# CoDeL: A Human Co-detection and Labeling Framework

Jianping Shi* Renjie Liao* Jiaya Jia
The Chinese University of Hong Kong
{jpshi, rjliao, leojia}@cse.cuhk.edu.hk

## Abstract

*We propose a co-detection and labeling (CoDeL) framework to identify persons that contain self-consistent appearance in multiple images. Our CoDeL model builds upon the deformable part-based model to detect human hypotheses and exploits cross-image correspondence via a matching classifier. Relying on a Gaussian process, this matching classifier models the similarity of two hypotheses and efficiently captures the relative importance contributed by various visual features, reducing the adverse effect of scattered occlusion. Further, the detector and matching classifier together make our model fit into a semi-supervised co-training framework, which can get enhanced results with a small amount of labeled training data. Our CoDeL model achieves decent performance on existing and new benchmark datasets.*

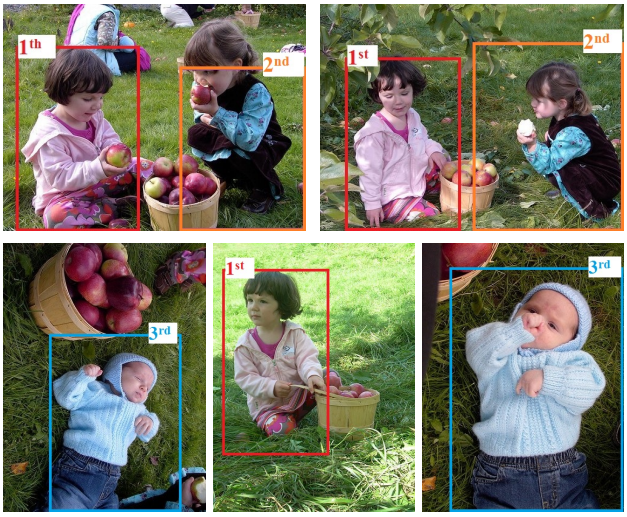


Figure 1. Human co-detection. The cross-image correspondence in color and texture features can improve detection, and provides a natural extension for individual grouping.

## 1. Introduction

We in this paper tackle the *human co-detection* problem, which can be defined in the following way. Given $N$ images $\mathcal{I} = \{I_1, \ldots, I_N\}$, which contain a group of $M$ persons denoted as $\mathcal{H} = \{H_1, \ldots, H_M\}$, the objective includes detecting human in the image set and labeling them into groups by their identities.

Human co-detection has its notable merit in many practical computer vision applications. For example, it can help group personal photos not only with face similarity, but also based on respective appearance. Fig. 1 shows an example. Current commercial systems, such as Picasa or facebook, have already provided the human grouping function based on face similarity. These methods work well for frontal faces, but could be less stable for others. Previous human identity grouping research [27, 20, 1] extends faces to torsos, given the fact that a person appearing in multiple images taken in the same day or during the same event often wears the same clothes. Since face detector is vulnerable to head-pose variation, not to mention occlusion or back views. This makes existing approaches have various limitations when handling challenging data. In this regard, a reliable human detector would be vastly valuable to the community.

For most classical single-image human detectors [5, 29, 6], input images generally contain pedestrians in standing or walking poses. Our method relaxes this latent constraint in detecting and grouping persons, thus working on data that could fail conventional human detectors and human template matching. Moreover, the possible high variation of backgrounds in different images would make the human template matching really challenging.

Our system makes three assumptions following common knowledge to make the co-detection problem tractable. First, each image in $\mathcal{I}$ contains a subset of humans in $\mathcal{H}$. Second, each person can only appear once in one image. Third, only a person with self-consistent appearance in multiple images can be grouped. Obviously, if one is with different dressings across images, identification based on body appearance is almost impossible.

To efficiently utilize the human co-occurrence information in multiple images, we develop a human *co-detection*

*Both authors contributed equally to this work.

*and labeling* (CoDeL) framework. It is in a semi-supervised learning manner, since the number of manually annotated regions is limited [8] and most existing human detection datasets [6] seldom offer labeling about whether two detected persons actually correspond to the same one.

Our CoDeL follows a co-training scheme [4, 30] to consider insufficiently labeled data. It trains two classifiers, including a detector and a matching classifier, based on two feature sets, which are conditionally independent given the class labels. The detected region gives an indication on which part of the image is used for matching. Meanwhile, the matching process among different images can help retrieve missing data as well as rejecting false alarm for the detector.

