Visual Phrases for Exemplar Face Detection

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

Abstract

Recently, exemplar based approaches [13, 22] have been successfully applied for face detection in the wild. Contrary to traditional approaches that model face variations from a large and diverse set of training examples, exemplar-based approaches use a collection of discriminatively trained exemplars for detection. In this paradigm, each exemplar casts a vote using retrieval framework and generalized Hough voting, to locate the faces in the target image. The advantage of this approach is that by having a large database that covers all possible variations, faces in challenging conditions can be detected without having to learn explicit models for different variations.

Current schemes, however, make an assumption of independence between the visual words, ignoring their relations in the process. They also ignore the spatial consistency of the visual words. Consequently, every exemplar word contributes equally during voting regardless of its location. In this paper, we propose a novel approach that incorporates higher order information in the voting process. We discover visual phrases that contain semantically related visual words and exploit them for detection along with the visual words. For spatial consistency, we estimate the spatial distribution of visual words and phrases from the entire database and then weigh their occurrence in exemplars. This ensures that a visual word or a phrase in an exemplar makes a major contribution only if it occurs at its semantic location, thereby suppressing the noise significantly. We perform extensive experiments on standard FDDB, AFW and G-album datasets and show significant improvement over previous exemplar approaches.

1. Introduction

Face detection is one of the classical computer vision problems that finds extensive applications in a variety of commercial systems. Despite years of research, it still remains a challenging and unsolved problem. Though the current algorithms have matured for near-frontal faces, they are yet to achieve a satisfactory performance for unconstrained face images, popularly known as *in the wild* faces.

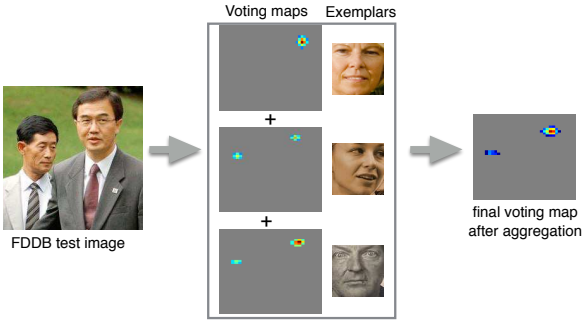


Figure 1. **Ensemble of Exemplars for Face detection:** A large database of diverse exemplars is collected, and indexed using a BoW representation. During testing, each exemplar casts a vote on the test image at multiple scales. The votes from different exemplars are then aggregated to detect the faces.

Most of the popular algorithms for detecting objects and faces are either based on cascaded AdaBoost classifiers [28] or deformable part models (DPM) [5]. The Viola and Jones cascade [28] of discriminatively trained AdaBoost classifiers is extremely efficient and is very effective for frontal faces. DPM based approaches [5, 18, 30, 37], on the other hand, handle intra-class variations by learning the individual parts of an object along with their deformations. Both approaches aim to learn face variations from a large set of training examples seeking good generalization performance. However, it is extremely difficult to capture all possible object variations in a compact model, whether holistic or part based.

Contrary to the above approaches, recently proposed exemplar based detectors [13, 22] do not explicitly model the face variations. Instead, it follows Bag-of-Words (BoW) retrieval technique and Hough voting [6, 12] to detect the faces efficiently. In this paradigm, a large database of exemplars that cover significant face variations are collected. Local features (such as SIFT) are extracted, quantized and indexed using traditional BoW technique. For detection, each exemplar casts a vote on the given target image at multiple scales after which the votes are aggregated (Figure 1). Since each exemplar is *specific* to particular variation, it is possible to detect faces in challenging conditions using a sufficiently large database with diverse exemplars. This ap-

proach which avoids exhaustive sliding-window search is efficient, scalable, easily parallelizable and offers flexibility to add more exemplars without additional training required.

Current exemplar approaches treat each exemplar as a collection of independent visual words that capture facial features from different regions. It is however apparent that many visual features co-occur in faces. For *e.g*., stable visual features that describe eyes and nose occur together with greater probability. Thus, the current exemplar schemes fail to capture such semantic relations among visual features, unlike model-based approaches, which are designed to capture higher order spatial relations. We propose to incorporate such higher order information using "visual phrases" in the exemplar framework. Visual phrase is a group of highly correlated and stable visual words that co-occur in faces frequently. We discover such visual phrases from an exemplar database. As it is computationally expensive to model all possible dependencies for a large vocabulary, we employ a popular association rule mining technique [2] and obtain large candidate visual phrases that occur frequently. We then retain only those phrases that are suited for detection through a discriminative training process.

We also introduce a domain-specific similarity function that considers the spatial consistency of visual features along with their discriminative ability. This is in contrast to non-discriminative inverse document frequency (IDF) based function used in current schemes which ignores the spatial information. Our approach is based on the observation that a stable visual word or a phrase appears at *consistent locations* and in *consistent exemplars*. We leverage the availability of a large database to estimate the spatial distribution of words and phrases and weigh their individual occurrences in each exemplar based on this distribution. This ensures that visual words and phrases in exemplars cast a strong vote only if they occur at their globally consistent locations. This suppresses the contribution of noisy features introduced due to imperfect feature extraction and quantization processes.

