25 subjects performing these actions under 4 different scenes: outdoors (s1), outdoor with scale variations (s2), outdoors with different clothes (s3), and indoors (s4). In total, there are 598 sequences with image resolution 160×120. Sample frames are illustrated in the first row of Fig. 10.

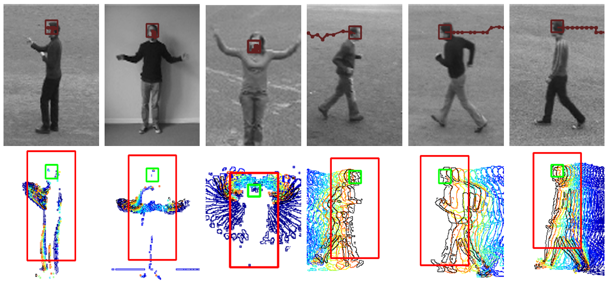


Figure 10. Sample frames of the KTH dataset (1st row) and the corresponding MEHIs ($\alpha = 0.8$) (2nd row).

We perform 5-fold cross-validation on the KTH dataset, where the sequences of 20 persons are used in training and those of the other 5 persons are for testing. The number of Haar features in the Adaboost classifiers is up to 600. The accuracies of different scenes and the all-in-one test are averaged over 5 folds and listed in Table. 3 with the average confusion matrices present in Fig. 11. We compare with recent approaches in terms of the average recognition accuracies in Table 4. This comparison is quite indicative since the ways to split data and perform cross-validation vary in different methods. Our results are quite competitive in that we employ weak assumptions, *i.e.* not assuming that the human figures are well aligned and stabilized or one video clip contains a single instance of an action. Some video clips in the KTH dataset have very low image contrast which degrades our tracking performance sometimes.

Accuracy	s1	s2	s3	s4	all-in-one
frame-based	73.9%	71.0%	73.6%	78.9%	74.4%
video-based	83.7%	84.4%	82.6%	92.4%	87.3%

Table 3. The action recognition accuracy on the KTH dataset.

Avg. accuracy	KTH		Weizmann	
methods	per-frame	per-video	per-frame	per-video
Our method	**74.4%**	**87.3%**	**89.2%**	**99.4%**
Schindler et al. [25]	88.0%	92.7%	99.6%	100%
Fathi et al. [6]	N/A	90.5%	99.9%	100%
Jhuang et al. [9]	N/A	91.7%	N/A	98.8%
Niebles et al. [23, 22]	N/A	81.5%	55.5%	72.8%
Dollár et al. [4]	N/A	81.2%	N/A	N/A
Schüldt et al. [26]	N/A	71.7%	N/A	N/A

Table 4. Illustrative comparison of the action recognition performance.

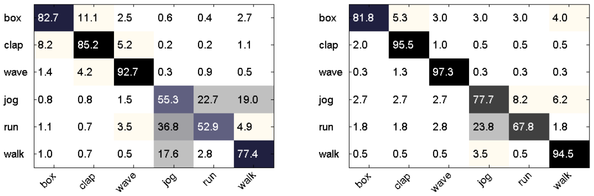


Figure 11. The average confusion matrices of action recognition results on the KTH dataset: per-frame results (on the left) and per-video results (on the right).

4.5. Discussion

Our approach shows promising results in detecting the 3 actions in the TRECVID dataset, however the false positive rates are still high. From our observations, there are two primary reasons: the semantic gap between motion patterns and actions, and the cluttered motion background. Some false detections are reasonable in the sense that the subtleties of the motion patterns are too hard to discern. For example, fixing hair may be confused with *CelltoEar*, the motion of getting an object is identical to that of putting an object, and many actions involve movements of arms similar as *Pointing*. Some typical false detections are shown in Fig. 12. The other complication is that though our method is not sensitive to cluttered background, the cluttered motion background, *e.g.* motions of a crowd on the background, threatens correct detections severely.

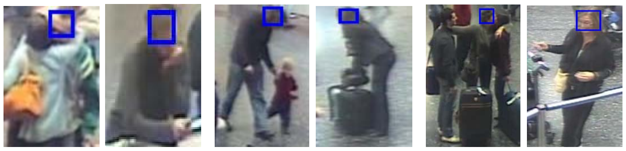


Figure 12. Sample false detections. From left to right: 2 false positives for *CelltoEar*, *ObjectPut*, and *Pointing*, respectively.

5. Conclusion

In this paper, we demonstrate the effectiveness of an efficient action representation based on a set of motion edge history images and present a complete human action detection system for real surveillance videos. We show encouraging detection performance of 3 basic actions on the challenging 2008 TRECVID event detection dataset. We believe this challenging dataset will greatly propel the research on action detection in realistic settings. In addition, the method generalizes well in recognizing 16 actions on the KTH and Weizmann datasets and achieves competitive performance in comparison with the state-of-the-art algorithms. Our future work will include improving the motion features to alleviate the restriction of slow camera motion.