In particular, for the detection part, we resort to the deformable part-based model [10, 8], which exploits edge information, like HOG [5], to distinguish human from background. Part-based model represents a human hypothesis as multiple flexible parts. It reduces background noise and is also robust to deformation of human body, which often occurs. For the matching process, we build a potential function through a Gaussian process [15]. It not only incorporates the similarity between any two parts [2] but also measures the similarity between two human hypotheses, thus being robust to partial appearance variation. It functions if two candidate regions have similar color and texture features. From another point of view, the two feature sets used by the two classifiers are conditionally independent given human labels.

With the initial annotated human regions and labeled matching region pairs, we train the part-based human detector and matching classifier respectively. We regard the positive outputs of one classifier as the weak positive samples of the other, and iterate this process until reliable classifiers are yielded. In testing, given the trained detector and matching classifier, we apply CoDeL to detect and label human regions.

Our main contributions are as follows. First, we propose an iterative co-training framework for human co-detection and labeling. Second, we design a new matching classifier to capture occurrence of the same person. Finally, we conduct experiments on two datasets, including a new one built by us with ground-truth labels. Our empirical results are satisfactory, with performance better than other alternatives.

## 2. Related Work

Previous work for human identity grouping [27, 20, 1] usually extracts visual features from face regions and clothes. Performance of the face detector is important in these methods. If faces are heading in different directions or are occluded, the detector could fail. Another stream of human identity identification [18, 21] is to handle videos via trackers. Moreover, Garg *et al*. [9] matches human in crowd images given the user input as initial label to retrieve under a small-motion assumption. Beyond matching given detected results, Sivic *et al*. [19] extended the contextual information in family album to improve detection. It can avoid missing persons from face detector. It applies a pictorial structure model on human parts (hair, face and torso), which can be regarded as a special version in our general framework.

Detecting human is a substantial hot topic in computer vision. Started with [5, 29], the problem is addressed via a two-stage framework including feature extraction and classifier building. Dollár *et al*. [6] provided a comprehensive survey on it. From the feature perspective, histogram of gradients (HOG) [5] forms a prominent type. Follow-up methods extend it to combination with color [14], texture feature [23, 25], etc. In terms of classifiers, linear SVM [5, 25], Ada-boost [29] and partial least square analysis [17] are among the mainstreams. Most previous human detectors concentrate on pedestrians, where sliding windows are adopted. For general human bodies with large deformation, object detector trained on human datasets performs better. Representative methods include implicit shape model [26], latent hough transform [16] and deformable part-based model [8] where the latter one provides leading performance as reported in several recent VOC competitions.

Recently, image sets with similar foregrounds were used in several applications. Kim *et al*. [11] proposed a multiple foreground co-segmentation method, where images are captured for the same group of humans or in the same scene. Bao *et al*. [2] introduced object co-detection, which finds matched objects from two or multiple related images. It provides a promising direction where the similar-foreground assumption gives an essential clue to improve detection. Although this method incorporates a unified energy function on both detection and matching, the two steps are optimized separately. Besides, this model requires a relatively large amount of labeled matching objects in training.

## 3. Co-Detection and Labeling

Given a general human detection training set and an additional small set with matching labels, we aim to build a human *co-detection and labeling* (CoDeL) solution. We start with the human representation.

### 3.1. Human Representation

Following the convention of star models in [8], our representation contains a root filter $r$ and $K$ part filters denoted as $\mathcal{P} = \{p^1, \ldots, p^K\}$. Since face is potentially important as reported in [27, 20, 1] and the technique for detecting faces [24] is mature, we add the face filter $f$ as an additional constraint for human hypothesis. As long as the ratio of overlapping area between human bounding box

and face exceeds a predefined threshold (set to 0.5 in our experiments), face and human are grouped together. The overall human model is defined as $H = (r, \mathcal{P}, f)$.

### 3.2. Energy Function for CoDeL

The goal of our CoDeL model is to incorporate the human detecting and matching classifiers in the same framework, so that the two classifiers could help improve each other by adding weak positive samples according to their classification results. The input contains $N$ images denoted as $\mathcal{I} = \{I_1, \ldots, I_N\}$. Our CoDeL framework aims to detect human regions as $\mathcal{H} = \{H_1, \ldots, H_M\}$, and give pair-wise matching scores via the matching classifier. Its energy function contains two parts, expressed as

$$E(\mathcal{H},\mathcal{I}) = \sum_{n=1}^{N} \sum_{H_i \in I_n} \Big\{ E_u(H_i, I_n) + \sum_{l=n+1}^{N} E_m(H_i, I_n, \mathcal{H}_l, I_l) \Big\}, \tag{1}$$

where $H_i$ is the $i^{th}$ human hypothesis in $\mathcal{H}$. The restriction $H_i \in I_n$ confirms that $H_i$ is detected within image $I_n$. $\mathcal{H}_l$ is the set of all human hypotheses in image $I_l$. $E_u$ in Eq. (1) is the unary potential term, which measures the compatibility between human hypothesis $H_i$ and observed image $I_n$. $E_m$ is the matching potential term, measuring pairwise similarity between $H_i$ in image $I_n$ and human hypotheses set $\mathcal{H}_l$ in image $I_l$.

