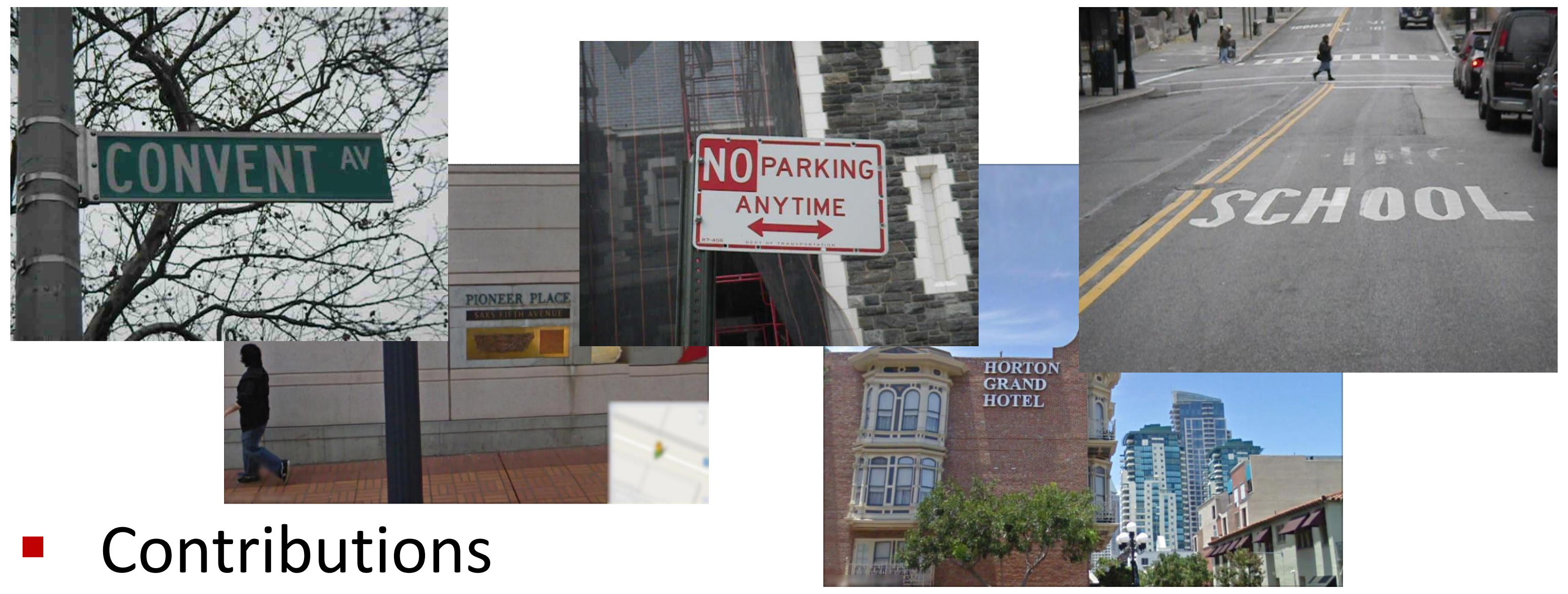
Scene Text Recognition in Multiple Frames Based on Text Tracking

Xuejian Rong[1], Chucai Yi[2], Xiaodong Yang[1] and Yingli Tian[1,2]

