Vector Array based Multi-view Face Detection with Compound Exemplars

Kai Ma and Jezekiel Ben-Arie
University of Illinois at Chicago
Chicago, IL 60607
kma5,benarie@uic.edu

Abstract

We address the problem of Multiple View Face Detection (MVFD) in unconstrained environments. In order to achieve generalized face detection we use part-based image representations by tessellation of small image patches, which are typified by 2D vector arrays. Faces are detected by a method named Vector Array Recognition by Indexing and Sequencing (VARIS). VARIS is designed to find the optimal similarity matching between the input image and stored exemplars while allowing wide geometrical variations that are limited only by topological constraints. Aggregated similarity is further enhanced by matching the input images with compound exemplars. The novel compounding procedure also reduces the number of exemplars necessary for each class representation. VARIS with compounding performs efficient parallel classification and has polynomial computational complexity.

1. Introduction

Face detection is a classical computer vision problem that has been one of the most widely researched topics. Due to the success of modern appearance-based machine learning approaches, the field of face detection has made significant progress both in speed and accuracy. However, recent technical surveys [13, 26] have revealed that the performance of state-of-the-art face detectors is dropping vastly in a completely unconstrained environment. They also point out that the major obstacle is the problem caused by the pose variation and partial occlusion.

During the last decade, a number of studies have addressed the multi-view object detection problem, such as human faces [24, 25] and pedestrians [8, 9]. These methods collect a large image data set and label images as positives and background negatives, which are then used to train sophisticated classifiers that can optimally separate the two classes. Although these generalized approaches usually work well, there is still a big gap between their performance and ground truth.

In this paper, we concentrate on the Multi-View Face Detection (MVFD) problem and try to fill the performance gap by introducing a new approach that combines two popular frameworks, the exemplar-based method and the part-based approach. Our system, Vector Array Recognition by Indexing and Sequencing (VARIS), views the images as 2D vector arrays (tensors of order 3). VARIS is a multi-dimensional extension of 1D algorithm called RISq (Recognition by Indexing and Sequencing) [3] that has been developed for applications such as speech [10] or human activity recognition, which can be represented by 1D vector arrays.

VARIS is designed to fulfill two major, sometimes conflicting, requirements of object detection: *Generalization* and *Reliability*. Generalization requires recognizing a very large number of different appearances. For this reason, the desired system should have a flexible representation that can tolerate large geometrical variations. We collect a set of face exemplars and represent each exemplar by a tessellation of small image patches. Geometrical flexibility is achieved by the ability of VARIS to dynamically modify the location of patches. Reliability is maintained by the indexing and sequencing steps in VARIS. The indexing step retrieves the stored exemplar patches most similar to each input patch. The sequencing step finds the optimal matching of the indexed patches to the input while preserving the mutual topology.

The basic idea of VARIS is to maximize the similarity between the input and a face exemplar which is composed of the optimal combination of classified feature patches. The optimization is constrained by the requirements of preserving the topology of both input and exemplars. In this work, we define a non-parametric part representation, which enables a dynamical assembly of local exemplar features. Moreover, parts from different exemplars of the same class can be optimally selected and assembled into a new compound exemplar that has the highest similarity with the input. The general flow chart of VARIS with compounding is shown in Figure 1. Based on a much smaller training set, our system achieves better results than the current state-of-the-art object detection systems [9, 19] with a reasonable

