

Action Detection in Complex Scenes with Spatial and Temporal Ambiguities

Yuxiao Hu^{§,†,*}, Liangliang Cao[§], Fengjun Lv[†], Shuicheng Yan[◇], Yihong Gong[†], and Thomas S. Huang[§]

[†]NEC Laboratories America, Inc., Cupertino, CA

[‡] Microsoft Cooperation, One Microsoft Way, Redmond, WA

[◇]Department of ECE, National University of Singapore, Singapore

[§]Beckman Institute and Coordinated Science Lab, Department of ECE, UIUC

Abstract

In this paper, we investigate the detection of semantic human actions in complex scenes. Unlike conventional action recognition in well-controlled environments, action detection in complex scenes suffers from cluttered backgrounds, heavy crowds, occluded bodies, and spatial-temporal boundary ambiguities caused by imperfect human detection and tracking. Conventional algorithms are likely to fail with such spatial-temporal ambiguities. In this work, the candidate regions of an action are treated as a bag of instances. Then a novel multiple-instance learning framework, named SMILE-SVM (Simulated annealing Multiple Instance LEarning Support Vector Machines), is presented for learning human action detector based on imprecise action locations. SMILE-SVM is extensively evaluated with satisfactory performances on two tasks: 1) human action detection on a public video action database with cluttered backgrounds, and 2) a real world problem of detecting whether the customers in a shopping mall show an intention to purchase the merchandise on shelf (even if they didn't buy it eventually). In addition, the complementary nature of motion and appearance features in action detection are also validated, demonstrating a boosted performance in our experiments.

1. Introduction

In the past few years, computer vision researchers have witnessed a surge of interest in human action analysis through videos. Researchers have built several public action data sets (e.g., KTH [17], Weizmann [3]), which provide good test beds for algorithm evaluation. Although these data sets have become very popular, there exists a considerable gap between these staged samples and real world scenarios. The majority of the action data sets [17] [3] are col-



Figure 1. Illustration of the action detection problem in complex scenes. The image in the first row shows a crowded scene of a shopping mall. The left five images in the second row are the detected actions of reaching (first two), pointing, squatting, and bending to merchandise on shelf. The last image in the second row is a negative sample as the customer is only walking in front of the shelf.

lected in well-controlled environments, while the real world actions often happen in much more complex scenes.

In most current human action data sets [17] [3], the human actions are generally recorded with clean backgrounds, and each video clip generally involves only one type of action (e.g., running or jogging) and only one person, who keeps doing this action within the whole video clip. However, in real surveillance scenarios, the background is often cluttered, and the surveillance system has to detect the human actions of interest from a crowd. Fig. 1 shows such an example of action detection in complex scene. In contrast to classic actions such as running and jumping, we expect to know whether the customers in a shopping mall intend to get the merchandise from the shelf. The action detection in complex scenes is much more difficult than in simple labo-

*The work of Y. Hu was done in part during his internship with NEC Laboratories America, Inc.

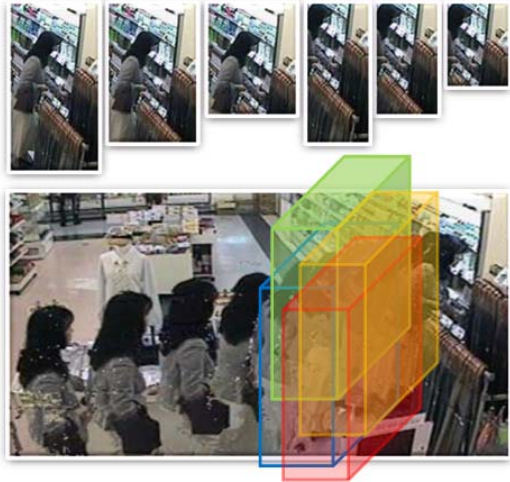


Figure 2. Illustration of multi-instance learning for human action detection in complex scenes. Top: spatial ambiguities in scale and position. Bottom: temporal ambiguities in time domain. Different color boxes indicate different candidate instances. For better viewing, please see the original pdf file.

ratory environments.

In complex scenes, *e.g.*, with cluttered backgrounds or partially occluded crowds, it is very difficult to locate human body precisely. When trying to crop an object from a complex scene, we often have to endure substantial misalignment or occasional drifting if no human interaction is involved. In addition, ambiguities may also exist in temporal domain. A large portion of real world actions happen only once and the duration is short. Since the human motion is continuous and the speed vary greatly even within the same action category, it is not easy to decide the start or end point of these actions of interest, even the duration of each action in real world scenarios. The ambiguities in temporal domain are not recognized in repetitive actions, such as running and jogging, but they may greatly affect the detection performance when handling non-repetitive actions such as picking up an item, taking a photo, and pushing an elevator button. Such spatial and temporal ambiguities bring serious difficulty into the action detection task.