Specifically, the unary potential $E_u$ for human hypothesis $H_i$ in image $I_n$, which is the detection classifier in our CoDeL model, is defined as

$$E_u(H_i, I_n) = E_f(f_i, I_n) + E_h(r_i, \mathcal{P}_i, I_n), \tag{2}$$

where $E_f$ is the potential which indicates the likelihood of containing a face in the area. We give this face potential score via the statistical explanation of Ada-boost in [28] as $E_f = 1/(\sum_i \exp\{-y_i g(f_i, w_f)\}+1)$, where $g(f_i, w_f)$ is a convex function defined on the face region as the sum over weighted outputs of weak classifiers, $w_f$ is its parameter set, and $y_i$ is the classifier label in this region. The higher the value $E_f$ is, the more likely the region contains a face. $E_h$ measures the compatibility between image $I_n$ and part-based human hypothesis $H_i$ represented by $\{r_i, \mathcal{P}_i\}$. We adopt the star model in part-based representation as

$$E_h(r_i,\mathcal{P}_i,I_n) = E_r(r_i,I_n) + \sum_{k=1}^{K} E_p(p_i^k,I_n) + \sum_{k=1}^{K} E_c(r_i,p_i^k,I_n), \tag{3}$$

where $E_r$ and $E_p$ are the unary potentials for the root and part filters respectively. $E_c$ provides the connecting potential for deformation cost between root $r_i$ and each part $p_i^k$. We define the energy $E_r$, $E_p$ and $E_c$ following those of [8].

The second term $E_m(H_i,I_n,\mathcal{H}_l,I_l)$ in Eq. (1) defines the human hypothesis level matching potential between $H_i$ in image $I_n$ and the set of hypotheses $\mathcal{H}_l$ in $I_l$. Perfect matching of the same person contributes a large value to this term. We model matching as

$$E_m(H_i, I_n, \mathcal{H}_l, I_l) = \mathcal{T}(\max_{H_j \in I_l} \hat{E}_m(H_i, H_j), t), \tag{4}$$

where $\mathcal{T}(x,t)$ is a threshold function to measure the similarity between $H_i$ and the best matched human hypothesis $H_j$ in image $I_l$. Matching is established when $\mathcal{T}(x,t) = x$ given the best matching score $x \geq t$; otherwise $\mathcal{T}(x,t) = 0$. With this threshold, only human hypothesis pairs with similarity scores larger than $t$ can contribute to the final energy. It can avoid establishing excessive or incorrect matching linkage between any two human pairs. $\hat{E}_m(H_i, H_j)$ reports the similarity between two human hypotheses $H_i$ and $H_j$. Based on common sense that each person only appears once in an image, only the largest potential of $H_i$ with respect to all $H_j$ in image $I_l$ has the chance to compete for matching. We define $\hat{E}_m(H_i, H_j)$ as a biased log marginal likelihood:

$$\hat{E}_m(H_i, H_j) = \log p(y_{ij} = 1 \,|H_i, H_j\,) + C, \tag{5}$$

where $C$ is a constant to ensure positive $\hat{E}_m(H_i, H_j)$ and $y_{ij}$ is a label, setting to 1 when $H_i$ and $H_j$ are matched, $-1$ otherwise. To describe the marginal likelihood explicitly, we introduce a latent function $\lambda$ and transform the matching value to obtain a valid probability measure as

$$p(y_{ij} = 1 \,|H_i, H_j\,) = \sigma(\lambda(H_i, H_j)), \tag{6}$$

where $\sigma$ is a logistic function. For modeling the latent function $\lambda$ effectively, we adopt *Gaussian process* (GP) [15] as a nonparametric prior, making the overall marginal likelihood a Gaussian process classifier. In particular, the input of $\lambda$ is defined as the difference between two stacked feature vectors extracted from parts of human hypotheses respectively. The correspondences of parts for two human hypotheses are obtained similarly as in the part-base model [10]. The covariance function of GP classifier captures the relative importance of both different features and different parts. It is robust against scattered occlusion.

In the overall energy Eq. (1), the unary potential corresponds to a classifier based on face and part-based human detectors, which largely rely on edge information. The matching potential, differently, contains a classifier taking part-level similarity scores measured as difference upon color and texture features. The feature sets (edge vs. color-texture) are conditionally independent given human labels, since edges are used to distinguish between human and non-human while color and texture are responsible for measuring similarity of two human hypotheses. The two classifiers supplement each other. When the labeled training data are not enough, we can use positive samples produced by one classifier as weak positive ones for updating the other.