[1]The City College, [2]The Graduate Center, City University of New York

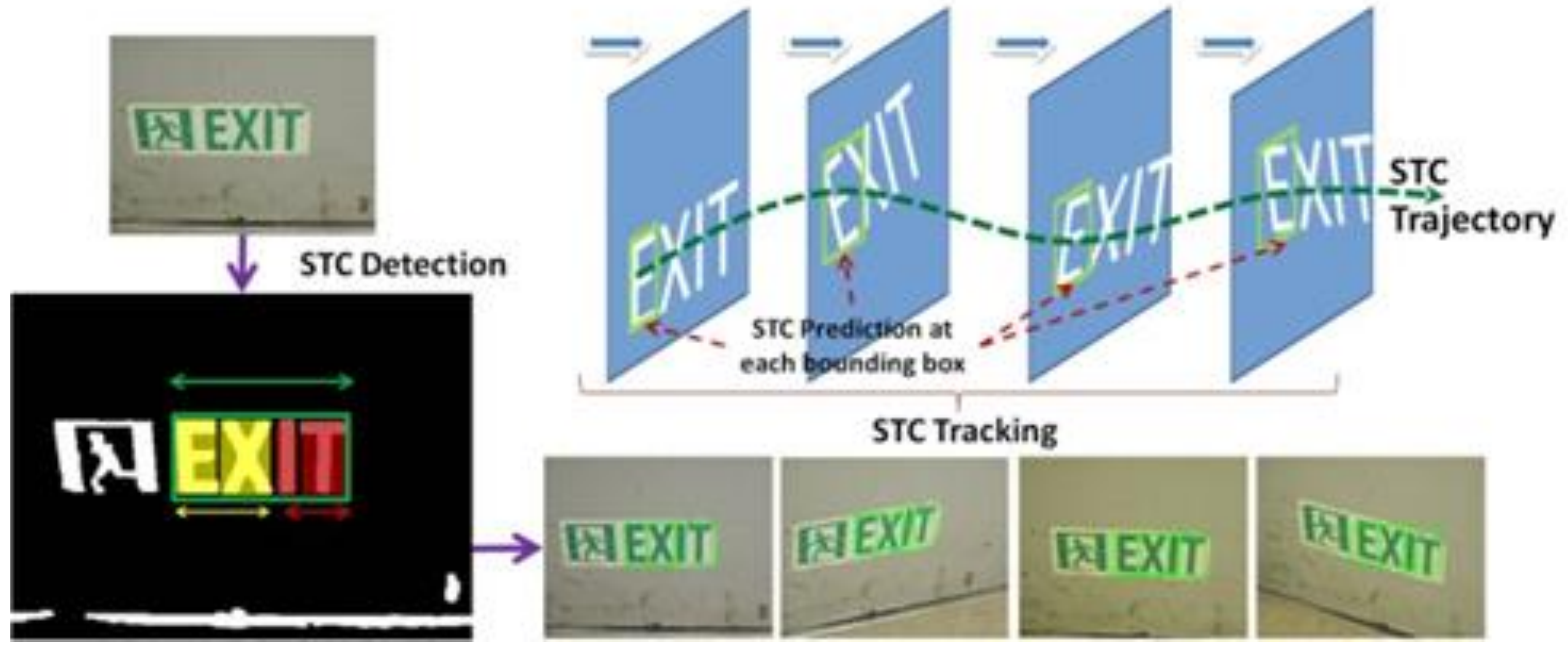
Overview

- Scene text recognition in multiple frames
  - More and more real-world raw data and applications of **text information retrieval** are based on video frames rather than a single scene image
  - **Frame relationships** are ignored in traditional single image based scene text recognition methods
  - **Text tracking** algorithms improve detection results a lot.
- Our goal: make Scene Text Character (**STC**) prediction and word configuration in multiple frames based on text tracking to improve the performance of scene text recognition



- Contributions
  - Uniform tracking-based video text recognition framework
  - Novel feature representation of STC with dense sampled SIFT descriptors and Fisher Vector
  - New collected Video Text Reading dataset (**VTR**)