To overcome these ambiguities, one naive approach is to ask human laborers for accurate labels. The labelers need to provide the bounding boxes of the objects and the starting/ending frames of an action instance. This labeling work is extremely tedious. For a video as long as several hours, the labeling process might take several weeks or even longer. In the detection stage, we may also encounter troubles in aligning the actions. Because the boundaries between continuous actions are usually fuzzy and the background is often cluttered, it is difficult to obtain well-aligned action instances to feed into the classifier. If the spatial location, scale and shape of the human action are inconsistent

with those of the training data, the performance of human action detection may be degraded much.

In this work, we employ multi-instance learning (MIL) based Support Vector Machine (SVM) to handle these ambiguities in both spatial and temporal domains. Fig. 2 illustrates the main idea of multi-instance learning. Although we do not know exactly where and when the target action happens, we may estimate a "bag" covering more than one potential region and time slice. A bag can be positive (target action happen somewhere in the bag) or negative (absolutely no interesting action happens). There must be at least one positive instance in one positive bag, while all instances in one negative bag are non-action instances. This multi-instance method provides a way to not only recognize the action of interest, but also locate the exact position and time period of the action.

To avoid the local minimum trap caused by the unbalanced data during the iteration of MIL, simulated annealing (SA) is introduced to search for the global optimum in the learning process. We called the proposed algorithm as Simulated annealing Multiple Instance Learning (SMILE).

2. Background and Motivations

The most popular approach for human action recognition is to employ spatial-temporal interest points [16][9], modeled in generative [21] and discriminative [23] [11] manners. Much effort have been put in enriching spatial-temporal interest points with additional information, *e.g.*, hierarchical structures [13], implicit shapes [25], local contexts [26], 3D spin images [19], and 3D cube [28]. Employing spatial-temporal interest points makes it easier to distinguish the periodic actions such as running and jogging, where we need not to consider the alignment problem in temporal domain. However, spatial-temporal interest points focus the local information instead of global motion, and the detection of *real* spatial-temporal interest points on human bodies in complex scenes might fall on cluttered backgrounds if the camera is not fixed.

Most previous algorithms for human action detection not based on spatial-temporal interest points are constrained to well-controlled environments. Boiman and Irani [5] proposed to extract densely sampled local video patches for detecting irregular actions in videos with simple background. Rodriguez *et al.*[22] designed a novel filter to analyze the filtering responses of different actions. This approach has difficulties in aligning non-repetitive actions in complex scenes. Moreover, some researchers [29] [12] tried to model the configuration of human body and its evolvement in time domain. In [4], Bobick *et al.* reveals the difficulties of recognizing actions as the variability of how the movements are made and how the causal relationship appears between human and environment, which confirms our difficulties of spatial-temporal ambiguity.

Beyond the limitations, we have observed that although some works aimed to solve problems different with ours, their approaches are still valuable to help action detection in complex scenes. Efros *et al.* studied the actions in sports games [10] with relatively simple backgrounds. Ke *et al.* proposed the volumetric features to correlate spatial-temporal shapes to segmented video clips [14]. Combined with flow-based correlation techniques [10], Ke’s algorithm can detect a wide range of actions in videos, based on action exemplars obtained by manual segmentation. In [8], Davis and Bobick also tackled the action recognition problem in the scenes with simple backgrounds. For the action of interest in complex scenes, their approach suffers from the ambiguities in spatial and temporal domains.

3. System Overview

To collect the data for building an action classifier, we manually labeled the video sequences to obtain the training samples. Only rough positions of the human heads and the approximate frame where the action happens need to be specified. This labeling process can be further simplified where automatic human head detection/tracking is available. It is unnecessary to provide the exact frame by frame labeling, since our algorithm in Section-4 will automatically exclude the non-interesting portions of the labeled data. As shown in our experiments, by allowing the ambiguities in the labeling process, great reduction of labeling labor can be achieved without sacrificing algorithmic performance.

After the labeling process, the labeled video sequences are further cropped at variant locations/scales within the frame and different start/end frame number in timeline so that each action (referred as a bag hereafter) will generate multiple segments (referred as instances hereafter). These positive and negative bags are fed into our proposed learning algorithm for action detector training. Here each positive bag includes a target as the action of interest, while a negative bag does not.