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

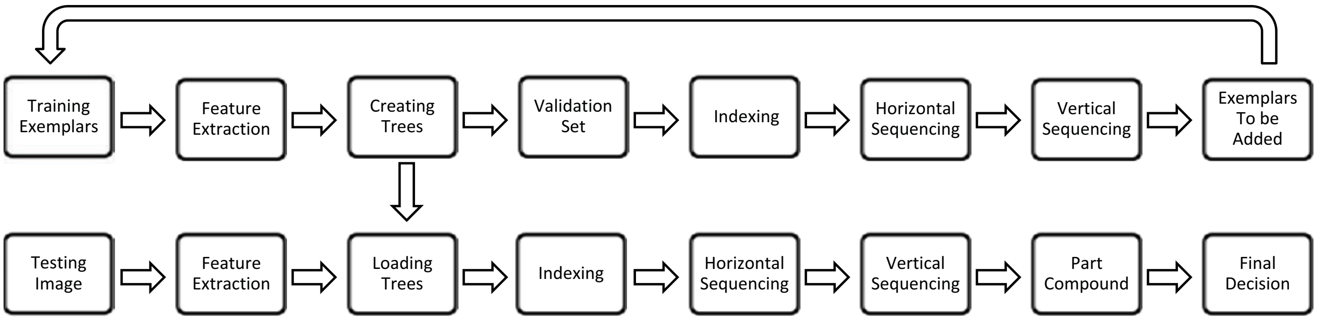

Figure 1. Flow chart of VARIS with compounding. Top: Validation stage. New face exemplars are iteratively added to the initial training dataset until no further improvement is achieved; Bottom: Testing stage. Similar exemplars are selected by applying VARIS on the final dataset and new compound exemplars with higher similarity are assembled with parts from different exemplars.

computational cost. In contrast to other methods, VARIS with compounding is also less sensitive to geometrical distortion, noise and partial occlusion.

The rest of the paper is structured as follows. Section 2 summarizes relevant previous work on multi-view object class detection. Section 3 and 4 describe details of VARIS and exemplar compounding. Section 5 displays the experimental results as well as the comparisons with other approaches. Section 6 concludes and discusses future work.

2. Related Work

A major obstacle of MVFD is the significant variations in the appearance of different face views. Hence, it is very difficult to train a single detector for all face poses. A straightforward solution is to use multiple detectors, each for a pre-defined face direction [24, 25]. The multiple-detector approach requires a significant amount of training data. Both collecting and labeling the images become expensive with the increasing number of poses. Furthermore, the learned system is difficult to extend. A handful of changes to the training data set, *e.g*. adding or deleting few examples, may require retraining, which is also time consuming. Moreover, the improvement of the retrained system is not guaranteed since the modification is sometimes so small compared to the full data set that it can barely affect the learned parameters.

Exemplar-based approaches [7, 18] are recently getting more attention in the computer vision area since they can overcome the problems mentioned above. Instead of learning a sophisticated classifier, these approaches concern similarities between an input image and previously saved instances, *i.e.* the exemplars. For example, Frome *et al*. [12] introduced a framework that learns a distance function for each training exemplar as a combination of weighted distances between small similar patches. Although this work falls into the object recognition domain, the results show that the method can be used to detect object classes. The major problem, however, is that the structural relationships of the ranked patches are ignored. Later work [5, 9] have demonstrated that the geometric information plays an important role in the patch based approaches as well as the appearance.

Another recent and interesting work was proposed by Malisiewicz *et al*. [19]. Their object detection approach is built on a large collection of linear Support Vector Machine (SVM) classifiers and each of them is trained with a single positive exemplar and millions of negatives. They claim that a better method to represent an object class is to use a non-parametric approach to express the positives and a parametric one for the negatives. However, our work is based on a different hypothesis that object class instances with some constraints might be sufficient to characterize their class. This is a widespread idea in human cognitive science as well [1]. Moreover, [19] collects a large number of exemplars as positive instances in order to account for the intra-class variation. In comparison, we use a much smaller number of positive instances because our compounding approach can collect parts from different exemplars of the same class and assemble them into an optimized compound exemplar. Theoretically the number of compound exemplars assembled during the testing stage is much larger than the number of original exemplars.

Compared to holistic approaches, part-based approaches have many advantages in terms of face detection: Parts of the face are less sensitive than the whole face to the difference in appearance; Parts also have better tolerance to lighting variations and partial occlusion. Another inspiring work was proposed by Thomas *et al*. in [23]. Their system consists of a strong part-based object detector, referred to as Implicit Shape Model, and a pose recognition system that enables the same part vote across multiple poses. Similarly, our approach allows features from multiple classes to be selected simultaneously. Unlike [15], however, we use densely sampled features to represent each part instead of building a codebook by clustering sparse interest points.