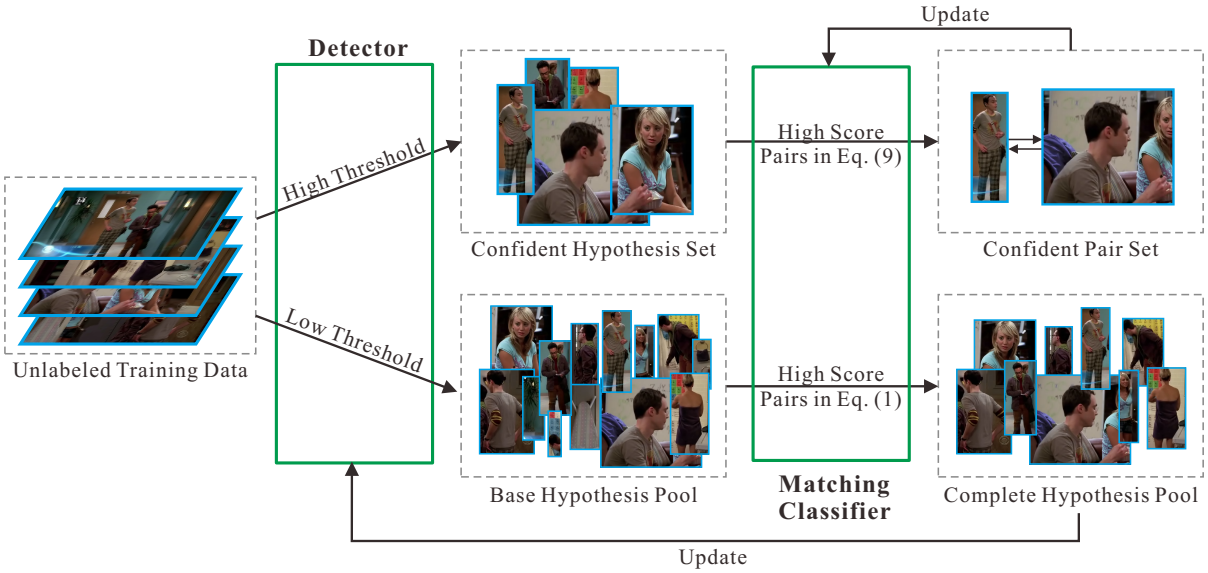


Figure 2. CoDeL co-training framework.

This semi-supervised manner is named *co-training* [4, 30] in literatures, which explains why our CoDeL framework works.

### 3.3. Model Learning

In model learning, given a group of training data with part of them containing detected bounding boxes and a subset with labeled human correspondences, we learn the parameters for Eq. (1), i.e., the human detector and matching classifier. To maximize Eq. (1), we first train the initial classifiers with labeled training data, and then update the unary and matching terms iteratively in a co-training manner by exploring unlabeled training data. In each round, the two terms are optimized as follows.

Since frontal faces do not often appear in our dataset, we first obtain the parameter $w_f$ in the face detector through training Ada boost [24] on common face dataset and fix it in remaining iterations. Therefore, maximizing the unary term in Eq. (2) is equivalent to

$$\begin{aligned}\arg \min_{w_r,w_p,w_c} & \frac{1}{2}(\|w_r\|_2^2 + \|w_p\|_2^2 + \|w_c\|_2^2) \\ & + \sum_{i=1}^{M} \max(0, 1 - y_i \max_{\mathcal{P}_i} E_h(r_i, \mathcal{P}_i, I_n)),\end{aligned} \tag{7}$$

where $w_r$, $w_p$, and $w_c$ are the parameters for root, part, and connecting potentials in Eq. (3) respectively. We train the part-based detector for $w_r$, $w_p$, and $w_c$ following the setting in [10, 8]. The detector is a two-component mixture model as [8] – visually one for full body and one for upper body.

For the matching term in Eq. (5), the input matched pair of GP classifier is represented as a concatenated vector of feature difference. During training GP, we choose the logistic function as the likelihood and resort to Laplace approximation for calculating the desired posterior of the latent function $\lambda$. More specifically, the posterior $p(\lambda|\, H_i, H_j, y_{ij})$ is approximated by a Gaussian $\mathcal{N}(\lambda|\, \hat{\lambda}, K)$. $\hat{\lambda}$ and $K$ are the approximated mean function and covariance matrix respectively. Our GP covariance function is a full squared exponential. By Bayes rule, the log-posterior $\Psi(\lambda) = \log p(\lambda|\, H_i, H_j, y_{ij})$ is expressed as

$$\Psi(\lambda) \propto \log p(y_{ij}\,|\lambda) + \log p(\lambda|\, H_i, H_j), \tag{8}$$

where $p(y_{ij}\,|\lambda)$ is the logistic likelihood function. $\hat{\lambda}$ and $K$ are given by

$$\begin{aligned}\hat{\lambda} &= \arg\max_{\lambda} \Psi(\lambda), \\ K &= -\nabla\nabla\, \Psi(\lambda)|_{\lambda=\hat{\lambda}},\end{aligned} \tag{9}$$

where $\nabla\nabla$ is the Hessian operator. Note Eq. (9) can be solved efficiently via the Newton-Raphson method [15].

