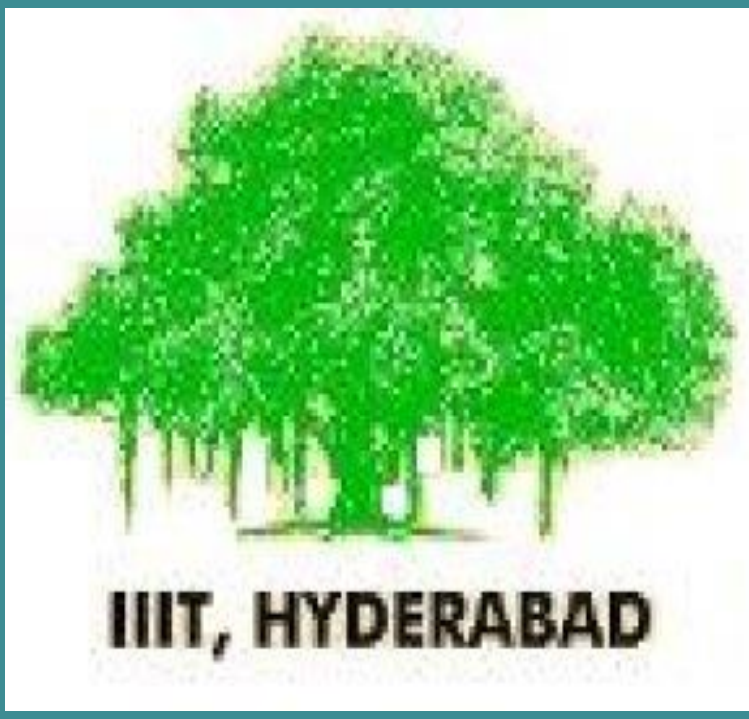




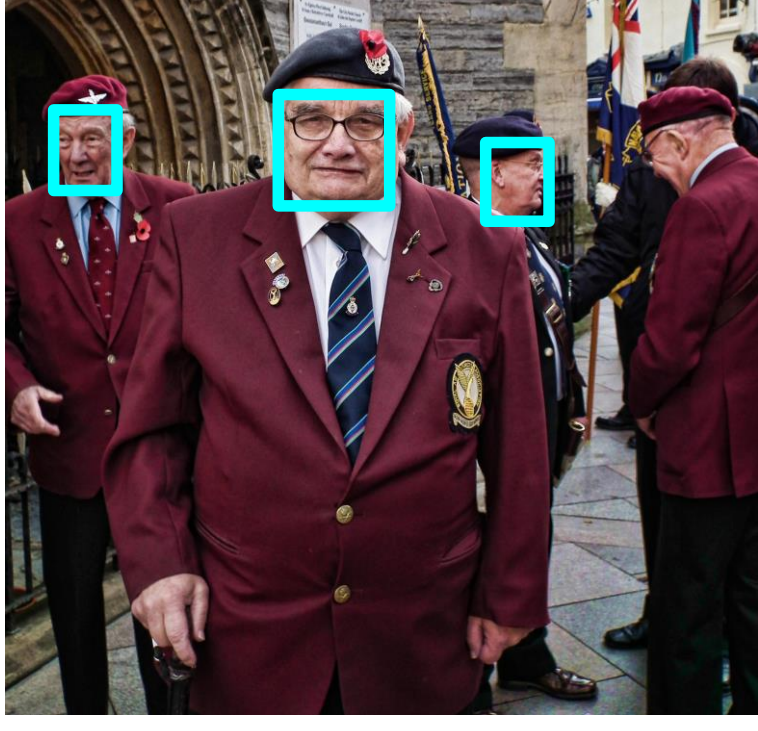
Visual Phrases for Exemplar Face Detection

Vijay Kumar, Anoop M. Namboodiri and C. V. Jawahar
CVIT, IIIT Hyderabad, India



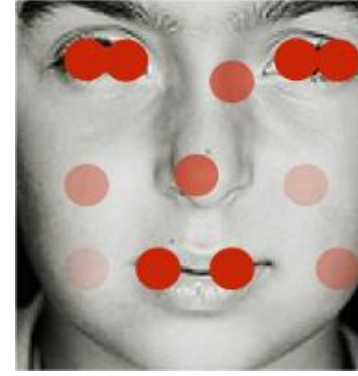
More details can be found at our project page: <http://cvit.iiit.ac.in/projects/exemplar>