3. Indexing

VARIS inherits the characteristics of part-based approaches. It segments the face image into a tessellation of small image patches, where each patch is typified by a multi-dimensional vector. Indexing this vector retrieves similar patches of previously stored exemplars. Moreover, an object part is defined as an array of low level image patches and this part definition has more flexibility than other approaches in achieving global similarities.

3.1. Feature Extraction

Inspired by the substantial properties of local rotation-invariant image features [16], many modern part-based approaches apply the interest point detector as the way to discover major object parts. However, the interest point detector always prefers the salient features such as eyes, nose and mouth while ignoring other less informative ones like weak edges. Therefore, if one or few parts are not detected due to blur or occlusion, the system performance drops drastically. On the other hand, holistic detection system proposed by Dalal and Triggs [8] shows superior results with densely sampled features. Later Felzenszwalb *et al.* proposed a part-based approach [9] with similar representation and achieved the current state-of-the-art performance in general object detection. In this aspect, we apply the same strategy of densely sampling each training image into a normalized Histogram of Oriented Gradient (HOG) grid map. During the testing stage, features on the input image are indexed to find similar features from pre-stored exemplars. To build a better correspondence between the input image and exemplars, we set up a middle layer that dynamically assembles features into a part representation. During the training stage, the face part is defined as a vector array $\mathbb{A}$ consisting of all features on the same row of the same exemplar.

3.2. Feature Indexing

We apply a kd-tree based technique to index the input feature vectors. Initially proposed by Bentley [4], the kd-tree has been widely adopted to build a balanced binary tree data structure for nearest neighbor indexing due to the efficient and accurate search ability in low dimensions. However, the efficiency vanishes when the data dimensions increase because the backtrack in high dimensions becomes computationally expensive. Therefore, we implement the priority kd-tree [2] for the feature indexing, which keeps a high probability of finding the true nearest neighbors while the backtracking is limited. There are many other nearest-neighbor search algorithms that have similar indexing functionality, and a thorough study can be found in [14].

Tree construction: Our initial training data set contains face exemplars in different poses. We manually separate the exemplars into subclasses according to their pose, and keep 10 to 15 exemplars for each subclass. HOG features in the exemplars are defined as $\mathbf{v}_{p,q}^{m,n}$, where p and q are the row and column indices of that feature, and m and n denote the subclass and exemplar indices. We have previously defined the part $\mathbb{A}$ as a one-dimensional array of feature vectors:

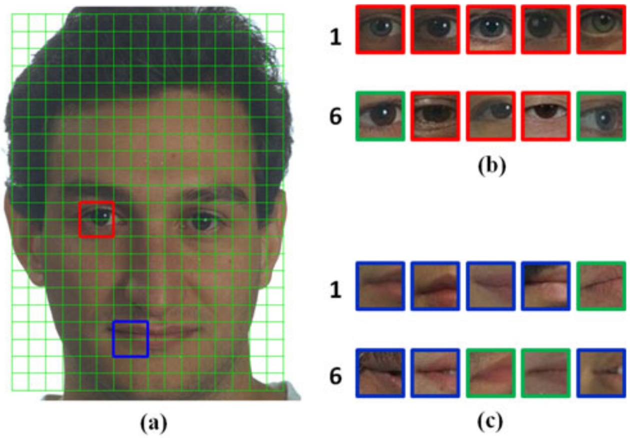

Figure 2. Examples of feature indexing. (a) A test image with overlapping HOG feature grid. The eye in red square and the mouth in blue square are two indexed patches; (b) 10 nearest neighbors of the indexed eye. The eyes in green squares (6 and 10) are from non-frontal face classes; (c) 10 nearest neighbors of the indexed mouth. The patches in green squares (5, 8 and 9) are from non-frontal face classes.

$$\mathbb{A}^{m,n}(p) = \{\mathbf{v}_{p,q}^{m,n} | q = 1, 2, \cdots, Q\} \tag{1}$$