The contributions of this paper are as follows:

- We propose an approach to discover and incorporate visual phrases that capture higher order information into the voting framework.
- We introduce spatial consistency of visual words and phrases that weighs their occurrence in exemplars according to their location. This also helps in identifying and removing noisy features in the exemplars which reduces the memory requirements.
- We achieve near state-of-the-art results on the challenging FDDB [9], AFW [37] and G-album [7] datasets, and achieve significant improvements over baseline exemplar [22] and Boosted exemplar [13] approaches, respectively.

2. Related Work

The models proposed for face detection fall into three broad categories: Global discriminative models, Part based models, and Exemplar based models. The first category is the simplest and most efficient of which the Viola-Jones (VJ) face detector [28] is the most popular one. Zhang and Zhang [34] presents a detailed survey of the variants of VJ along with several features. Due of its speed and openly available implementations, it has been extensively used in commercial applications and consumer devices. However, the performance of the vanilla VJ detector degrades significantly for challenging *in the wild* faces. SURF cascade detector [15], and SquaresChnFeatures [18, 32] are currently the best performing VJ variants. These methods use much more richer and informative SURF and integral channel features to achieve superior performance.

Deformable part model (DPM) based techniques [18, 30, 37] which are very effective object detectors, have enjoyed similar success for face detection in the recent years. Mathias *et al.* [18] have shown recently that a properly trained vanilla DPM using a large database can achieve state-of-the-art results on various face benchmarks. Both cascaded detectors and part-based models distill compact models of faces from large training database that captures most common variations in pose, expression, lighting, etc.

Exemplar based techniques, on the other hand, do not learn such global models but instead allow each exemplar to contribute for the task at hand. Exemplar-SVM [17] proposed for object detection learns a linear model for each positive exemplar with large pool of negative examples and evaluates each model during testing. Similarly, per-class exemplar detectors provide object cues in Image Parsing [26]. Exemplar based approaches [13, 16, 22] were applied recently for face detection. Ma *et al.* [16] incorporates ideas from DPM into the exemplar approach, in which parts from different exemplars are combined to obtain an aggregated similarity between an input image and the compound exemplar. The approach offers flexibility to face variations, occlusions and requires minimal training data. The approaches in [13, 22] combine retrieval and Hough voting schemes [6, 12] for detection where each exemplar votes for presence of faces in test image. While a large database is used in [22], a much compact database is selected in [13] through a discriminative boosting framework. Exemplar based approaches have the advantage of being easier to adapt compared to other models, even though the detection performance has been slightly below the state-of-the art. We show that it is possible to improve this using the spatial information of visual words and their dependencies.

In our work, we have started with the original algorithm in [22] and avoided [13] as it involves manual selection of thresholds for domain partitioned classifiers for each exemplar. We make several improvements over [22]. We incor-

porate visual word relations through visual phrases along with spatial weighting of features into the voting framework. Our similarity function is much more discriminative compared to IDF-based scoring used in [13, 22]. As we show in experiments, this approach results in a significant performance improvement over [13, 22].

Some of the works in the area of content-based retrieval have used similar insights. In [3, 10, 24], visual word dependencies in a database with multiple objects and scenes are discovered. While such dependencies are suppressed for retrieval tasks [3, 10, 24], we exploit them as positive cues for detection. In [29], the contextual weighting of the features is proposed but for sparse local features. The work of Yuan *et al.* [33] is closely related to ours. They demonstrate an approach to discover meaningful *visual phrase lexicons* with spatially consistent visual words given a large database. Visual phrases are also applied in image retrieval [35, 36], object recognition [35] and detection [20] tasks.

3. Proposed Approach

3.1. Exemplar Framework for Face Detection

In the exemplar framework [13, 22], local features such as dense-SIFT are extracted from a large exemplar database and a k-means based vocabulary is constructed followed by feature quantization. Term frequencies (TF) and inverse document frequencies (IDF) are calculated and inverted files are created similar to BoW retrieval scheme [25].

During testing, all the exemplars collectively participate in the Hough-based voting [6, 12] process that uses the spatial locations of features to locate the faces in a given image. Each exemplar generates a voting map (at multiple scales), where each location in the map indicates the similarity score between the exemplar and the image sub-region at that location (Figure 1). The similarity measure between an exemplar e_i and the rectangular region centered at location p of the test image x is given as [21, 22]:

$$S(p, e_i) = \sum_k \sum_{\substack{f \in R_x(p), g \in e_i \\ w(f)=w(g)=k \\ ||\mathbb{T}(L(f))-L(g)||<\epsilon}} \frac{F(w(g), L(g))}{tf_{e_i}(k) \cdot tf_x(k)}, \quad (1)$$

where x is the test image, $R_x(p)$ is the sub-image region of x centered at p. f and g are the local features and $L(f)$ and $L(g)$ are their corresponding locations from x and e_i, respectively. $w(f)$ and $w(g)$ are the quantized visual words of features f and g respectively. $w(f) = w(g) = k$ indicates that only the matched visual words are considered for voting. The spatial constraint $||\mathbb{T}(L(f)) - L(g)|| < \epsilon$ further ensures that matched features should be closer under some unknown transformation $\mathbb{T}$. $F(w(g), L(g))$ is the weightage given to each matched feature pair quantized to visual word k. To handle burstiness, weights are divided by