Based on the above two updating steps, we perform co-training to generate new weak labeled positive samples. The flowchart is shown in Fig. 2. We first learn two initial classifiers and use the initial trained human detector to test new unlabeled images. Since output of the detector contains no label, they cannot be directly employed by the successive matching classifier. To overcome this problem, we build a confidence criterion based on the probabilistic property of the GP classifier, which lets the mean prediction of the GP learned in last round determine whether two new hypotheses produced by the detector match. The mean prediction of the two human hypotheses $H_i^*$ and $H_j^*$ in the

GP classifier is defined as

$$\bar{y}^*_{ij} = \int p(y_{ij}\,|\lambda^*)p(\,\lambda^*|\,\mathcal{H}, H^*_i, H^*_j)d\lambda^* \qquad (10)$$

where $\lambda^*$ is the current latent function corresponding to the test pair and $\mathcal{H}$ is the initial training human hypotheses set. When $\bar{y}^*_{ij}$ is above a threshold (0.8 in our experiment), these two human hypotheses are denoted as weak positive and are added to the training data of matching in the next round. This confidence criterion essentially shares the same insight with the rejection option studied in GP classification [15, 3] and provides a very useful hint on finding positive data. Statistical properties of this confidence criterion are demonstrated by the error-reject curve in Section 4.3. Since these input hypotheses are selected from output of the detector with high unary scores, we pass these hypotheses pairs to retrain the matching classifier, illustrated in Fig. 2.

Given the updated matching classifier, we retrieve weak positive human hypotheses to train the detector, shown in the bottom row in Fig. 2. First, a base hypotheses pool is generated by a human detector with a low unary score threshold, thus with high recall. For each pair of hypotheses in this base pool, we calculate the total energy in Eq. (1) and discard those with low scores to construct the final complete hypothesis pool. Since the total energy indicates the confidence of a human region, we retrain our detector with data remaining in this complete hypothesis pool.

These two steps iterate in a way to test new data and add them to the training set when confident. It stops when the maximum number of iterations is reached or performance is not improved anymore. Thus, with only a small amount of labeled training data, we can efficiently learn our CoDeL framework via the semi-supervised co-training setting.

### 3.4. Model Inference

In model inference, our goal is to detect human hypotheses and report their corresponding labels on new data given the human detector and matching classifier. We use the face and part-based human detectors to find candidates. The detector is tuned to a low threshold to achieve high recall so that most candidates are included. Then we adopt the GP classifier on each pair of human hypotheses to get the matching score via Eq. (10).

If the total score for a particular human body in Eq. (1) is above 1.5, we label it as a human body. Therefore, the final confidence includes both the unary and matching scores. If a human region finds similar ones in other images, which are also labeled as human, it becomes more confident. Meanwhile, the detected false-alarm regions have relatively low unary scores and could hardly find matches among other human regions.

With this process, we quickly increase precision and preserve the recall of detection. After we obtain all detected human regions and their pairwise matching potential scores, we assign the human labels via hierarchical clustering [22], which assures the difference of scores within a cluster is less than a predefined distance.

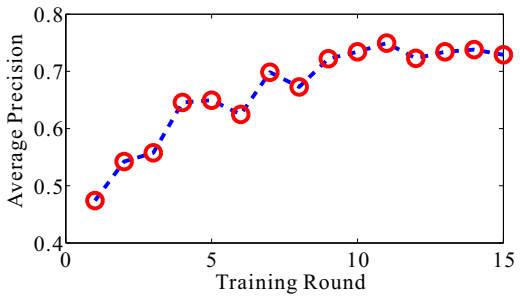


Figure 3. Co-training evaluation. The detection performance increases along with growth of semi-supervised training rounds.

## 4. Experiments

In this section, after describing our experimental settings, we evaluate our method in the following four aspects. 1) How does the co-training setting improve the performance of CoDeL in a semi-supervised manner? 2) How useful is the criteria of confidence in Eq. (10)? 3) Can our matching classifier correctly distinguish between matched pairs and others? 4) How to compare our detection results under the co-detection setting with previous single-image detection methods?