We construct the kd-tree $T(p)$ with $\mathbb{A}(p)$ from all exemplars, and in total we get P kd-trees, where P is the number of total row arrays in one exemplar. Since exemplars are randomly selected under different conditions, the parts are not well aligned. To better cope with this spatial variation, each kd-tree also includes adjacent row arrays. Then we redefine $T(p)$ as:

$$\begin{aligned} T(p) = \{\mathbb{A}^{m,n}(b) | m = 1, \cdots, M; n = 1, \cdots, N; \\ b = p - l, \cdots, p, \cdots, p + l\} \end{aligned} \tag{2}$$

where $p > l > 0$ and $p+l < P$. l defines how much vertical shift that the part can tolerate.

Tree indexing: For each training exemplar, we create a matrix to preserve the indexing information. Given an input image I, we extract the HOG features first and then do a k nearest-neighbor search for each $\mathbf{x}_{i,j}$ within $T(i)$. The search returns with k nearest neighbors according to the Euclidean distance. If vector $\mathbf{v}$ is selected as $\mathbf{x}_{i,j}$'s neighbor, we add one entry to the corresponding element of the matrix with the indexing information: the row and column indexes, i and j, and the distance value. To abide by the principle of maximum aggregated similarity, we convert each distance value to a similarity score with the Gaussian function:

$$S(\mathbf{x},\mathbf{v}) = exp(-\frac{1}{2\sigma^2}d(\mathbf{x},\mathbf{v})) \tag{3}$$

where $d(\mathbf{x},\mathbf{v}) = ||\mathbf{x}-\mathbf{v}||^2$ and σ is estimated in the validation step. We would like to comment here that aggregating similarity scores is much more effective than minimizing accumulated distances [10].

Our results show that HOG features are highly discriminative and most of the salient ones have neighbors falling into the same class as the input (see Figure 2). For the negative images, neighbors of the indexed feature vectors are more uniformly distributed over all subclasses. After the indexing stage, each element of the matrix may have zero, one or multiple entries depending on the similarities of the corresponding feature vectors. To further utilize the similarity information, a 2D sequencing step is applied to select the optimal feature set $\mathbb{W}$ from the non-empty elements. More details about the selection will be revealed in section 4.

3.3. Feature Importance Weighting

Since our approach only checks the similarities between the input image and positive exemplars, the max-margin framework [17, 19] cannot be adopted here to learn the feature weights. In addition, VARIS dynamically selects the optimal feature subset during the testing stage, therefore a traditional learning framework for the part-based approach [6, 11] is not feasible either. To emphasize discriminative features in an exemplar, we simply use the power of the gradient magnitude to represent the feature importance. Because our training data set has a limited number of exemplars, we can manually remove high power features that are outside the face, to make the weighting even more precise. During the experiments, we observe that the false positive cases emerge when many non-salient features are indexed with high-similarity neighbors, which also contribute enough similarity scores to the final summation. To alleviate this effect, we set a threshold t_w that turns the similarity scores of those low power features to negative scores. As a result, equation 3 is modified to:

$$S(\mathbf{x}_{i,j},\mathbf{v}^{m,n}_{p,q}) = \begin{cases} G_{i,j} \times G^{m,n}_{p,q} \times exp(-\frac{1}{2\sigma^2}d(\mathbf{x}_{i,j},\mathbf{v}^{m,n}_{p,q})) \\ \qquad \text{if } G^{m,n}_{p,q} \geq t_w \\ \\ -G_{i,j} \times G^{m,n}_{p,q} \times d(\mathbf{x}_{i,j},\mathbf{v}^{m,n}_{p,q}) \\ \qquad \text{if } G^{m,n}_{p,q} < t_w \end{cases} \tag{4}$$

where $G^{m,n}_{p,q}$ is the power of gradient magnitude for each $\mathbf{v}^{m,n}_{p,q}$ and $G_{i,j}$ is the corresponding power for each $x_{i,j}$. Therefore, the salient features contribute positive scores to the total similarity while the non-salient features, on the other hand, contribute negative scores or 0 in the best case. This negative weighting mechanism significantly reduces the false alarm rate of the detection.