1. Motivation



- Pose, extreme lighting, occlusion, scale present serious challenges for face detection
- Traditional model based methods that generalizes large number of training examples may not perform well with unseen variation
- Exemplar-based approaches, on the other hand, uses large number of 'exemplar' images which vote independently for locating the faces
- We further improve the exemplar approach using two key ideas:

- We propose better weighting scheme for visual words that suppresses the noisy visual words
- We discover higher-order dependencies from the large database and exploit them for detection



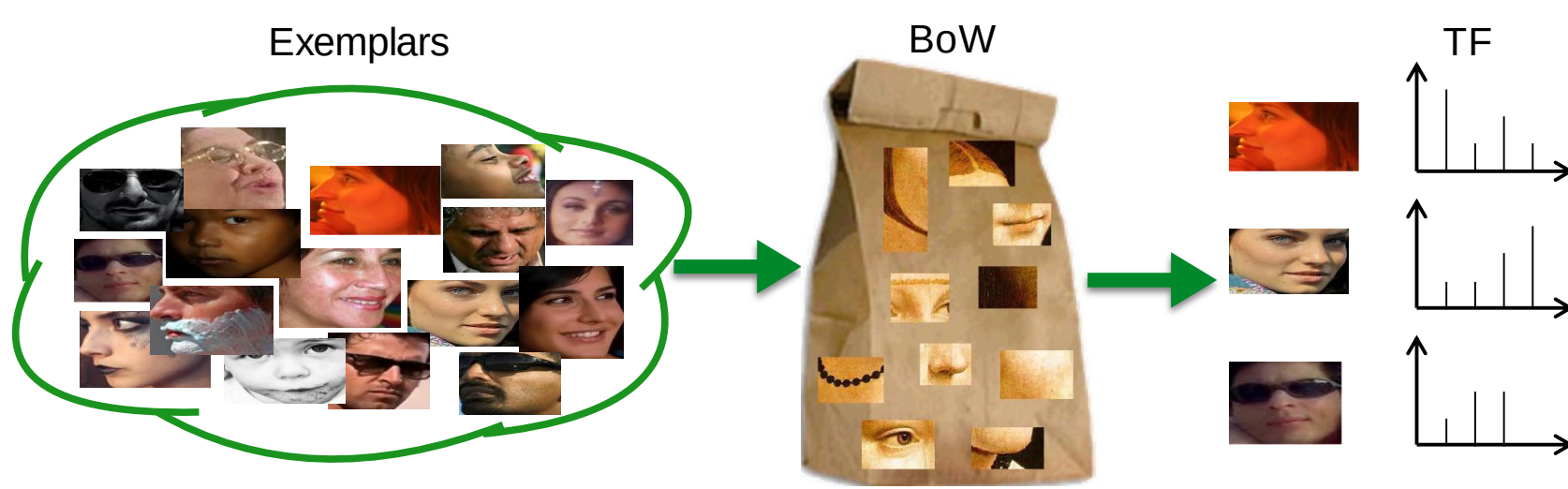
Contextual Weighting of Features



Visual Word Co-occurrences (Visual Phrases)