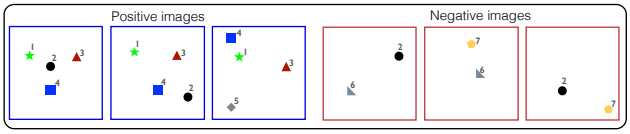


Figure 2. **Motivating reasons:** Consider 3 positive and negative examples. Current schemes which ignore spatial location assign an high IDF weight for words 1, 3 and 4. However, word 4 in third positive example is occurring at inconsistent location possibly due to noise which should be given slightly less weightage. Also, word 2 appears in both positive and negative examples and should be given less weightage. Word 4 in positive example 3 do not occur at its globally consistent location, hence should contribute less.

$tf_{e_i}(k)$ and $tf_x(k)$, which denote the TF of the visual word k in the exemplar and test image, respectively [21].

Suppose that we are interested in detecting faces of size $N_x \times N_x$ in the test image[1]. The location p where the vote is cast is calculated as follows.

$$p = L(f) + \frac{N_x}{N_{e_i}}(C_{e_i} - L(g)), \quad (2)$$

where C_{e_i} and N_{e_i} are the center and size of the exemplar e_i, respectively. The voting maps are then subtracted with an exemplar specific threshold and aggregated to obtain the final voting map [22]:

$$S(x) = \sum_{i:s_i(x)>\rho_i} (s_i(x) - \rho_i), \quad (3)$$

where $s_i(x)$ is the similarity score between x and e_i, and ρ_i is the discriminatively trained threshold for exemplar e_i obtained during training.

3.2. Contextual Weighting of Features

Current exemplar detectors compute the similarity scores between the exemplar and a target image sub-region as [22],

$$F(w(g) = k, L(g)) = idf^2(k), \quad (4)$$

where $idf(k)$ is the IDF of the visual word k. The voting scheme with above similarity score has two issues. First, the use of IDF computed from only the positive exemplars makes it less discriminative for detection tasks. Second, the approach assumes that exemplar words are noise-free and considers all the visual words equally important when computing the similarity score. However, a noisy feature that is wrongly assigned to a visual word with high IDF may significantly affect the voting process.

Figure 2 illustrates these issues with a simple example with 3 positive and 3 negative exemplars. Current exemplar approaches consider only positive exemplars and will give a high IDF to vocabulary elements $1, 3$ and 4. This will also

[1] with an aspect ratio of 1:1 for exemplars and target faces.

Figure 3. **Spatial Context of visual words:** (a) and (b) shows the location of two visual words in different images. Notice how the visual word in (a) is *highly localized* with consistent locations while the word in (b) appears at random locations (left). The global distributions of each visual word over the entire database (middle) is used to weight their occurrences in individual exemplars. Its overlay on the mean exemplar face (right), shows strong localization for stable words. Unstable words occurs at diverse locations and are down weighted.

assign a high IDF to the vocabulary element 2, even though it occurs with similar probability in both positive and negative images. Another issue is that, a highly discriminative word occurring at an incorrect location in an exemplar may cast a wrong vote. In Figure 2, the visual word 4 is discriminative as it occurs in consistent locations in positive exemplars 1 and 2. However, a feature in exemplar 3 is wrongly assigned to visual word 4 due to noise, and if we ignore its location, may contribute incorrectly during voting.

Motivated by these observations, we address the following questions: How can we down-weight less discriminative vocabulary elements? and How can we discover noisy features in exemplars and down-weight their contribution during voting? Our modification is based on the argument that a visual word that is stable and discriminative tends to occur consistently in *similar locations* and in *similar exemplars*. Similarly, a visual word that is noisy or less discriminative with very high probability occurs at random locations. This is illustrated in Figure 3, where a stable visual word that describes the appearance of nose in a particular view (here frontal) appears consistently at the same location in other similar exemplars, or in other words, it is highly *localized*. We estimate the distribution of each visual word from the entire database and use it to weight their occurrence in exemplars. Based on this, visual words appearing at their globally consistent location get more weightage while those appearing at random locations get less weightage.