In the testing phase, our proposed algorithm will handle the ambiguities of locations of human actions in both spatial and temporal domains. Our proposed algorithm allows multiple candidates in a short sequence as inputs, and infers whether the action of interest happens. It does not require accurate tracker or human detections. On the contrary, the outputs of face detectors or probabilistic trackers can be used as estimations of the human bodies. Also it does not assume to know the exact start or end frames of the human actions. Instead, it can take multiple possibilities into account and estimate where the action truly happens.

To obtain discriminative features for action detection, we first consider the motion features to distinguish the actions of interest from the others. Since the traditional optical flow is prone to noise, we employ the Davis and Bobick’s motion history image (MHI) feature [8], which accumulates the mo-

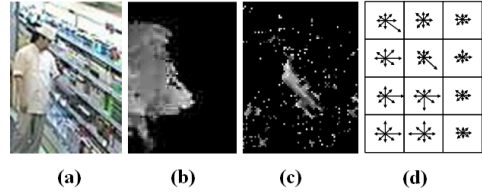


Figure 3. Illustration of appearance features and motion features: (a) original window, (b) motion history image feature (MHI), (c) foreground image feature (FI), and (d) histogram of oriented gradients.

tion information of several frames. In our system, we compute the MHI feature for each instance window and down-sample it to the size of 10×10 pixels, that is, a feature vector of length 100.

Although MHI is an intuitive feature for motion description, it is not discriminative enough to characterize complex actions. We propose to combine both motion and appearance information for better characterization of human actions. We employ two kinds of appearance-based features, which are combined with MHI as a more discriminative feature for action recognition. The first appearance based feature is foreground image (FI), obtained by background subtraction [15], and the second one is the histogram of oriented gradients feature (HOG) [7], which characterizes the directions and magnitudes of edges and corners. Given a image region of an instance, the FI feature is normalized to 10×10 pixels. To obtain the HOG feature, the image region is divided into $3 \times 4 = 12$ sub-windows and then eight bins of gradient directions are used for computing the histogram, which generates an HOG feature vector of 96 dimension for each instance. We extract HOG for each patch instead of the dense sampling in [7] so that this patch-based histogram feature is robust to alignment errors.

The motion features (MHI) and appearance features (FI and HOG) characterize the human actions from different aspects and are complementary to each other. Appearance features capture the spatial shape of human bodies during actions, while the motion features focus on capturing the direction and intensity of the moving body parts. Fig. 3 illustrates examples of these features. The experiments introduced later show that the combination of these two types of features yields to a better performance compared with single features.

4. SMILE-SVM Model

In recent years, multiple instance learning has attracted much research interest in the field of machine learning and computer vision. People have applied multiple instance learning for different tasks, *e.g.*, scene classification [20], drug activity prediction [30], image annotation [27], and face detection [24]. In this paper, we will show that multi-

ple instance learning can be used to handle the ambiguities in action detection in both spatial and temporal domain.

Following popular work by Andrews, Tsochantaridis and Hofmann [2], we employ the Supporting Vector Machine (SVM) as the basic classifier for multiple instance learning. However, the work in [2] does not guarantee the convergence and may suffer from the errors made in earlier iterations. In this paper, we propose an algorithm named SMILE-SVM (Simulated annealing Multiple Instance LEarning Support Vector Machines), which aims to obtain a global optimum via simulated annealing method, thus not relies on model initialization to avoid falling into local minima.

Given a set of input patterns x_1, x_2, \dots, x_N grouped into bags B_1, \dots, B_M , with $B_m = \{x_i : i \in I_m\}$ for given index sets $I_m \subseteq \{1, \dots, N\}$. Each bag B_m is associated with a label Y_m , where $Y_m = 1$ means the bag is positive and at least one instance $x_i \in B_m$ is a positive example of the class. In contrast, $Y_m = -1$ means that the bag is negative, and all the instances $x_i \in B_m$ are negative examples. If we denote the label for each instance as y_i , we have $\forall y_i = -1$, for $i \in I_m$, if $Y_m = -1$, for $m = 1, 2, \dots, M$. Otherwise if $Y_m = 1$, $\exists y_i = 1$, for $i \in I_m$. Note that this constraint can also be written as $\sum_{i \in I_m} \frac{y_i + 1}{2} \geq 1$.

SVM based multiple-instance learning can be naturally formulated as to minimize the object functions as below,

$$\min_{y_i} \min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + c \sum_i \xi_i, \quad (1)$$

subject to

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad (2)$$

$$\text{If } Y_m = 1, \sum_{i \in I_m} \frac{y_i + 1}{2} \geq 1,$$

$$\text{If } Y_m = -1, \forall i \in I_m, y_i = -1,$$

where ξ_i is the estimation error and \mathbf{w} determines the size of the margin. Unlike traditional SVM which is a quadratic optimization problem, minimizing (1) also involves a mixture of combination optimization under the constraint of (2). This objective function is difficult to minimize directly because of the large number of possible choices of y_i .