4. Sequencing

In the previous step, each input vector $\mathbf{x}$ finds k neighbors by being indexed within a kd-tree. The mapping information is recorded in the data matrices. Since one input feature vector $\mathbf{x}$ could match multiple feature vectors $\mathbf{v}$ in the same exemplar and only one such matching is allowed, we want to find an optimal subset $\mathbb{W} \subseteq \mathbb{V}$ that maximizes the similarity with the input features $\mathbb{X}$ for each exemplar's feature set $\mathbb{V}$.

$$\hat{\mathbb{W}} = \arg\max_{\mathbb{W}} \sum_{\mathbf{v}\in\mathbb{W}} S(\mathbf{x},\mathbf{v}) \tag{5}$$

where $S(\bullet)$ is the similarity function in equation 4. The optimization is constrained by the requirements of preserving the topology of both input image and exemplars.

An unsophisticated way of finding the optimal set is to examine all the combinations and choose the one with the highest value, but this is impractical due to the large computational cost. In this paper, we implement a 2D sequencing approach to search for the optimal set $\mathbb{W}$. Our sequencing approach is based on dynamic programming, which is a much more efficient method for discrete sequence optimization. Since we may have negative similarity scores from the previous step, the traditional dynamic programming that maximizes the total score is no longer valid. We make a tradeoff between the similarity and the topology, where the system takes the negative scores into account and maximizes the feature similarity as well as the topology:

$$\hat{\mathbb{W}} = \arg\max_{\mathbb{W}} \sum_{\mathbf{v}\in\mathbb{W}} S(\mathbf{x},\mathbf{v}) + \sum_{\mathbf{v}\notin\mathbb{W}} L(\mathbf{v}) \tag{6}$$

where $L(\mathbf{v}) = \alpha \times G_v$ is a loss function that penalizes the final similarity score if $\mathbf{v}$ is not selected in the sequence.

The approach is 2D because the same sequencing search runs recursively in horizontal and vertical directions. We first implement the horizontal search for an optimal sequence of features that best represent object rows (parts). Since $\mathbf{x}_{i,j}$ is indexed not only to $\mathbf{v}$ in the same row i but also to $\mathbf{v}$ in adjacent rows, each input row is compared to multiple rows in an exemplar. Then we implement the vertical search to select the optimal sequence of the rows. Because we treat the feature selection process in a sequential manner, the one-dimensional sequencing approach has the following spatial restrictions:

- A strict one-to-one matching is applied. In particular, if an input vector $\mathbf{x}$ has multiple matched $\mathbf{v}$ in the same row and one $\mathbf{v}$ has been selected to be in the sequence, other matched $\mathbf{v}$ cannot be included in the sequence.

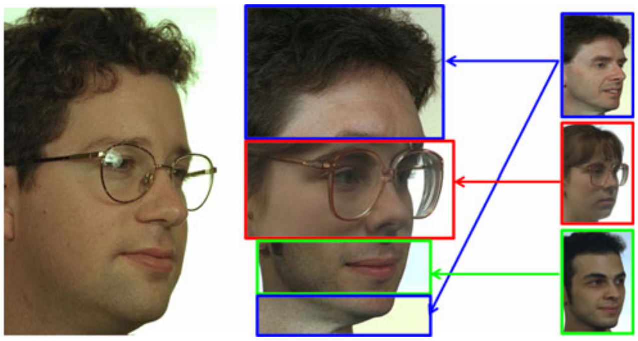

Figure 3. Example of a compound exemplar. Left: test image; Middle: the compound exemplar that is assembled with parts from different exemplars of the same class. It has a higher similarity with the test image than each of the exemplars on the right; Right: exemplar candidates that provide the compounding parts.

- Any two matchings cannot lie across each other. For example, if an input vector $\mathbf{x}_i$ is selected to match $\mathbf{v}_j$, then next $\mathbf{x}_{i+1}$ can only match to $\mathbf{v}_{j+t}$, where $t \geq 1$.