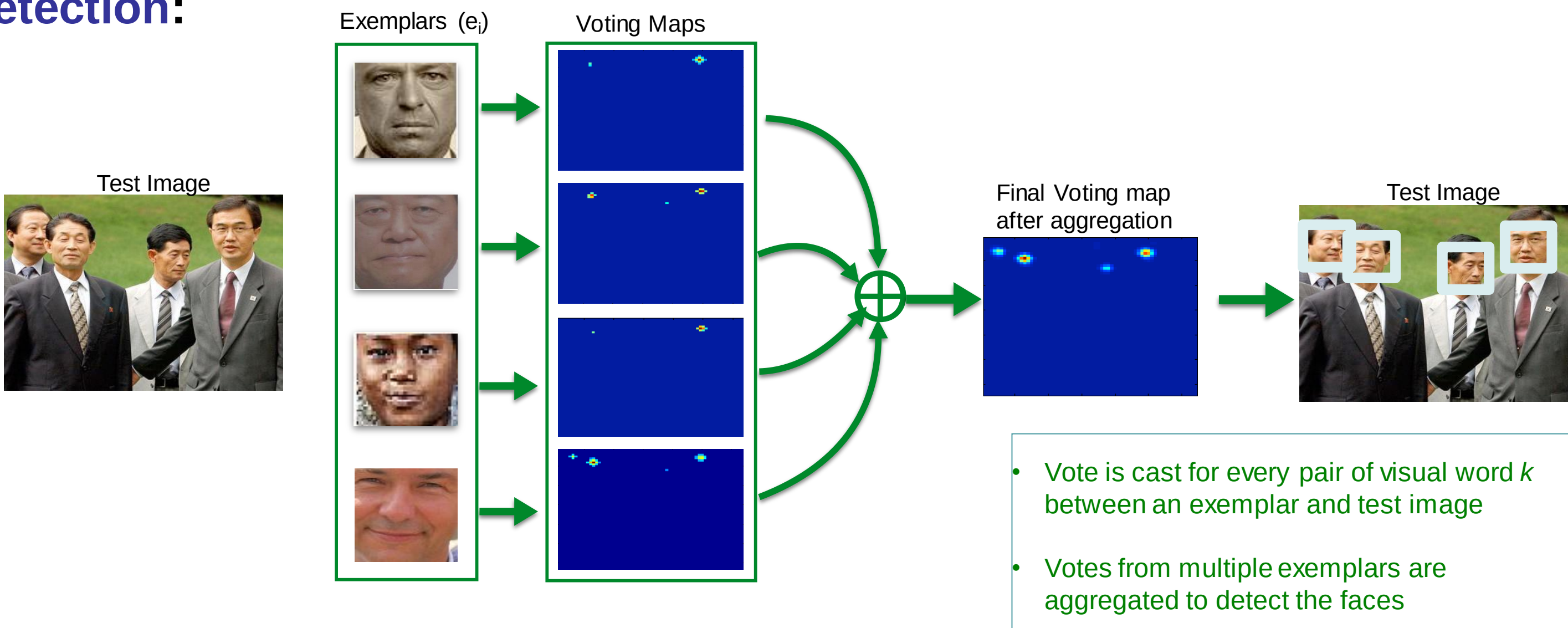
2. Exemplar Face Detection - Overview

Training:



- Large diverse set of Exemplars are collected
- Local features (dense-SIFT) are extracted and quantized using BoW technique
- TF and IDFs for each exemplar are calculated and stored.

Detection:



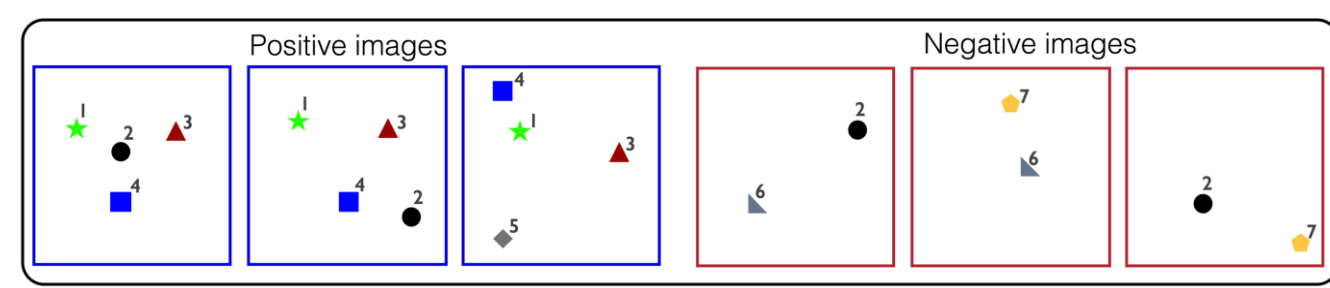
The similarity between exemplar e_i and test region p is measured by [1, 2],

$$S(p, e_i) = \sum_k \sum_{\substack{f \in R_x(p), g \in e_i \\ w(f) = w(g) = k \\ ||T(L(f)) - L(g)|| < \epsilon}} \frac{F(w(g), L(g))}{tf_{e_i}(k) \cdot tf_x(k)},$$