Let, $w(L(g), e_i)$ denote the visual word corresponding to feature g at location $L(g) = (L^{\mathrm{x}}(g), L^{\mathrm{y}}(g))$ in the exemplar e_i. We estimate the distribution of each vocabulary element k from the entire exemplar database as,

$$P_e(k|(\mathrm{x},\mathrm{y})) = \frac{1}{N_e}\sum_{e_i}\mathcal{I}(w(L(g),e_i) == k), \quad (5)$$

where N_e denote total number of exemplars, (x,y) denote the location and $\mathcal{I}(\cdot)$ is an indicator function whose value is 1 if the condition is satisfied, otherwise 0. In the practical cases, however, there will be some misalignments between the exemplars. To handle the misalignments, we convolve the distribution with a 8×8 Gaussian filter H ($exp(-d/\sigma^2)$ and $\sigma^2 = 2.5$) to obtain the spatial weightage for each vocabulary element as,

$$W((\mathrm{x},\mathrm{y}),k) = P_e(k|(\mathrm{x},\mathrm{y})) * H \quad (6)$$

We show such weightage obtained for a stable and non-stable word in Figure 3. It also suggests to suppress the unstable words which would otherwise affect the voting process. Similarly, we estimate distribution of each vocabulary element on a large corpus of negative images n_i as

$$P_n(k|(\mathrm{x},\mathrm{y})) = \frac{1}{N_n}\sum_{n_i}\mathcal{I}(w(L(g),n_i) == k), \quad (7)$$

where N_n denote the total number of negative images. We then compute the global *discriminative score* $\mathcal{D}$ for each visual word k as,

$$\mathcal{D}(k) = \max_{(\mathrm{x},\mathrm{y})} \frac{P_e(k|(\mathrm{x},\mathrm{y}))}{P_n(k|(\mathrm{x},\mathrm{y}))} \quad (8)$$

The above score $\mathcal{D}(\cdot)$ is discriminative and also considers the spatial location of features, hence is suited for detection. Finally, our scoring function for every matched feature pair between exemplar and target sub-region is given as,

$$F(w(g) = k, L(g)) = W(L(g),k) \cdot \mathcal{D}^2(k), \quad (9)$$

where $W(\cdot)$ denote the context-aware weightage given to exemplar feature and $\mathcal{D}(\cdot)$ is the discriminative score.

4. Visual Phrases for Detection

Due to the independent assumption in previous exemplar approaches, each visual word independently votes for the target image. However, for faces, it is intuitively obvious

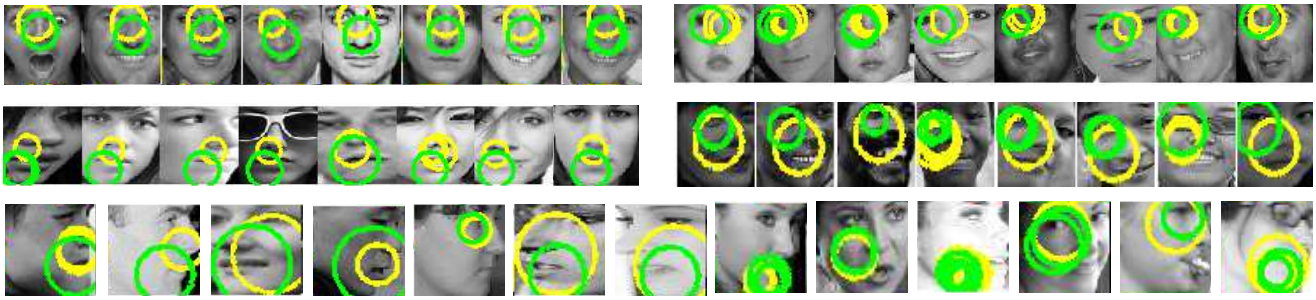


Figure 4. **Visual Phrases for Faces:** The top two rows (left and right) shows 4 different visual phrases that capture relation among two visual words. Notice how the stable visual phrases capture semantic relation among different visual features. Visual phrases are highly localized and appear at similar locations in similar exemplars. Bottom row shows few other visual phrases discovered from the database.

that many visual words are *highly correlated* and *co-occur* together. The current schemes fail to capture such semantic relations among visual features, unlike in model-based approaches which capture much complex relations. Though the terms in the denominator of Eqn 1. handles burstiness, it does not consider the relation among the visual words.

We propose to incorporate higher order information using so called visual phrases in the exemplar framework. A visual phrase is a group of spatially consistent and semantically related visual words that co-occur in faces. We leverage the presence of large database to discover such visual phrases. Given a large vocabulary, it is however, computationally expensive to find all such dependencies. To this end, we resort to a popular data mining technique, *association rule mining* [2] to obtain the candidate visual phrases that occur frequently in the database. We then prune the candidate set and retain only those visual phrases that are well suited for detection.

It is worth to note that, such relations are earlier exploited in computer vision for retrieval tasks [3, 10, 24]. In these tasks, images usually contain multiple objects and scenes and a similarity function with independence assumption tends to over-weight the regions containing highly correlated words [3, 24]. Therefore such correlated words are down-weighted for better retrieval. However, we exploit such relations among visual words for detection as they provide strong cues about the existence of a face region.

We now formally discuss the proposed approach to discover visual phrases. Let $V = \{v_1, v_2, \ldots, v_n\}$ denote the vocabulary and e_i be the exemplar containing subset of vocabulary elements *i.e.* $e_i \subseteq V$. An association rule [2] is an implication of the form $X \Longrightarrow Y$, where X and Y are the itemsets (*visual phrases*) that satisfy $X \subset V$, $Y \subset V$ and $X \cap Y = \varnothing$. The implication rule basically checks with what proportion the itemsets X and Y occur together in an image e_i. The result is a list of all possible combination of words with a support[2] greater than user-specified threshold.