The advantage of the 2D sequencing approach is that the geometrical flexibility is achieved by dynamically assembling features in the testing phase. Those deformable templates have large tolerance to spatial variations and occlusions.

After the 2D sequencing step, we implement the part compounding procedure. Parts from different exemplars that belong to the same subclass can be merged together to generate a compound exemplar with a higher similarity than any other original exemplars (see Figure 3). This idea is inspired by the observation during our experiment that many face exemplars are partially similar to the input image but none of them has a distinguishing score that can be positively recognized. Although the compounding approach increases the final scores of the true positives, the scores of negative inputs are also raised. Therefore, we set a restriction that exemplars could offer parts to be compounded only when their similarity scores pass a pre-defined threshold t_c, where t_c is usually set to a value two times smaller than the final detection threshold. The diagram in Figure 4 illustrates the 2D sequencing and the compounding procedures.

5. Evaluation and Results

In this paper, the training images are from two separate sources, the color FERET data set [21] and the FDDB data set [13]. We evaluate our system with the FERET data set for pose recognition first, and then conduct experiments with the FDDB data set for face detection. We also compare our results with several baselines. All the results shown in this section are averaged over 10 runs on a 2.4GHz Intel i7 processor with 8GB memory. The major part of the algorithm is written in Matlab $2010b$ without extra optimization and run in a single thread.

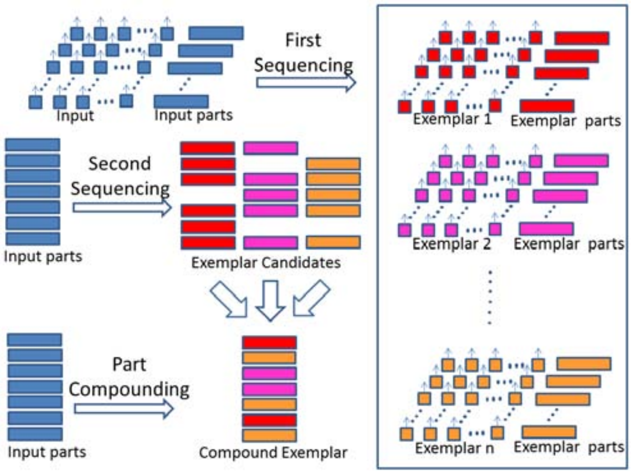

Figure 4. Diagram of 2D sequencing and compounding procedures. In the first phase, the 1D sequencing searches for the optimized exemplar rows matching to each input row. In the second phase all input and exemplar rows are assembled into columns. Next, 1D sequencing finds the optimal matching of the input column to each exemplar column. Compounding takes the best rows from different exemplars and creates a new exemplar that has a higher similarity.

The initial exemplars of the training data set are selected from the FERET data set. The full data set consists of 7810 single face images with face poses spanning from $-90°$ to $90°$ in yaw. It also provides ground truth of pose angle and locations of eyes, nose and mouth for each face. Based on ground truth, we use different rectangle bounding-boxes to extract face exemplars. We manually define 9 subclasses to represent 9 different yaw angles, which are $0°$, $\pm 20°$, $\pm 45°$, $\pm 70°$ and $\pm 90°$ and fill each subclass with 10 to 15 face exemplars. Since this initial data set is insufficient to cover all face poses, we iteratively run the algorithm on a validation data set to fill up the missing cases. The validation data set is selected from the FDDB data set.

The FDDB data set includes 2845 real life images with 5171 faces (labeled with ellipses as ground truth) originally partitioned into 10 fixed folds. In our experiments, each time we use one of the folds as the validation set and average the results of 10-fold cross validations. New exemplars that improve system performance are added to the final training set. To avoid the overfitting problem and reduce the computational cost, we limit the total number of exemplars stored in the final training set to 200. If the number is exceeded, we replace the least used exemplar with a new one from the rest of the current validation fold and continue the iteration until no further improvement is obtained. After the validation stage, we have 200 exemplars in about 18 poses. Figure 5 shows some face examples of the final training set.