In [2], Andrew *et al.* employed an algorithm named MI-SVM to solve (1). Their algorithm starts by an initialized SVM model, which is followed by a relabeling of the instances in positive bags using the learned model. If a positive bag contains no instances labeled as positive, then the instance that with the largest SVM score in that bag is relabeled as positive. The SVM is then retrained with the new labels. The process of relabeling and retraining is repeated until no labels are changed. As agreed by the author [2], this approach is a heuristic approach and might often get trapped in local minimum.

Our SMILE-SVM model aims to increase the recognition accuracy of bags and maximize the margin of the classifier simultaneously. Since the size of the classification margin can be measured by $1/\|\mathbf{w}\|^2$ [6], we define a new objective function as

$$S = \max_{\mathbf{w}, b, y_i} nc + \frac{k}{\|\mathbf{w}\|^2}, \quad (3)$$

where nc is the correct rate of bag classification, and k is the parameter controlling the weight of margin measure. In our implementation, k is set to be 0.5.

As in the general simulated annealing algorithm [1], SMILE-SVM employs a parameter T (called temperature), which controls the probability whether a new score S is acceptable or not. At the early steps of the learning process, T is set to be high enough to allow a candidate solution to change to another state with a lower score. During the learning process, T is gradually decreased such that the probability to switching to another states of lower S is reduced. The system will converge as T approaches zero.

SMILE-SVM searches for the optimal score S_{opt} in an iterative way. In the t -th iteration, SMILE-SVM generates a neighboring state $\{y_i^*\}$ which will be fed to the next iteration for training a new SVM classifier. Here we prefer that the decision boundary of the new classifier is similar to the old one, so we only introduce a random small perturbation to generate a new state. More specifically,

$$y_i^* = \begin{cases} -\text{sign}(f_i^t), & \text{if } |f_i^t| < \text{thresh and } i \in I_{rand}^t, \\ \text{sign}(f_i^t), & \text{otherwise,} \end{cases}$$

where I_{rand}^t is the random set at the t -th iteration, and $|f_i^t|$ is the classification confidence estimated by (\mathbf{w}^t, b^t) .

After the neighboring state $\{y_i^*\}$ is generated, SMILE-SVM then decides whether to accept it as the training set for next iteration or refuse it. First the constraints in (2) are verified. If (2) is satisfied, the system will compute the score S_t by (3) based on the classifier trained according to $\{y_i^*\}$. Here the probability to accept $\{y_i^*\}$ as the state $\{y_i^{t+1}\}$ for next iteration is decided by the comparison of S_t and a random number. If $\{y_i^*\}$ is not accepted or (2) is not satisfied, another neighboring state will be generated. Above steps are summarized in Algorithm 1.

5. Experiments

In this section, we systematically evaluate the effectiveness of the proposed SMILE-SVM algorithm with two sets of experiments. The first is conducted on the CMU human action dataset ¹ collected by Ke *et al.*[14]. For the second one, we consider a real world application, namely to detect whether the customers show an intention to purchase the merchandise on shelf in a shopping mall.

¹We are thankful to Yan Ke for providing the dataset.

Algorithm 1 : SMILE-SVM Procedure.

- 1: **Initialize:** For each bag indexed as m , let $y_i^0 = Y_m$, with $i \in I_m$. Initialize $S_{opt} = \text{inf}$.
 - 2: **for each** annealing temperature T , **do**
 - 3: **for** $t = 1, 2, \dots, n$
 - 4: Compute the linear SVM solution (\mathbf{w}, b) for dataset $\{\mathbf{x}_i, y_i^{t-1}\}$, with outputs $f_i = \mathbf{w}^T \mathbf{x}_i + b$ for each example \mathbf{x}_i .
 - 5: Estimate the label of each bag F_m based on the classifier, and compute the number of correctly classified bag nc .
 - 6: Compute the classification score S_t using (3).
 - 7: **if** $S_t < S_{opt}$
 - 8: Let $(\mathbf{w}_{opt}, b_{opt}) = (\mathbf{w}, b)$,
 - 9: $S_{opt} = S_t$.
 - 10: **endif**
 - 11: **if** $\text{rand} < \exp(-\frac{S_{opt}-S_t}{T})$
 - 12: Assign $f_i^* = f_i$ for each sample i .
 - 13: Find a neighboring state $\{y_i^t\}$ based on $\{f_i^*\}$.
 - 14: **else**
 - 15: Keep $y_i^t = y_i^{t-1}$;
 - 16: **endif**
 - 17: **end for**
 - 18: Decrease $T = \rho T$, where $\rho = 0.8$.
 - 19: **end for**
- Output:** Output the SVM classifier with $(\mathbf{w}_{opt}, b_{opt})$.
-