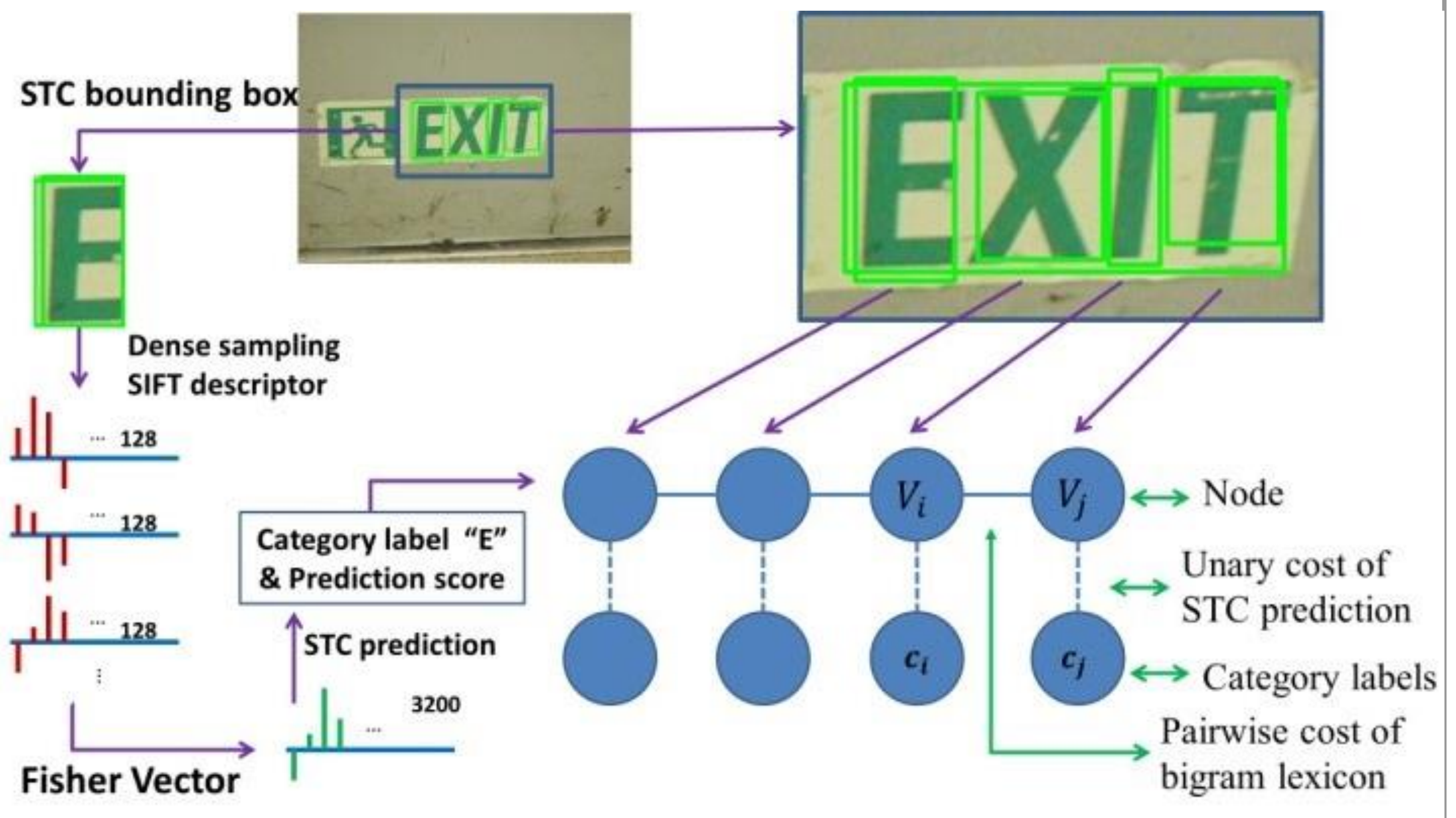
Scene Text Detection and Tracking



After initial text regions detection, each STC is independently tracked with multi-object tracking methods. A trajectory is then estimated by merging the tracked STC bounding boxes in each frame.

Scene Text Character Representation

- Word configuration in conditional random field (**CRF**)



1. STC extracted by detection and tracking and then transformed into Fisher Vector feature, which has the following merits compared with Bag-of-words (BOW):
   1) Fisher Vector performs quite well with simple linear classifiers which are efficient in both training and testing.
   2) The Fisher Vector can be computed upon a much smaller visual vocabulary which significantly reduces the computational cost.
2. SVM-based STC predictor is applied to obtain it category label and prediction score
3. In a tracking frame of scene text, each STC bounding box is defined as a node in CRF model

Cost function of CRF model for STC bounding box

$$L(V = c) = \sum_{i=1}^{|V|} L_i(V_i = c_i) + \lambda \sum_{(i,j)\in E} L_{ij}(V_i = c_i, V_j = c_j)$$

