Transferring a Generic Pedestrian Detector Towards Specific Scenes

Meng Wang, Wei Li and Xiaogang Wang

Department of Electronic Engineering, The Chinese University of Hong Kong

香港中文大學
The Chinese University of Hong Kong

Motivations

- ***A generic detector doesn't work well on a specific scene.***
- Adapting a generic pedestrian detector to a specific scene requires automatically selecting reliable training samples from the target scene using contextual information (such as motion and scene structure).
- Imperfection of the automatic labeler results in mislabeling, which causes drifting or slow convergence of the scene-specific detector.
- ***Our Goal***: Automatically train an error-resilient scene-specific detector with fast convergence under the *transfer learning* framework.

Strategies

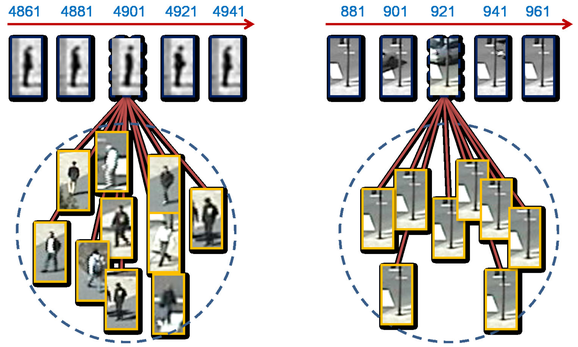
- The source samples used to train the generic detector are re-weighted to match the sample distribution in the target scene.
- Contextual information is used to compute confidence scores of samples from the target scene to guide transfer learning.
- Soft labels with confidence scores are resilient to occasional errors brought out by hard thresholding.
- Confidence scores propagate among samples on a graph according to the underlying visual structures of samples.
- All these considerations are formulated under a single objective function called ***Confidence-Encoded SVM***.

Confidence-Encoded SVM

$$
\begin{aligned}
\min_{\mathbf{c},\mathbf{w},b} \quad & \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n_s}(\nu_i\xi_i^s)^2 + C\sum_{j=1}^{n_t}(c_j\xi_j^t)^2 + \\
& \frac{\mu}{2}\mathbf{c}^T\mathbf{L}\mathbf{c} + \frac{\lambda}{2}(\mathbf{c}-\mathbf{c}_0)^T\mathbf{A}(\mathbf{c}-\mathbf{c}_0) \\
\text{s.t.} \quad & y_i^s(\mathbf{w}^T\mathbf{x}_i^s + b) \geq 1 - \xi_i^s, i = 1, \cdots, n_s \\
& y_j^t(\mathbf{w}^T\mathbf{x}_j^t + b) \geq 1 - \xi_j^t, j = 1, \cdots, n_t \\
& \xi_i^s \geq 0, i = 1, \cdots, n_s \\
& \xi_j^t \geq 0, j = 1, \cdots, n_t.
\end{aligned} \tag{1}
$$

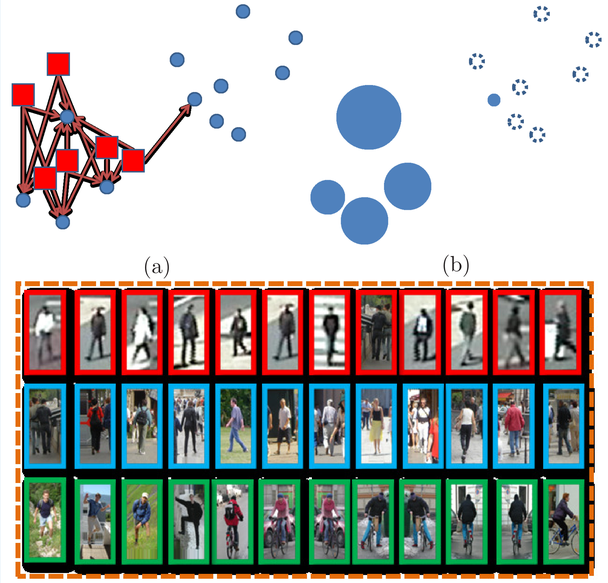
- Confidence-based soft labels $\nu_i, c_j \in [-1, 1]$ are used in lieu of hard labels.
- It is more error-resilient and robust to drifting.
- Confidence is propagated through appearance similarity using the Graph Laplacian term.