[2]Support is the number of transactions (images) in the database that contain the itemset (phrase) or simply the frequency count of a phrase.

The candidate visual phrase set obtained from the above algorithm on a large database is usually huge containing many redundant phrases. It may also be possible that many of the visual words occur together by chance. Also, the mining technique does not consider the spatial location of words due to which many of the candidate visual phrases are not discriminative for detection task. Due to these reasons, we need to prune the candidate phrases obtained from the rule mining and select only those discriminative phrases that are suitable for detection. We achieve this using the concept of spatial consistency introduced earlier for visual words. We consider the visual phrase as stable and discriminative if all the words associated with it appear in consistent locations in the exemplars, and occur rarely in negative images.

Let, $\Omega = \{\eta_i \mid \forall i, \eta_i \subset V\}$ be the list of candidate visual phrases discovered from association rule mining and $|\eta_i|$ denote the number of words associated with the visual phrase η_i. We assign a score for each candidate visual phrase η_i as,

$$\mathcal{Q}(\eta_i) = log\Big(\frac{1+\Psi^+}{1+\Psi^-}\Big), \tag{10}$$

where

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

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

The terms Ψ^+ and Ψ^- measure the spatial consistency of the words that constitute visual phrase in positive and negative images, respectively. The score $\mathcal{Q}$ in Eqn 10 will be large for those visual phrases that capture the relation of stable visual words, and less for non-discriminative and noisy phrases that occur at random locations. We finally retain the visual phrases whose $\mathcal{Q}$ score exceeds a threshold *i.e.* $\omega = \{\eta_i \mid \mathcal{Q}(\eta_i) > \rho\}$ (see Section 6.1). We show few visual phrases discovered from the exemplar database in Fig 4. Notice how the visual phrases capture the neighbourhood (spatial and scale) relations due to multi-scale

dense feature extraction (*e.g.*, bottom row 5th and 8th image). Once the phrases are discovered, we index their occurrences in exemplars and incorporate them into the voting framework. The spatial location $L(\cdot)$ and discriminative score $\mathcal{D}(\cdot)$ of the selected visual phrases ($\eta_i \in \omega$) are obtained using the mean location of visual words and sum of their individual discriminative scores, respectively.

$$L(\eta_i) = \frac{1}{|\eta_i|} \sum_{\substack{w(g)==k \\ \forall k, k \in \eta_i}} L(g) \tag{11}$$

$$\mathcal{D}(\eta_i) = \sum_{\forall k, k \in \eta_i} \mathcal{D}(k) \tag{12}$$

5. Time and Memory Complexity

When compared with baseline exemplars, the proposed approach requires additional memory for indexing the visual phrases and the contextual weights of visual words and phrases. The average number of visual phrases discovered per exemplar was around 8. In the case of a database of 15k images, this results in an additional memory of 1MB to index the visual phrases and their locations following the representation in [21]. The contextual weights are quantized and stored using 1 byte integer which requires additional 10MB of memory for 80×80 exemplar with 700 visual words and phrases on an average. One could reduce the memory footprint by removing those visual features and phrases with very low contextual weights as they make limited contribution in the voting process. As we demonstrate later, it is possible to remove upto 30% of features with a slight drop in performance. Compared to previous approaches, the only additional time required is to find the dependencies in the target image. Since the target image is indexed and TFs are computed already for voting process, dependencies can be found much faster. This usually takes less than 2 seconds in our unoptimized MATLAB code. Also, similar to [21], we can achieve a further speedup by ignoring the words with high TFs, as their contributions are limited according to Eqn 1. Our MATLAB implementation[3] of the entire detection pipeline without tiling [13] usually takes $10 - 12$ secs for 1280×1280 image most of which is spent in feature extraction and quantization. As reported in [22], it is possible to achieve near real-time efficiency with custom implementation of these steps. Our future work will be focused to develop a C/C++ implementation to achieve near real-time performance.

6. Experiments and Results

We implemented the exemplar detector [22] upon which our improvements are made. The performance of our baseline exemplar closely matches with [22] as shown in Fig 11.

[3]Code is available at http://cvit.iiit.ac.in/projects/exemplar/

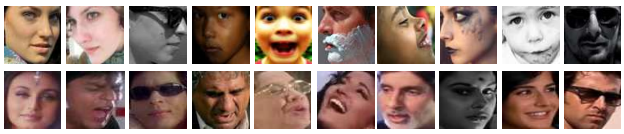


Figure 5. Few images from our database built from AFLW (top) and IMFDB (bottom).

6.1. Implementation details

Exemplars: We collected the exemplar images from AFLW [11] and IMFDB [23] databases. AFLW contains around 25k images and IMFDB contains 34512 images. We randomly sample 10k images from AFLW and 5k images from IMFDB to create our 15k exemplar database. All the exemplars are resized to a fixed size of 80×80. Few exemplars from our database are shown in Fig 5.