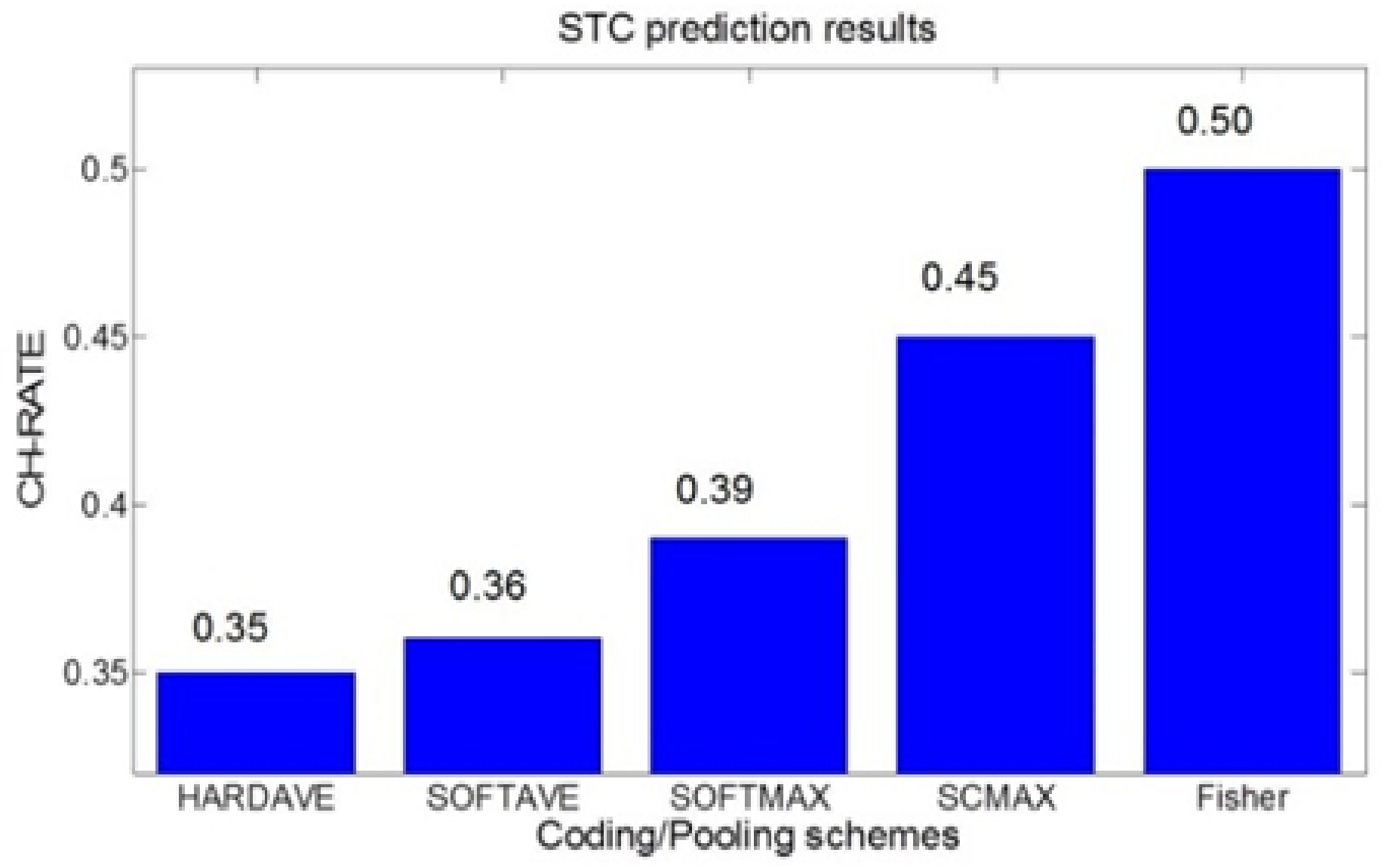
V : the set of all nodes, c : category label, L_i & L_{ij} : cost function of single & neighboring node