$tf(k)$ denote the term frequency of the visual word k

$L(f)$ and $L(g)$ are the locations of the features quantized to visual word k in test region and exemplar, respectively

3. Contextual Weighting of Features



Current exemplar schemes [2, 3] compute the similarity score using,

$$F(w(g) = k, L(g)) = idf^2(k),$$

We leverage the presence of large database and estimate the distribution of each visual word over the entire database.

$$P_e(k|\langle x, y \rangle) = \frac{1}{N_e} \sum_{e_i} \mathcal{I}(w(L(g), e_i) == k),$$

The distribution of each visual word is used to weight their individual occurrences in exemplars.

$$W(\langle x, y \rangle, k) = P_e(k|\langle x, y \rangle) * H$$

A global discriminative score for visual word k is computed as,

$$\mathcal{D}(k) = \max_{\langle x, y \rangle} \frac{P_e(k|\langle x, y \rangle)}{P_n(k|\langle x, y \rangle)}$$

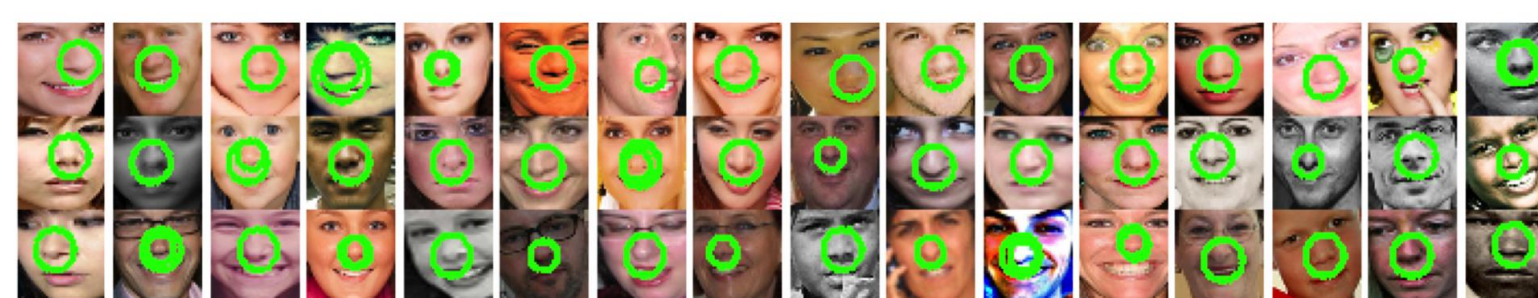
The similarity is finally computed as,

$$F(w(g) = k, L(g)) = W(L(g), k) \cdot \mathcal{D}^2(k),$$

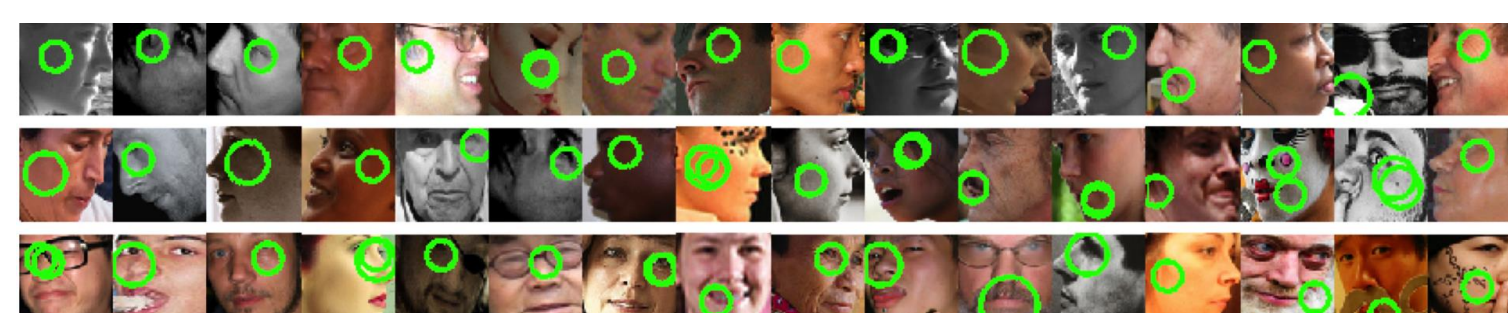
It is discriminative and down-weights noisy features