Dense Features and Vocabulary: We densely extract patches of size 24×24 with a stride of 3 pixels and 128D root-SIFT representation is computed. We extract the features and their locations at 12 scales by resizing the original image with a scaling factor of $\sqrt{2}$. We construct the 50k-vocabulary using fast approximate nearest neighbour (ANN) k-means [19]. We used the publicly available software VLFEAT [27] for both these tasks.

Visual Phrases: We use the publicly available apriori software [1] to obtain the initial candidate phrases with a minimum support of 100. This resulted in 5837 candidate visual phrases containing 4880 2-visual word phrases, 736 3-visual word phrases and 221 4-visual word phrases. We used 50k 80×80 negative patches [4] with a threshold of $\rho = 0.5$ for discriminative training which finally resulted in 1282 2-visual phrases. Few visual phrases are shown in Fig 4. We noticed that 3 and 4- word phrases were noisy and inconsistent ($\mathcal{Q}(\cdot) < \rho$) and hence are not considered.

Voting and Thresholds: We considered a voting map of size 64×64 similar to [13, 22] and obtained the corresponding grid size using smallest image dimension. To avoid quantization errors, maps are smoothened using a 5×5 Gaussian filter $exp(-d/\sigma^2)$ and $\sigma^2 = 2.5$. The gating threshold for each exemplar is obtained by selecting the maximum score on 1000 negatives images [4] when voted using the same exemplar [22].

Detection: For better performance, test images are upscaled to have a size of atleast 1280 [22]. For memory efficiency, we follow tile-based detection and divide the upscaled image into tiles of size 640×640 with an overlapping stride of 140 pixels [13]. The detection operation on each tile is performed at 3 different scales $(1, 0.5, 0.3)$. At each scale, faces of 15 sizes with a base size of 80×80 and scaling factor of $2^{1/4}$ are detected. We vote only using top 3000 similar exemplars retrieved using BoW model to speed up the processing. A standard greedy non-maxima suppression (NMS) with an overlap threshold of 0.25 is applied to sup-

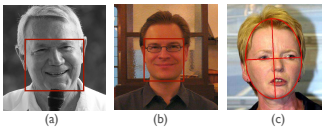


Figure 6. **Annotation mismatch:** Notice the difference annotation strategies across (a) AFLW [11] (b) AFW [37] and (c) FDDB [9].

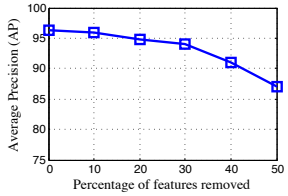


Figure 10. Merits of contextual weighting. It is possible to remove upto 30% of features without a significant performance change.

Figure 13. **Failure cases.** Due to bounding box misalignments (a) detected faces in green are considered false positives. (b) and (c) lack informative features due to extreme pose and low resolution.

press overlapping detections.

Bounding box adjustments: Different face detection benchmarks have followed different annotation strategies (see Fig 6). As in previous works [13, 18, 32] we also modify the detected face regions to better match the location and scale of the ground truth bounding boxes. For example, for the FDDB dataset [9], we convert a detected bounding box of size (w, h) to vertical ellipse with parameters $(\frac{0.9w}{\sqrt{2}}, \frac{h}{\sqrt{2}})$.

6.2. Datasets

We show our results on popular face detection benchmarks - FDDB [9] and AFW [37] and G-album [7]. All these datasets offer challenging scenarios for face detection. FDDB [9] contains a total of 5171 faces in 2845 images collected mainly from Yahoo news website. The dataset contains very low resolution images (smaller than 30×30) that truly tests the capability of algorithms. We use the ROC evaluation software that comes with FDDB database as recommended by the dataset creators and commonly followed by researchers. AFW [37] contains 468 faces present in 205 images. The database is characterized by cluttered background with pose, aging, and occlusion variations. G-album [7] dataset contains 589 family photos with 931 faces. We compare our results on AFW and G-album datasets in terms of precision-recall (PR).

6.3. Results

We compare the performance of the proposed approach with the previous exemplar schemes [13, 22] in Fig 8. We consistently outperform previous schemes on both FDDB and AFW datasets. Following FDDB protocol, we compare our results with all the previously published results in Fig 9. From the discrete curve in Fig 9(a), it is clear that our proposed approach, not only improves over exemplar schemes but also outperforms most of the previous non-exemplar schemes [8, 14, 15, 31, 37], except [18]. The contributions of contextual weighting and visual phrases to the performance improvement is given in Fig 11. While context helps to suppress noisy inconsistent features, visual phrases complement it with its ability to upweight the co-occurrence of visual words in faces. Thus a combination of the two approaches indeed helps as can be observed from Fig 11. We also show the continuous curve in Fig 9(b) which measures the bounding box overlap with ground truth. Unlike [18, 30] which fits oriented bounding boxes, we fit a vertical ellipse which results in a slightly lower score.