References

[1] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *CVPR'98*, pages 232–237, Santa Barbara, CA, June 23-25, 1998. 5

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV'05*, volume 2, pages 1395–1402, Beijing, China, Oct. 17-21, 2005. 1, 2, 6

[3] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Machine Intell.*, 23(3):257–267, Mar. 2001. 2, 3

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS'05*, pages 65–72, Beijing, Oct. 15-16, 2005. 1, 3, 6, 7

[5] A. A. Efros, A. C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV'03*, volume 2, pages 726–733, Nice, France, Oct. 13-16, 2003. 1, 2, 3

[6] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *CVPR'08*, Anchorage, AK, June 23-28, 2008. 1, 3, 5, 7

[7] M. Han, W. Xu, H. Tao, and Y. Gong. An algorithm for multiple object trajectory trackingt. In *CVPR'04*, volume 1, pages 864–871, Washington, DC, Jun.27-Jul.2 2004. 5

[8] S. Hongeng and R. Nevatia. Large-scale event detection using semi-hidden markov models. In *ICCV'03*, volume 2, pages 1455–1462, Nice, France, Oct. 13-16, 2003. 2

[9] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV'07*, Rio de Janeiro, Brazil, Oct. 14-21, 2007. 1, 3, 5, 7

[10] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *ICCV'05*, volume 1, pages 432–439, Beijing, China, Oct. 17-21, 2005. 1, 3

[11] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *ICCV'07*, Rio de Janeiro, Brazil, Oct. 14-21, 2007. 1, 2

[12] I. Laptev and T. Linderg. Space-time interest points. In *ICCV'03*, volume 1, pages 432–439, Nice, France, Oct. 13-16, 2003. 3, 7

[13] I. Laptev and P. Pérez. Retrieving actions in movies. In *ICCV'07*, Rio de Janeiro, Brazil, Oct. 14-21, 2005. 1, 3

[14] S. Lazebnik, C. Achmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR'06*, volume 2, pages 2169–2178, New York City, June17 - 22 2006. 6

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based leraning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, Nov. 1998. 5

[16] J. Liu, S. Ali, and M. Shah. Recognizing human action using multiple features. In *CVPR'08*, AK, June 23-28, 2008. 3

[17] D. G. Lowe. Distinctive image features from scale invariant keypoints. *Int'l Journal of Computer Vision*, 60(2):91–110, 2004. 2, 6

[18] F. Lv and R. Nevatia. Single view human action recognition using key pose matching and viterbi path searching. In *CVPR'07*, Minneapolis, MN, June 17-22, 2007. 2

[19] K. Mikolajczyk and H. Uemura. Action recognition with motion-appearance vocabulary forest. In *CVPR'08*, AK, June 23-28, 2008. 1, 3

[20] P. Natarajan and R. Nevatia. View and scale invariant action recognition using multiview shape-flow model. In *CVPR'08*, Anchorage, AK, June 23-28, 2008. 2

[21] National Institute of Standards and Technology (NIST): TRECVID 2008 Evaluation for Surveillance Event Detection. http://www.nist.gov/speech/tests/trecvid/2008/ and http://www.nist.gov/speech/tests/trecvid/2008/doc/eventdet08-evalplan-v07.htm#tasks, 2008. 1, 5

[22] J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *CVPR'07*, Minneapolis, MN, June 17-22, 2007. 1, 3, 6, 7

[23] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC'06*, Edinburgh, 4-7, 2006. 3, 6, 7

[24] M. D. Rodriguez, J. Ahmed, and M. Shah. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR'08*, Anchorage, AK, June 23-28, 2008. 1

[25] K. Schindler and L. V. Gool. Action Snippets: How many frames does human action recognition require? In *CVPR'08*, Anchorage, AK, June 23-28, 2008. 1, 2, 7

[26] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR'04*, volume 3, pages 32–36, Cambridge, UK, Aug. 23-26, 2004. 1, 2, 3, 7

[27] E. Shechtman and M. Irani. Space-time behavior based correlation. In *CVPR'05*, volume 1, pages 405–412, San Diego, CA, June 20-25, 2005. 1, 2

[28] Z. Tu. Probabilistic boosting-tree: learning discriminative models for classification, recognition, and clustering. 2, 3, 4

[29] P. Viola and M. Jones. Robust real-time object detection. *Int'l Journal of Computer Vision*, 57(2):137–154, 2001. 2, 4

[30] D. Weinland and E. Boyer. Action recognition using examplar-based embedding. In *CVPR'08*, Anchorage, AK, June 23-28, 2008. 1

[31] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3D examplars. In *ICCV'07*, Rio de Janeiro, Brazil, Oct. 14-21, 2007. 2

[32] A. Yilmaz and M. Shah. Actions sketch: A novel action representation. In *CVPR'05*, volume 1, pages 984–989, San Diego, CA, June 20-25, 2005. 1, 2