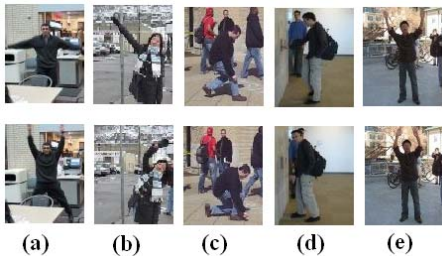


Figure 4. Some examples from the CMU action dataset. From (a) to (e) are examples of five action categories: jumping, one-hand waving, pick-up, pushing elevator buttons, and two-hand waving.

5.1. Results on CMU Action Dataset

Unlike videos in KTH [17] and Weizmann [3] datasets with clean background, the CMU action videos [14] were captured with a hand-held camera in crowded environments with moving people or cars in the backgrounds. There are five different types of human actions in the database, including jumping jacks, pick-up, two-handed wave, one-handed wave and pushing elevator buttons (as shown in Fig. 4). The duration of all the videos is about twenty minutes, which contains about one hundred actions of interest. The videos were down-scaled to 160×120 pixels in resolution. There are large variations in the ways the subjects performed the

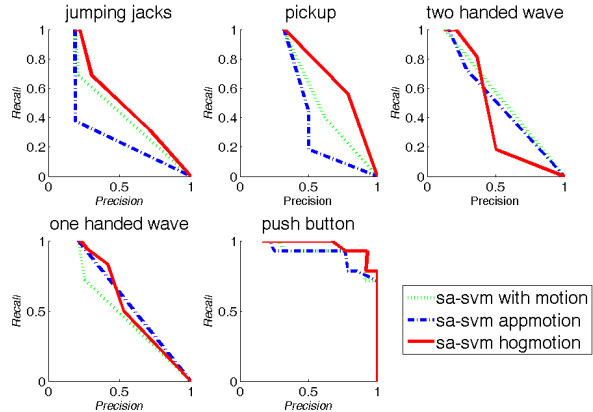


Figure 5. Comparison precision/Recall curves for a variety of actions and features. Mhi+HOG feature on SMILE-SVM outperforms other features in most cases.

actions. The background is cluttered and target actions sometimes are occluded by other people. All these variations introduce large ambiguities in both spatial and temporal domains.

The same training/testing setup as in [14] is used for evaluation. One example sequence performed by one subject is used for training where for all the five actions. About three to six other subjects performed each of these actions several times for testing. We use one vs. all strategy to train and test the five human action detectors respectively, so that we have one positive bag and four negative bags in the training phase and similar distribution of positive/negative bags in the testing phase. In Ke *et al.*'s original work [14], a user needs to interactively segment a spatial-temporal template for each action, which is then manually broken into parts. In contrast, our SMILE-SVM algorithm does not require such detailed ground truth labeling.

We conducted experiments to compare the performances of motion and appearance features and found that the combination of motion and appearance features outperforms other types of features. The performance of different features with the SMILE algorithm are compared in Fig. 5. Note the precision and recall ratio are calculated at bag (action) level in our results. As described above, if any of the instances in a bag is classified as positive instance, the whole bag will be classified as positive bag, i.e., detected as target actions. If all of the instances in a bag are classified as negative instances, this bag will be classified as negative bag, i.e., detected as non-action. These classification results are then compared with the ground truths to know whether they are overlapped with each other in spatial/temporal domain.

The best results from [14] and our method are compared in Table 1. SMILE-SVM significantly outperforms the best results of [14] for all actions except one (two-handed

Table 1. Detailed performance comparison between our proposed algorithm (mhi+HOG (SMILE-SVM)) and [14] (Shape+Flow (Parts)). Note that *R/P* means Recall vs. Precision. In each row, the first value is the recall rate, and other values are the precisions for different cases.