### 4.1. Experimental Settings

In our experiments, we use two datasets. One is the pedestrian dataset provided in [7] where the stereo image pairs serve as a natural source of matched pairs following the setting of [2]. The other dataset, denoted as *human co-detection dataset* (HCD dataset), is collected by us with some images from the co-segmentation dataset [11], representative frames in "Big Bang Theory" season 1, etc. It conforms to the assumptions of our human co-detection tasks – that is, each human appears in only part of the image set with consistent appearance. For the pedestrian set [7], we use 450 pairs of images to train and 354 to test as [7]. For HCD set, we provide around 400 images, 90% of which are for training and 10% for testing. This splitting is repeated 10 times for average accuracy.

To evaluate the detection performance, we report average precision (AP) following the criteria in PASCAL VOC challenge. The matching classifier is evaluated by the classification accuracy with respect to ground truth matching labels. For the part-based detector, we use the star model of DPM-v4 [8]. The training set combines training data from VOC2010, pedestrian set and HCD set. We have two threshold levels according to Fig. 2. The high threshold is

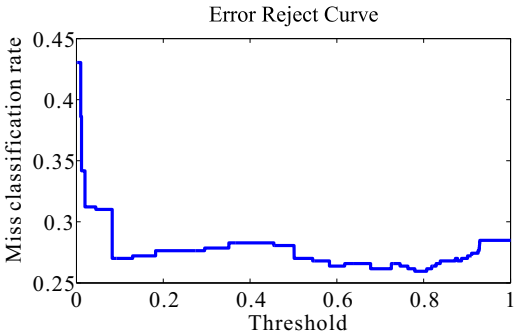


Figure 4. Statistics in the form of the error-reject curve.

| | Pedestrian [7] | HCD Dataset |
|---|---|---|
| SVM+Color | 60.39 | 52.29 |
| SVM+LBP | 54.29 | 53.21 |
| SVM+SIFT | 81.59 | 61.80 |
| SVM+Color+LBP+SIFT | 88.52 | 63.03 |
| GP+Color | 77.66 | 61.10 |
| GP+LBP | 82.08 | 62.76 |
| GP+SIFT | 88.57 | 62.73 |
| GP+Color+LBP+SIFT | **91.43** | **68.07** |

Table 1. Matching accuracy (%) for human co-detection on the pedestrian dataset [7] and our human co-detection (HCD) dataset.

set to 0.7 while the low threshold is -1.0. In the matching part, we use three sets of features to describe similarity in terms of color and texture as in Eq. (5). They include difference of color histograms (32D color histogram on the H channel of LSH color space), difference of LBP features [13], and the logarithm of matched SIFT [12] point number. Each feature represents a human hypothesis via a stacked vector among all parts. We transform the color and LBP features via PCA to 50D respectively. All parameters are tuned on a hold-out validation set.

### 4.2. Co-training Evaluation

In this section, we verify the effectiveness of our co-training process in a semi-supervised manner. We train CoDeL on an initial set of 20 images from HCD dataset with both bounding box labels and matching labels. Then in each round, we randomly pick 10 new unlabeled images from the remaining ones to conduct co-training described above. We iterate these procedures for 15 rounds and record the precision in each stage, which is evaluated on an independent testing set of 30 images. This process is repeated 10 times to compute AP. The result is shown in Fig. 3. With increase of unlabeled training data, our co-training system gradually enriches the training set by adding weak positive samples and improves the detection performance.

### 4.3. Error-Reject Curve

We study empirical properties of confidence criteria in Eq. (10), which is used to generate weak labeled human pairs to retrain the matching classifier. The mean prediction value in Eq. (10) of a testing sample is adopted as an indication of rejection. In particular, we calculate the values of all testing pairs in our HCD dataset and regard them as positive if they are above a threshold. With the threshold varying from 0.0 to 1.0, different misclassification error rates are recorded and shown in Fig. 4. We found that, when the threshold is very low, the misclassification error rate is rather high, since most negative samples are misclassified as positive ones. As the threshold goes up, the error rate drops, indicating rejection of more negative samples. Threshold around 0.8 gives the smallest error rate.

### 4.4. Matching Classifier

We evaluate the performance of our matching classifier, and show it in Table 1. All three kinds of features are tested alone with linear SVM and GP classifier. Features of matched SIFT yield good results on the pedestrian dataset because the scales of most persons are small and the majority of them are with dark clothes. In our HCD dataset, people poses are with a large variation and clothes are in different colors, making the performance of SIFT features drop. Other challenges introduced by this dataset include matching pair chosen among all images and background noise caused by deformable body parts. It leads to less perfect results for each type of features. Feature combination performs better on both datasets. Also the GP classifier is consistently better than the linear SVM classifier used in [2] due to its non-liner property.