For AFW, we used the evaluation software [18] to compensate for bounding box misalignments. Fig 12(a) shows the comparison of our approach with several academic (TSM [37], DPM, HeadHunter and SquaresChnFtrs [18] and Structured models [31]) and commercial solutions (face.com, Face++, Google Picasa). Our approach achieves very high performance reducing the gap between DPM and exemplar based approaches. The common reasons for failure are bounding box misalignment, extreme pose and low resolution. For images with extreme poses and low resolutions, lack of informative features around discriminative regions such as eye and nose causes exemplars not to match unlike holistic matching methods (see Fig. 13). Finally, we show the performance of our approach on G-album dataset. For this dataset, we compare with baseline exemplar [22] and DPM [18] using their trained model. Our approach not only improves upon exemplar method but matches the performance of DPM on this dataset as shown in Fig 12(b). As discussed earlier, it is possible to save memory by removing less consistent features using contextual weights. As shown in Fig 10 for the AFW dataset, it is possible to remove up to 30% of features without a significant drop in performance.

7. Conclusion

In this paper, we introduce visual phrases to capture the semantic relations among the visual words and propose a method to incorporate them into exemplar framework. We estimate the distribution of visual words and phrases from the database and then weigh their occurrences in exemplars based on their spatial consistency. Our domain-specific similarity score considers both spatial consistency and discriminative ability of visual words and phrases, and hence is suited for detection tasks. Finally, we show that incorporating visual phrases and contextual weights can significantly improve the performance of exemplar detectors on various face detection benchmarks.

Acknowledgement

This work is partly supported by the MCIT, New Delhi. Vijay Kumar is supported by TCS PhD research fellowship.

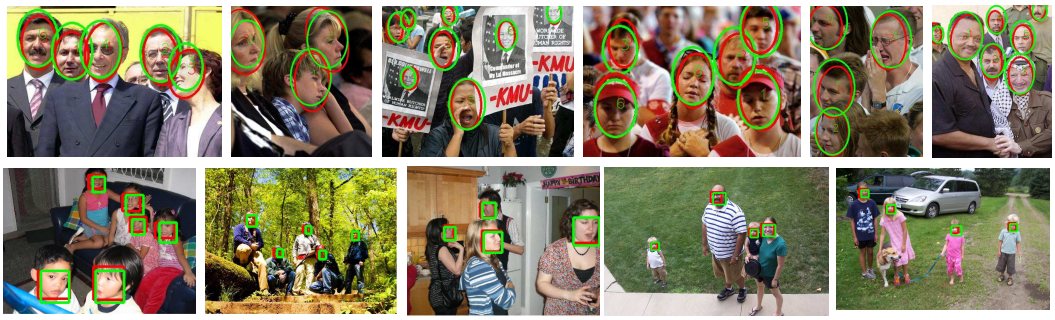
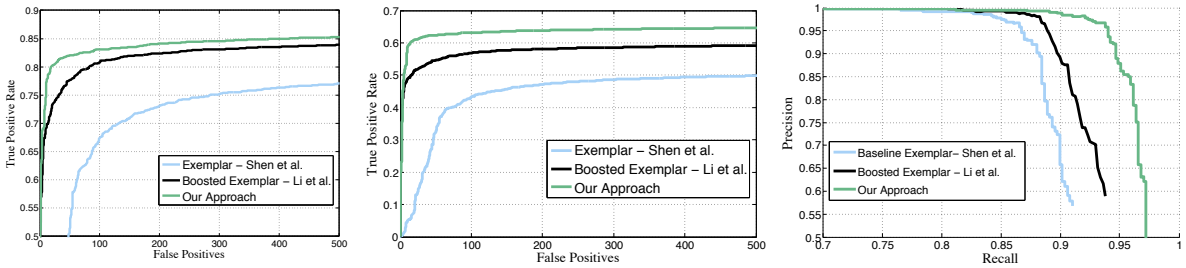
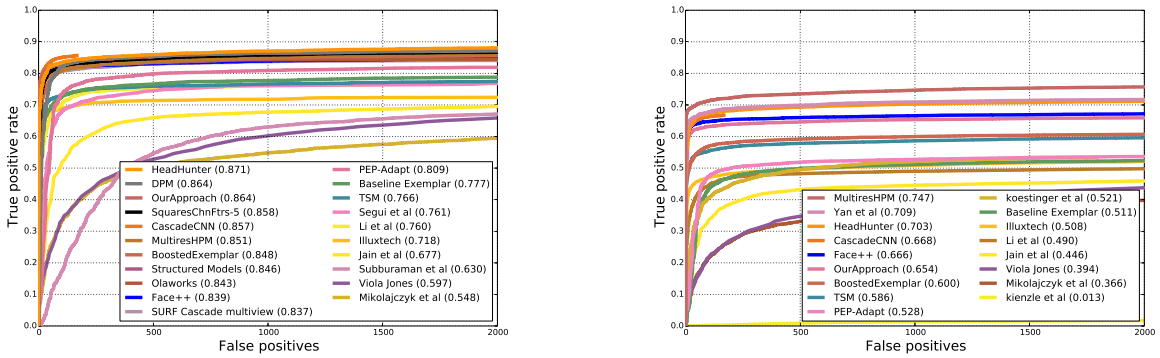


Figure 7. Qualitative results of our detector over FDDB (top), AFW (bottom - first three) and G-album (bottom - last two).