	Jumping jacks		pickup		two-handed wave		one-handed wave		pushing button	
R/P	[14]	ours	[14]	ours	[14]	ours	[14]	ours	[14]	ours
0.2	0.75	1.00	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.4	0.05	0.63	0.60	0.90	1.00	0.60	0.15	0.85	0.55	1.00
0.6	0.05	0.44	0.30	0.70	0.50	0.15	0.10	0.40	0.25	1.00
0.8	0.05	0.20	0.20	0.50	0.10	0.10	0.05	0.20	0.20	0.90

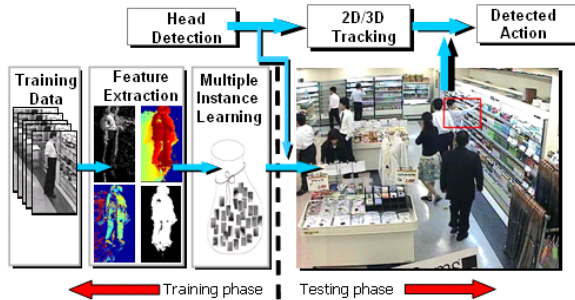


Figure 6. Overview of the system for action detection. Note that by combining our action detection result with a 3D tracker, the shop manager may know customers’ interests over different shelves.

wave). It validates the success of SMILE-SVM based on motion+appearance feature.

5.2. Results on Real World Scenario: Surveillance System in Shopping Malls

To further validate the effectiveness of our proposed algorithm, we test its performance in real world scenarios and demonstrate its usefulness in surveillance applications. The action of interest is whether a customer in crowded shopping mall shows intentions to purchase the merchandise on shelf. Such kind of intention actions include reaching/pointing at some merchandise using one or two hands or bending down to fetch/look at some merchandise. Merchants would like to track this kind of actions to know customers’ interests; more importantly, they want to further investigate which products attract customers but are not sold. Such actions of interest are quite different from typical periodic actions like walking/waving or simple actions involving only one consistent body motions, like pushing an elevator button. The database was collected in a typical shopping mall, which is usually crowded during the morning and noon hours. The customers move freely in front of the product shelf so that different shopping actions like walking/running, standing, squatting, pointing, reaching and bending can happen at different locations.

The video sequences captured at different times are used for training and testing, so that the same customer generally

do not appear in both the training and the testing data. The resolution of the camera is 320×240 pixels. The rough locations of the human heads are manually labeled. For each labeled human head, an action flag is marked when he/she focuses his/her attention on some merchandise on the shelf, *e.g.*, reaching or bending. Here the head location labels are not precise due to motion blur and occlusions. Also, the action flags are only rough estimates since the start and end points of an action are not clear during continuous movements.

We build a system to detect the action of interest. In our system, a Convolutional Neural Network (CNN) [18] is trained for detecting 2D human head candidates in each frame. Based on the outputs of head detector, the proposed action detection algorithm takes the position and size of each human head rectangle in a frame from the CNN human detector as input. Multiple windows of different locations are extracted around these head positions in spatial neighborhoods and the adjacent frames in temporal neighborhoods. The video features are then extracted for these instances. Based on the recognition results from the SMILE-SVM algorithm, target actions are detected if the estimation probability exceeds the learnt confidence threshold. When two actions are detected on the adjacent area in the temporal line, they are merged to form a longer action in the higher level. The system diagram is illustrated in Fig. 6. Note that in this system, we can integrate the action detection with 3D tracker (*e.g.*, obtained by combining the 2D tracking results from stereo cameras) to obtain specific information, *e.g.*, which part of the shelf attracts the customers’ attentions most. This kind of information is very useful to merchants.

About 20 minutes video are used for training and 40 minutes video are used for testing, which include about 150 positive action samples. On temporal aspect, each action is divided into many small segments of 10 frames in size at arbitrary points within the action. On spatial aspect, the action regions containing the human head and body are cropped

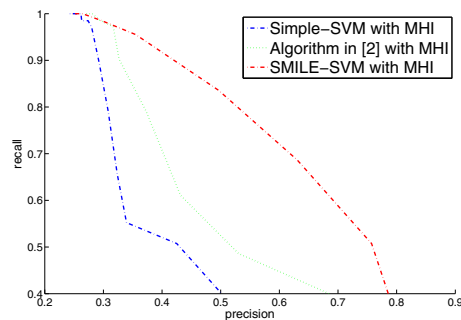


Figure 7. Performance comparison of simple SVM, the multi-instance learning algorithm in [2] and our SMILE-SVM.

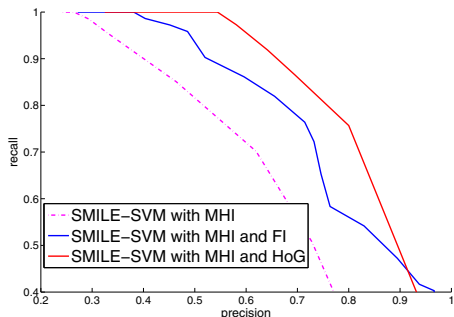


Figure 8. Performance comparison of the three kinds of features: 1) motion feature only (MHI), 2) combined motion feature (MHI) and appearance feature (FI), and 3) combined motion feature (MHI) and another appearance feature feature (HOG).