### 4.5. Co-Detection Results

We compare our method with the widely used face detector [24], one of state-of-the-art human detectors [8] and object co-detection method [2]. The results are reported in Table 2.

The face detector cannot deal with the situation that the face is partly or completely missing. It achieves high precision and low recall on our HCD dataset. We do not evaluate this detector on the pedestrian dataset, since faces can hardly be found. Compared to single-image human detector [8], our matching classifier can increase the score of unreliable human hypotheses when they have confident matches.

We also compare our method with object co-detection [2]. The gain is partly due to the matching potential, which captures more informative clues and is robust to deformation, as shown in Section 4.4. Further, our CoDeL framework yields a larger increase on the HCD dataset than that on the pedestrian one, since HCD provides more images with potential matching pairs. We

| | Pedestrian [7] (all) | Pedestrian [7] ($h > 120$) | HCD Dataset |
|---|---|---|---|
| Face | – | – | 5.95 |
| Part-based model [8] | 59.7 | 55.4 | 69.94 |
| Object co-detection [2] | 62.7 | 63.4 | – |
| CoDeL | 74.4 | 73.8 | 74.94 |

Table 2. Average precision (%) for human co-detection on the pedestrian dataset [7] and our human co-detection (HCD) dataset. We directly quote results of the object co-detection method reported in [2].

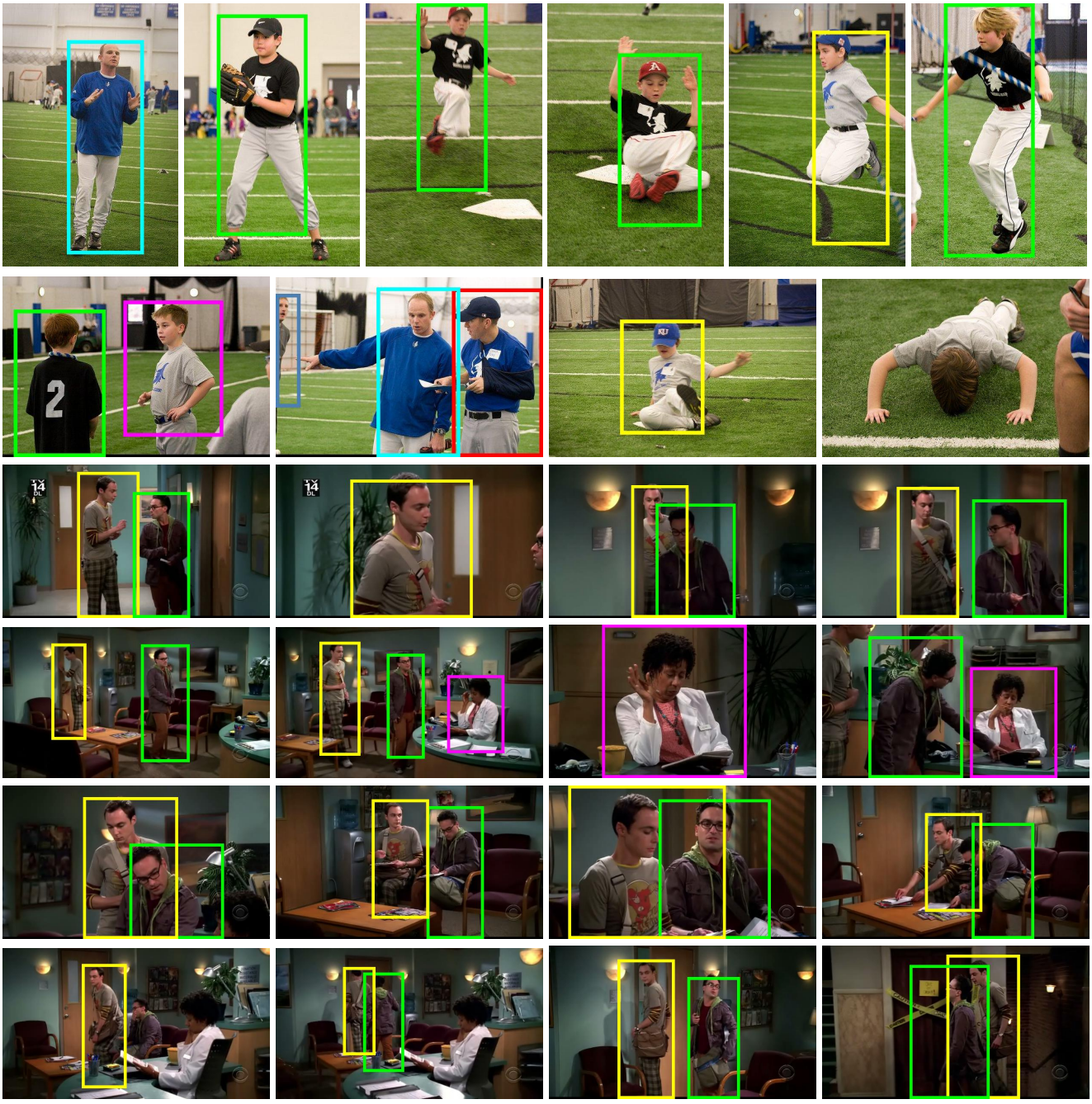

Figure 5. Visual examples for human co-detection and labeling in the baseball scene (rows 1-2) and big bang series scene (rows 3-6) in our HCD dataset. Persons are labeled in different colors.