4. Unary cost of STC prediction and pairwise cost of bigram lexicon is defined in the graphical model

Experimental Results

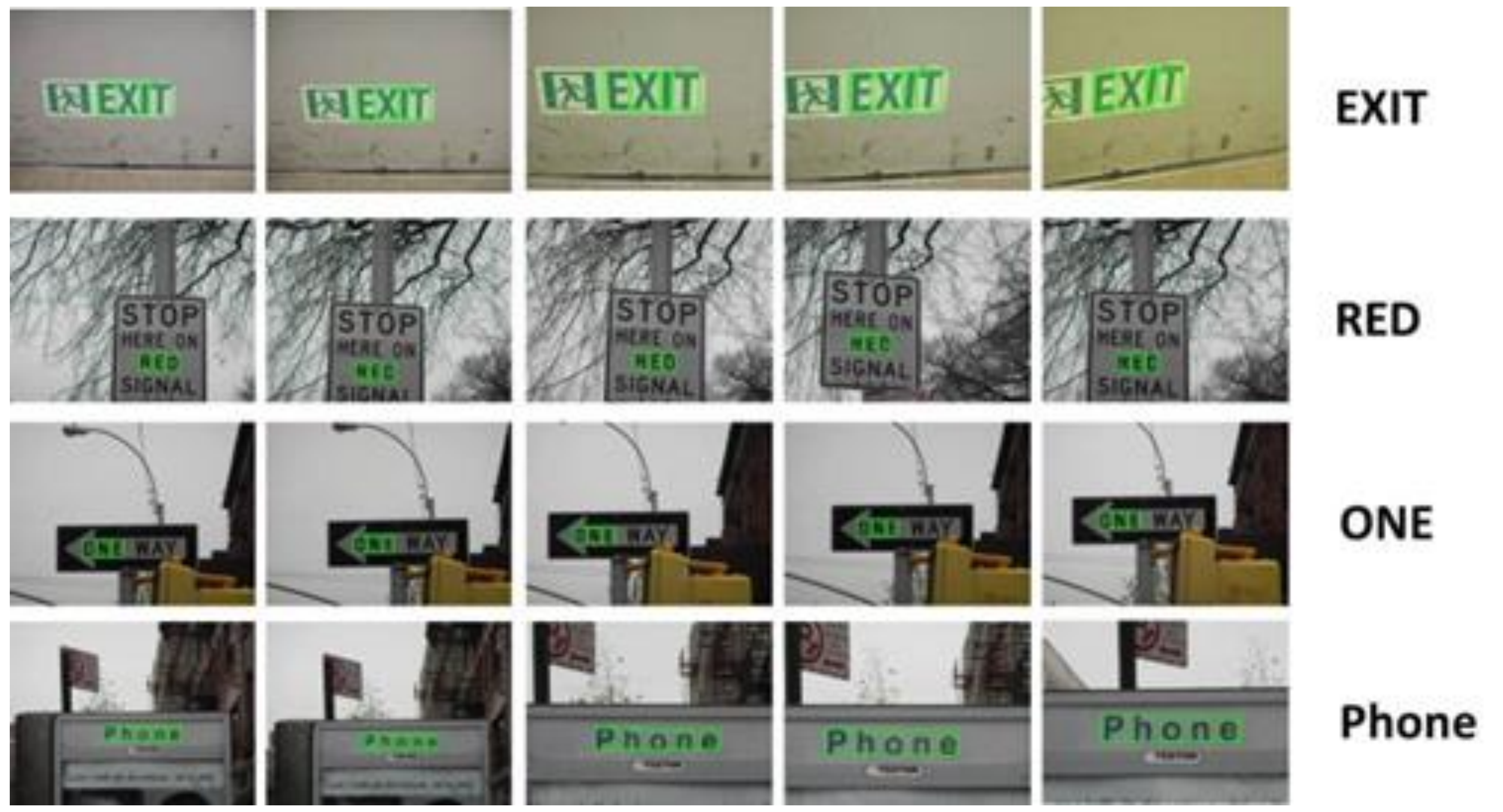
- Datasets
  - CHARS74K, ICDAR2003, ICDAR2011,
  - Our collected Video Text Reading dataset (VTR)
- Settings and Evaluations
  - Chars74k samples used to train STC predictor
  - First three datasets used to make statistics of bigram frequency in lexical analysis
  - VTR dataset used to validate the effectiveness
- Results

STC prediction results in Chars74K



	CH-RATE	ED0	ED1
Single Frame	0.640	0.333	0.583
Multi-Frame (Majority)	0.680	0.389	0.611
Multi-Frame (CRF)	0.713	0.389	0.722

Video-based Scene Text Recognition



- Conclusion

We have proposed an uniform effective tracking-based scene text recognition framework on multiple frames.

Future work will focus on designing more robust fusion methods to incorporate STC prediction scores of multiple frames to further improve the performance of word-level text recognition.

Sponsors: NSF grant EF1137172, IIP-1343402, and FHWA grant DTFH61-12-H-00002