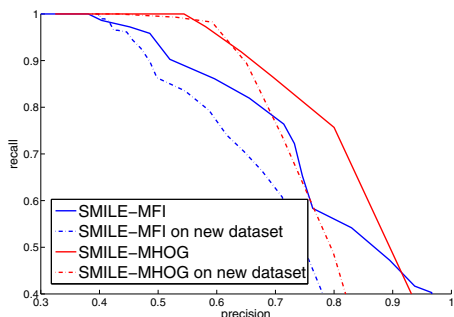


Figure 9. Generalization capability evaluation on a more challenging test video. Here MFI stands for the combined features of motion (MHI) and appearance (FI), and MHOG for the combined features of motion (MHI) and appearance (HOG).

with 12 different sizes and scales. Each above spatial-temporal sample becomes an instance and all the instances related to one action form a bag. This process will generate about 50 positive bags (containing 25k instances) and about 100 positive bags (containing 50k positive instances) in training and testing dataset respectively. For other negative actions like walking and standing (negative action samples), which are not of our interest, only 10% of them (382 negative bags) are randomly sampled to get similar number (34k and 79k) of negative instances in training and testing.

We first show the benefit of the proposed SMILE-SVM learning algorithm compared with the classical single instance SVM algorithm, and previous multiple instance learning algorithm in [2]. To make a fair comparison, we use the same motion feature (MHI) for different algorithms. When we apply the classical SVM for our problem, all the instances in a positive bag are treated as positive samples and all the instances in a negative bags are treated as negative samples. Fig. 7 compares the Precision-Recall curves of three algorithms. We can observe that the multiple instance learning algorithms obtain much better performance

than the classical SVM. In addition, the algorithm in [2] is not as good as our method, since our simulated annealing search strategy is less likely to get trapped in local optimum.

Our system is then further improved by combining motion feature with appearance features. As discussed in Section 3, both FI feature and HOG feature provide information complementary to motion feature, and hence we construct new features by combining motion with FI and HOG features, respectively. Fig. 8 shows the performance using three kinds of features: motion feature only (MHI) and two kinds combined motion feature (MHI+FI and MHI+HOG). We can observe that the combined features outperform the original motion feature. By combining MHI and HOG, we obtain the average of nearly 20% increase in recall rate (when the precision is 0.6) over using MHI only.

To test the generalization ability of our action detector, we apply it to new video sequences, which are taken in the shopping mall in a different date and time. Unlike the ori-

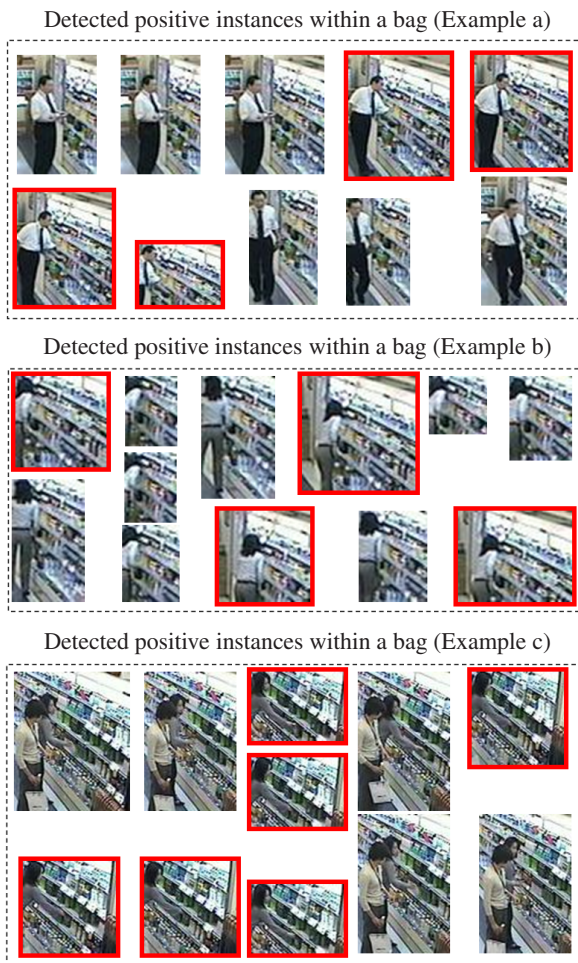


Figure 10. Illustrations of the detected actions. Each dashed box depicts an action bag, in which the positive instances are bounded by red rectangles.

gin testing set used in Fig. 7 and Fig. 8, the new video is captured during busy hours, with more customers and more actions. This new testing video contains 390 target actions as positive bags and 591 non-target as negative bags, which contains about 124k positive and 166k negative instances.