Confidence Propagation



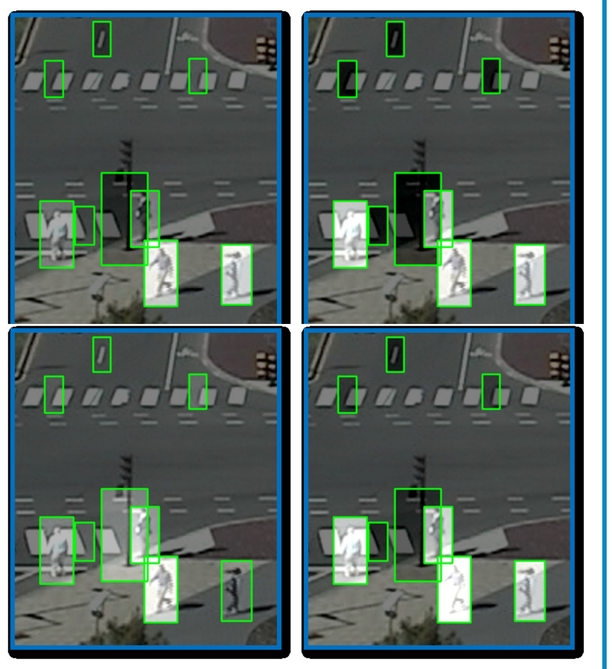
- **Positive Examples (left)**. A pedestrian is stationary for a long period and therefore is mistakenly labeled as a negative sample with a high initial high confidence score (close to -1) according to the motion cue. Its confidence score gets close to 0 after confidence propagation because many other samples with similar visual appearance to it are labeled as positive samples with high confidence scores. Therefore its negative influence is alleviated.
- **Negative Examples (right)**. A background patch is labeled as a negative sample with a low initial confidence score (close to 0) because a vehicle happens to pass by and causes motions. Its confidence score becomes higher (close to -1) after confidence propagation because some similar background patches are labeled as negative samples with high confidence scores.

Re-weighting Source Samples



- In (a) and (b) , red squares indicate target samples and blue points indicate source samples. Each target sample has K directed edges pointing towards it K nearest neighbors in the source set. If a source sample is outlier of the target set, it has a small indegree.
- In (c), target samples (first row), source samples with large indegrees (second row), and source samples with zero indegree (third row) are illustrated for the positive case.

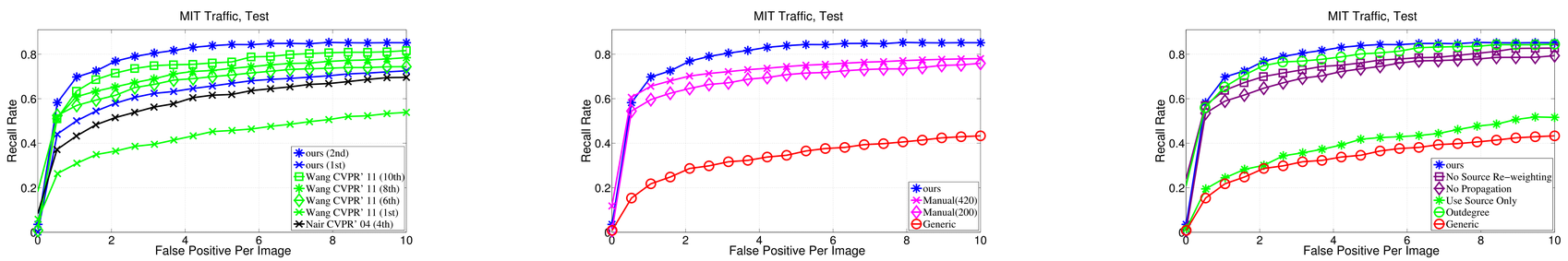
Iterative Optimization



(a) (b)

- The confidence scores (a) and detection scores by SVM (b) change after three iterations when optimizing the Confidence-Encoded SVM. A bright window indicates that the score is close to $+1$ and a dark window indicates that the score is close -1. After 3 iterations, the two types of scores look more consistent and correct.

Experiment Results



(a) (b) (c)

- Results on the MIT Traffic dataset. (a) compares with two automatic scene adaptation approaches (Wang CVPR'11 and Nair CVPR'04) on the testing sets after different rounds of training. Our approach quickly converged after two rounds of iteration. (b) compares with the generic detector and the scene-specific detector trained on different numbers of manually labeled frames. (c) shows the effectiveness of techniques, including the absence of source re-weighting, confidence propagation and the indegree graph.