Current exemplar approaches have two issues:

- Problem 1: Similarity score is based on idfs which is less-discriminative for detection tasks
- Problem 2: All the visual words in an exemplar are considered equally important. This ignores noisy features that needs to down-weighted



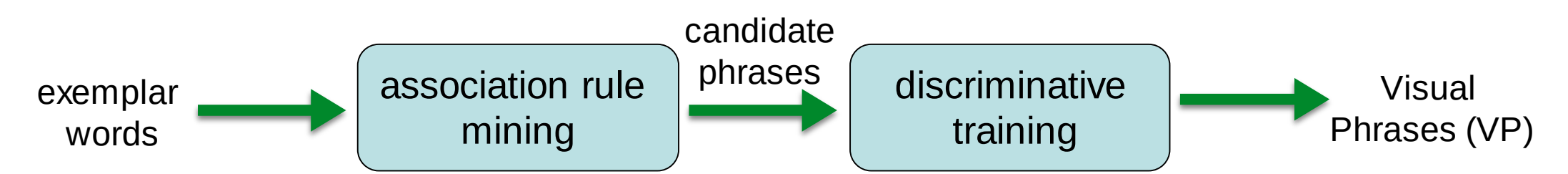
Stable visual words that occur at consistent locations and in consistent exemplars get high weightage



Non-stable visual words that occur at random locations get less weightage

4. Visual Phrases

- Visual words in faces are **highly correlated and co-occur** together
- Higher-order information provide **positive cues** about the presence of faces.
- We capture such dependencies in visual words from the large exemplar database
- Stable visual words that occur together consistently at similar locations form a **Visual phrase** (VP)



Discriminative Training:

- Visual phrases are selected through a discriminative training
- Phrases that occur at consistent locations and in consistent images are retained

$$\mathcal{Q}(\eta_i) = \log\left(\frac{1 + \Psi^+}{1 + \Psi^-}\right),$$

where

$$\Psi^+ = \max_{\forall k, k \in \eta_i} \sum_{\forall (x, y)} P_e(k|(x, y)) * H$$

$$\Psi^- = \max_{\forall k, k \in \eta_i} \sum_{\forall (x, y)} P_n(k|(x, y)) * H$$

In our experiments, phrases with $\mathcal{Q}(\eta_i) > \rho$ (0.5) are retained

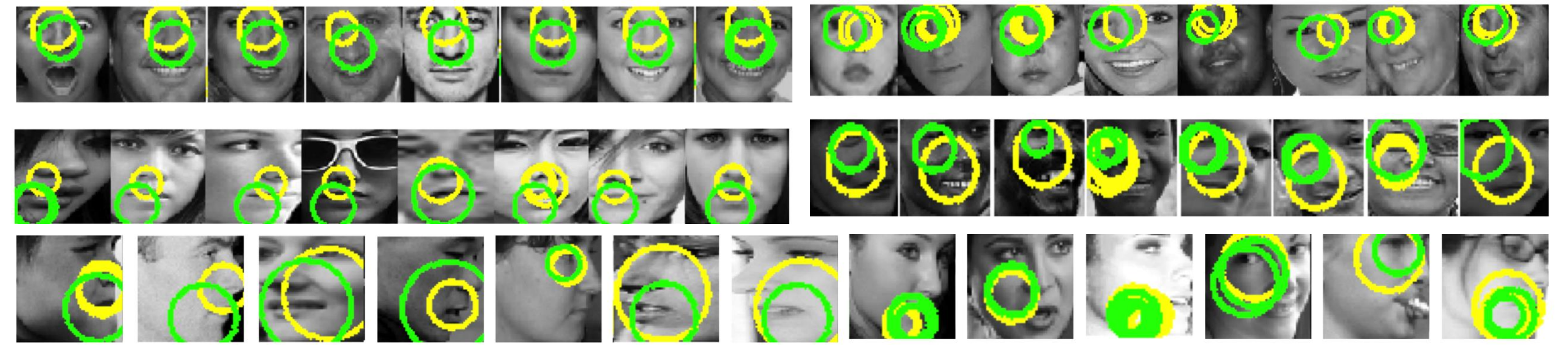
Incorporation into voting framework:

- Visual phrases are incorporated into the voting framework by finding the location and discriminative score similar to words

$$L(\eta_i) = \frac{1}{|\eta_i|} \sum_{w(g) = k} L(g)$$

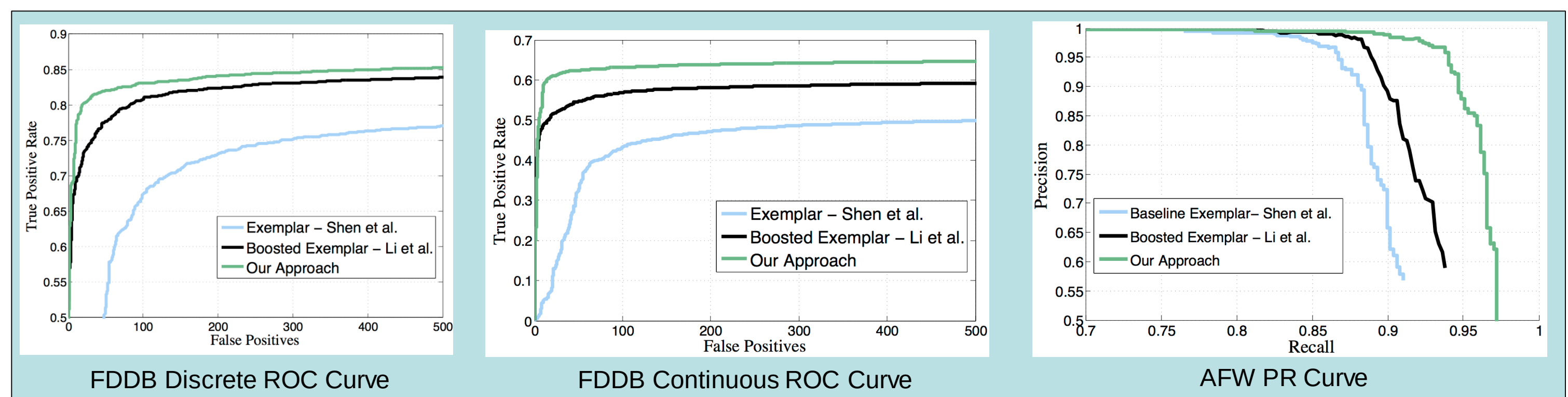
$$\mathcal{D}(\eta_i) = \sum_{\forall k, k \in \eta_i} \mathcal{D}(k)$$

Discovered Visual Phrases:

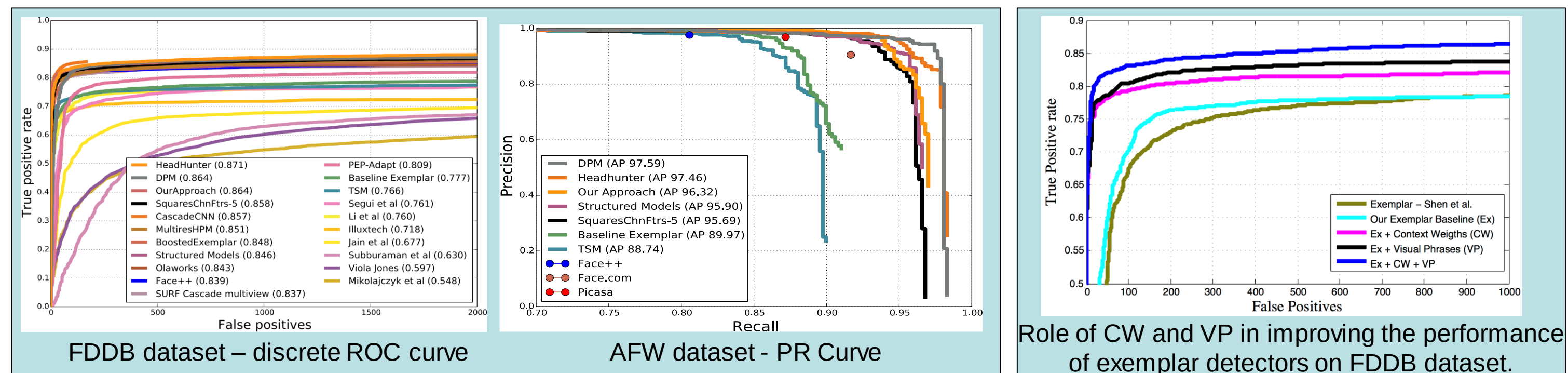


5. Results

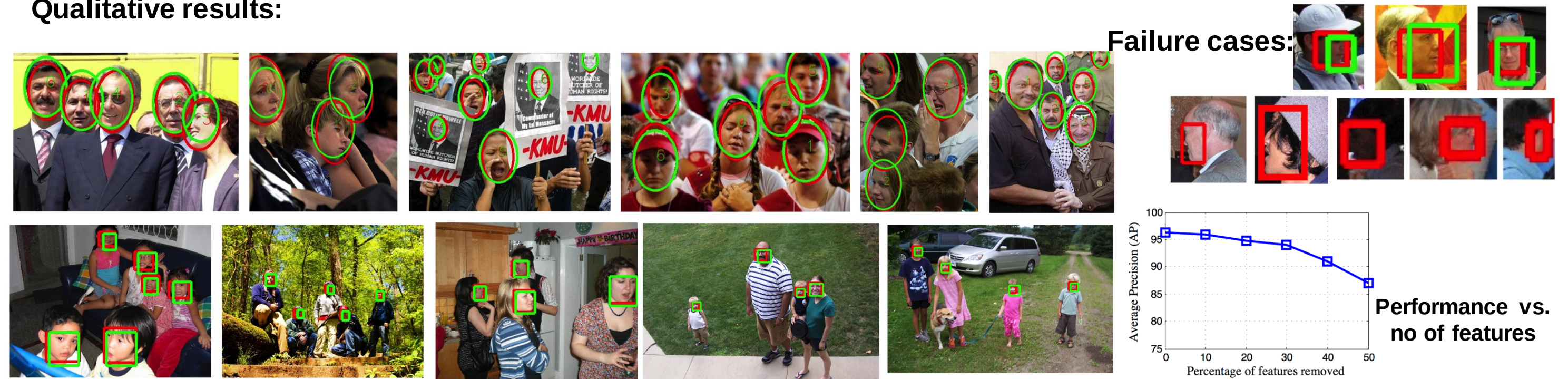
We achieve significant improvement over previous exemplar schemes



Comparison with other approaches:



Qualitative results:



6. References

- X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In *CVPR*, 2012
- X. Shen, Z. Lin, J. Brandt, and Y. Wu. Detecting and aligning faces by image retrieval. In *CVPR*, 2013.
- H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua. Efficient boosted exemplar-based face detection. In *CVPR*, 2014
- M. Shi, X. Sun, D. Tao, and C. Xu. Exploiting visual word co-occurrence for image retrieval. In *MM*, 2012.
- K. Ma and J. Ben-Arie. Vector array based multi-view face detection with compound exemplars. In *CVPR*, 2012
- M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011.

Vijay Kumar is supported by TCS PhD Research Fellowship 2012

Centre for Visual Information Technology, International Institute for Information Technology, Hyderabad, India
<http://cvit.iiit.ac.in>