Fig. 9 compares the precision recall curves on the original test data and the more challenging test data, using the combined MHI+FI and MHI +HOG features, respectively. We can observe that the overall performance degenerates slightly due to the effect of date difference. However, this error is not big and the resulting performance on new data set is still acceptable.

As stated in previous sections, all the instances in the non-action bags are negative samples, while only some instances in the action bags are recognized as positive. From these positive instances, we can not only recognize whether an interesting action happens in a bag, but also estimate the location and time period of the action. Fig. 10 demonstrates some example action detection results, where the positive instances are bounded in red boxes. Because of the way that our instances are structured, our algorithm can detect the interesting action along with its time period and location (Fig. 10.a and Fig. 10.b), even when the person is partially occluded (Fig. 10.c).²

6. Conclusions and Future Work

This paper studied the problem of human action detection in complex scenes, for which a framework of multi-instance learning was introduced to overcome the spatial and temporal ambiguities. We proposed the SMILE-SVM algorithm to avoid the local optimum issue of the traditional multi-instance learning. In addition, the mutual complementarity of the motion and appearance features were well validated for human action detection purpose. The proposed algorithm not only outperforms the state of art on public available CMU action database but also proves practical to real world surveillance applications. We build a system for detecting whether customers intend to reach the merchandise on shelf in crowded shopping malls, which provides valuable information for merchants to understand customers' interests. Our future work is to further extend the current system for more general applications, *e.g.*, in a cafeteria or at McDonald's.

Acknowledgement

We are thankful to the reviewers for their valuable comments. Cao and Huang are partially supported by U.S. Government VACE Program. Yan is partially supported by NRF/IDM grant NRF2008IDM-IDM004-029.

²Due to the limits of space, we cannot put more figures here. More results are available in <http://www.ifp.uiuc.edu/~cao4/miaction/>

References

- [1] E. Aarts. *Simulated Annealing: Theory and Applications*. Springer, 1987.
- [2] S. Andrews, I. Tsochantaris, and T. Hofmann. Support vector machines for multiple-instance learning. *NIPS*, 2002.
- [3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *ICCV*, 2005.
- [4] A. Bobick *et al.* Movement, Activity, and Action: The Role of Knowledge in the Perception of Motion. *Phil. Trans. Royal Society London B*, 352, pp.1257-1265, 1997
- [5] O. Boiman and M. Irani. Detecting irregularities in images and in video. *ICCV*, 2005.
- [6] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. *ACM Workshop on Computational Learning Theory*, 1992.
- [7] N. Dalai and B. Triggs. Histograms of oriented gradients for human detection. *ICPR*, 2006.
- [8] J. Davis and A. Bobick. The representation and recognition of human movement using temporal templates. *CVPR*, 1997.
- [9] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. *IEEE Workshop on VS-PETS*, 2005.
- [10] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. *ICCV*, 2003.
- [11] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. *CVPR*, 2008.
- [12] N. Ikinler and D. Forsyth. Searching video for complex activities with finite state models. *CVPR*, 2007.
- [13] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. *ICCV*, 2007.
- [14] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. *ICCV*, 2007.
- [15] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis. Real-time foreground-background segmentation using code-book model. *RealTime Image*, 11(3):172–185, 2005.
- [16] I. Laptev and T. Lindeberg. Space-time interest points. *ICCV*, 2003.
- [17] I. Laptev and T. Lindeberg. Velocity adaptation of space-time interest points. *ICPR*, 2004.
- [18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.
- [19] J. Liu, S. Ali, and M. Shah. Recognizing human actions using multiple features. *CVPR*, 2008.
- [20] O. Maron and A. Ratan. Multiple-instance learning for natural scene classification. *ICML*, 1998.
- [21] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *BMVC*, 2006.
- [22] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: A spatio-temporal maximum average correlation height filter for action recognition. *CVPR*, 2008.
- [23] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. *ICPR*, 2004.
- [24] P. Viola, J. Platt, and C. Zhang. Multiple instance boosting for object detection. *NIPS*, 2005.
- [25] S. Wong, T. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. *CVPR*, 2007.
- [26] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg. A Scalable Approach to Activity Recognition based on Object Use. In *ICCV*, 2007.
- [27] C. Yang, M. Dong, and J. Hua. Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning. *CVPR*, 2006.
- [28] J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. *CVPR*, 2009.
- [29] L. Zelnik-Manor and M. Irani. Statistical analysis of dynamic actions. *TPAMI*, 28(9):1530–1535, 2006.
- [30] Q. Zhang and S. Goldman. EM-DD: An improved multiple-instance learning technique. *NIPS*, 2001.