show a few co-detection and labeling results obtained with HCD in Fig. 5. It is notable that errors may occur when two different human hypotheses are of quite similar appearances, as illustrated in the first row of Fig. 5. Also, there are several hard cases that cannot be detected, e.g., the last image of the second row and the first two images in row 6. They contain back views or poses that rarely exist in the image data.

## 5. Conclusion and Future Work

We have proposed a human *co-detection and labeling* (CoDeL) framework. It is formed in a semi-supervised manner to boost performance given insufficient labels. Also we define our matching classifier via a Gaussian process on the human hypothesis level. Experiments demonstrate our approach produces reasonable matching and detection results compared with other methods.

Our future work includes extending the matching classifier by integrating spatial relationship among parts. Also we will build an online algorithm.

## Acknowledgments

This work is supported by a grant from the Research Grants Council of the Hong Kong SAR (project No. 413110) and by NSF of China (key project No. 61133009).

## References

[1] D. Anguelov, K. chih Lee, S. B. Gokturk, and B. Sumengen. Contextual identity recognition in personal photo albums. In *CVPR*, pages 1–7, 2007.

[2] S. Y. Bao, Y. Xiang, and S. Savarese. Object co-detection. In *ECCV*, 2012.

[3] P. L. Bartlett and M. H. Wegkamp. Classification with a reject option using a hinge loss. *The Journal of Machine Learning Research*, 9:1823–1840, 2008.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[6] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *PAMI*, 34(4):743–761, 2012.

[7] A. Ess, B. Leibe, and L. V. Gool. Depth and appearance for mobile scene analysis. In *ICCV*, 2007.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.

[9] R. Garg, D. Ramanan, S. M. Seitz, and N. Snavely. Where's Waldo: matching people in images of crowds. In *CVPR*, pages 1793–1800, 2011.

[10] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/ rbg/latent-release5/.

[11] G. Kim and E. P. Xing. On multiple foreground cosegmentation. In *CVPR*, pages 837–844, 2012.

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[13] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, 24(7):971–987, 2002.

[14] P. Ott and M. Everingham. Implicit color segmentation features for pedestrian and object detection. In *ICCV*, pages 723–730, 2009.

[15] C. E. Rasmussen. *Gaussian processes for machine learning*. Citeseer, 2006.

[16] N. Razavi, J. Gall, P. Kohli, and L. V. Gool. Latent hough transform for object detection. In *ECCV*, pages 312–325, 2012.

[17] W. R. Schwartz, A. Kembhavi, D. Harwood, and L. S. Davis. Human detection using partial least squares analysis. In *ICCV*, pages 24–31, 2009.

[18] J. Sivic, M. Everingham, and A. Zisserman. Who are you?-Learning person specific classifiers from video. In *CVPR*, pages 1145–1152, 2009.

[19] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *BMVC*, 2006.

[20] Y. Song and T. Leung. Context-aided human recognition–clustering. *ECCV*, pages 382–395, 2006.

[21] M. Tapaswi, M. Bauml, and R. Stiefelhagen. Knock! Knock! Who is it? probabilistic person identification in TV-series. In *CVPR*, pages 2658–2665, 2012.

[22] H. Trevor, T. Robert, and J. H. Friedman. *The elements of statistical learning*. Springer New York, 2001.

[23] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *PAMI*, 30(10):1713–1727, 2008.

[24] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[25] X. Wang, T. X. Han, and S. Yan. An HOG-LBP human detector with partial occlusion handling. In *ICCV*, pages 32–39, 2009.

[26] P. Wohlhart, M. Donoser, P. M. Roth, and H. Bischof. Detecting partially occluded objects with an implicit shape model random field. In *ACCV*, 2012.

[27] L. Zhang, L. Chen, M. Li, and H. Zhang. Automated annotation of human faces in family albums. In *ACM international conference on Multimedia*, pages 355–358, 2003.

[28] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.

[29] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *CVPR*, pages 1491–1498, 2006.

[30] X. Zhu and A. B. Goldberg. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130, 2009.