(a) FDDB Discrete ROC curve (b) FDDB Continuous ROC curve (c) AFW PR curve

Figure 8. Comparison with previous exemplar schemes. We outperform the baseline Exemplar [22] and Boosted Exemplar [13] on both FDDB ((a) and (b)) and AFW (c) datasets by a large margin.



(a) Discrete ROC curve (b) Continuous ROC curve

Figure 9. Comparison with other approaches on FDDB dataset. We achieve an average precision of 86.4% with a negligible difference compared to HeadHunter [18]. Our performance improves over the baseline exemplar approach [22] by almost 8%.

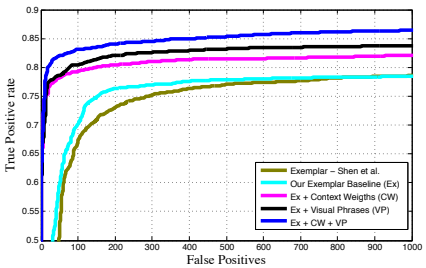
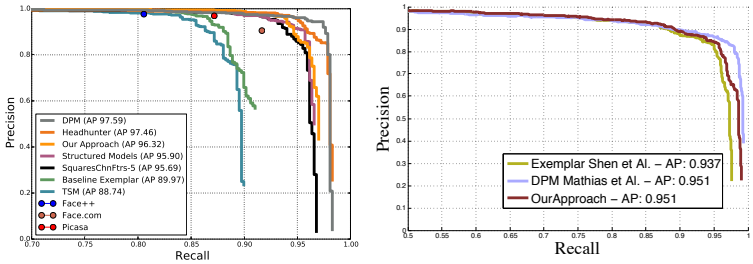


Figure 11. Role of contextual weights (CW) and visual phrases (VP) in improving the performance of exemplar detectors on FDDB dataset.



(a) AFW (b) G-album

Figure 12. Performance comparisons on AFW and G-album datasets. While our approach achieves superior performance on AFW compared to many academic and commercial approaches closely matching HeadHunter [18], the performance matches DPM [18] on G-album dataset.

References

[1] Apriori. http://www.borgelt.net/apriori.html.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, 1994.

[3] O. Chum and J. Matas. Unsupervised discovery of co-occurrence in sparse high dimensional data. In *CVPR*, 2010.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010.

[6] J. Gall, A. Yao, N. Razavi, L. J. V. Gool, and V. S. Lempitsky. Hough forests for object detection, tracking, and action recognition. *PAMI*, 2011.

[7] A. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *CVPR*, 2008.

[8] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Detecting and localizing occluded faces. *arXiv preprint arXiv:1506.08347*, 2015.

[9] V. Jain and E. Learned-Miller. FDDB: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[10] H. Jégou and O. Chum. Negative evidences and cooccurences in image retrieval: The benefit of PCA and whitening. In *ECCV*, 2012.

[11] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[12] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCV workshop on statistical learning in computer vision*, 2004.

[13] H. Li, Z. Lin, J. Brandt, X. Shen, and G. Hua. Efficient boosted exemplar-based face detection. In *CVPR*, 2014.

[14] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *CVPR*, 2015.

[15] J. Li and Y. Zhang. Learning SURF cascade for fast and accurate object detection. In *CVPR*, 2013.

[16] K. Ma and J. Ben-Arie. Vector array based multi-view face detection with compound exemplars. In *CVPR*, 2012.

[17] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[18] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, 2014.

[19] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISSAPP*, 2009.

[20] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011.

[21] 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.

[22] X. Shen, Z. Lin, J. Brandt, and Y. Wu. Detecting and aligning faces by image retrieval. In *CVPR*, 2013.

[23] S. Shetty *et al.* Indian Movie Face Database: A Benchmark for Face Recognition under wide variations. In *NCVPRIPG*, 2013.

[24] M. Shi, X. Sun, D. Tao, and C. Xu. Exploiting visual word co-occurrence for image retrieval. In *MM*, 2012.

[25] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[26] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013.

[27] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.

[28] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[29] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han. Contextual weighting for vocabulary tree based image retrieval. In *ICCV*, 2011.

[30] J. Yan, Z. Lei, L. Wen, and S. Z. Li. The fastest deformable part model for object detection. In *CVPR*, 2014.

[31] J. Yan, X. Zhang, Z. Lei, and S. Z. Li. Face detection by structural models. *Image Vision Comput.*, 2014.

[32] B. Yang, J. Yan, Z. Lei, and S. Z. Li. Aggregate channel features for multi-view face detection. In *IJCB*, 2014.

[33] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *CVPR*, 2007.

[34] C. Zhang and Z. Zhang. A survey of recent advances in face detection. Technical Report MSR-TR-2010-66, Microsoft Research, 2010.

[35] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li. Descriptive visual words and visual phrases for image applications. In *MM*, 2009.

[36] Y. Zhang, Z. Jia, and T. Chen. Image retrieval with geometry-preserving visual phrases. In *CVPR*, 2011.

[37] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.