Method	$-90°$	$-80°$	$-67.5°$	$-45°$	$-22.5°$	$0°\& \pm 10°$	$+22.5°$	$+45°$	$+67.5°$	$+90°$
VARIS	96%	90%	94%	97%	91%	98%	89%	94%	98%	99%
ESVM	87%	80%	84%	92%	84%	93%	81%	83%	87%	91%

Table 1. Pose recognition comparison between VARIS with compounding and ESVM.

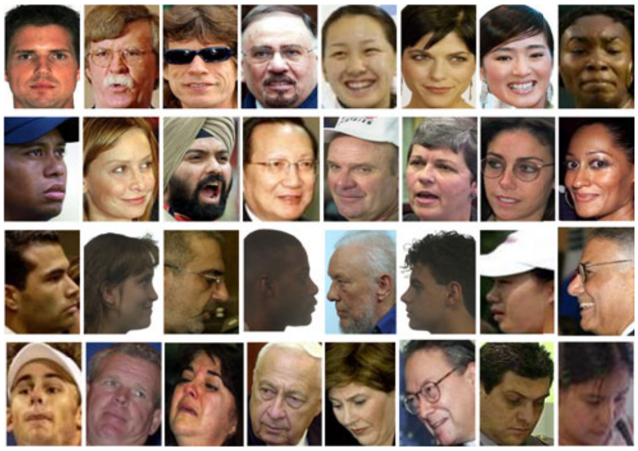

Figure 5. Examples of face exemplars. The final training data set includes human faces of different genders, ages, races, poses and appearance.

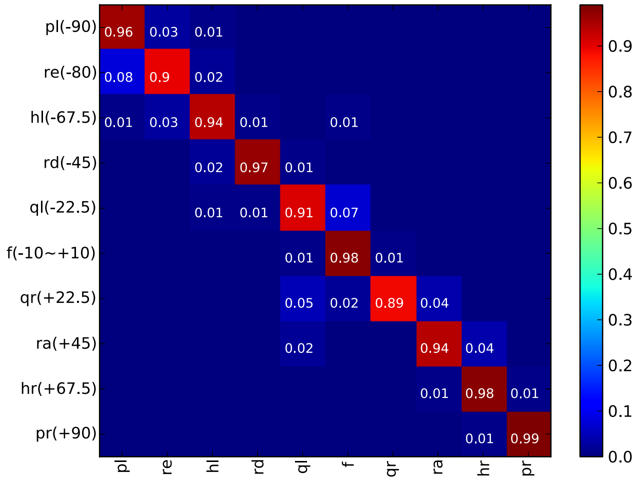

Figure 6. Pose recognition results. The confusion matrix shows the correct and incorrect pose estimations of VARIS with compounding on the FERET face dataset.

5.1. Pose Recognition

To evaluate the performance of our system, we first run the pose recognition test on the FERET data set. All the training exemplars are normalized to 96×72 pixels and then represented by 31-dimensional HOG features as in [9]. The pose of the winning subclass is compared against ground truth and marked as a correct recognition if the deviation is within $10°$. Figure 6 shows the confusion matrix of the results. The average pose recognition rate is around 94%.

We also evaluate the ESVM approach proposed by [19] with the FERET data set. All images in one fold (about 520 faces) of the FDDB data set are selected as the positive training exemplars. To make a fair comparison, we add the non-duplicated exemplars of our final training set to ESVM's training set. Then we use a second fold as the validation set to calibrate ESVM. The negative data is carefully selected from non-face images. Table 1 shows that VARIS with compounding completely defeats the ESVM approach in the discrete pose recognition test.

In the pose recognition experiment, we are also concerned with the effect of the parameter k on the system's recognition rate and the computational speed, where k determines the number of returned nearest neighbors for each input feature vector. Since the indexing step of VARIS is highly efficient and it only takes a few milliseconds to look up the neighbors, most of the computation time is spent on the sequencing step. A larger value of k might increase the probability of correct recognition but will definitely increase the computational complexity. Figure 6 shows the recognition rates versus the computational time at different values of k.

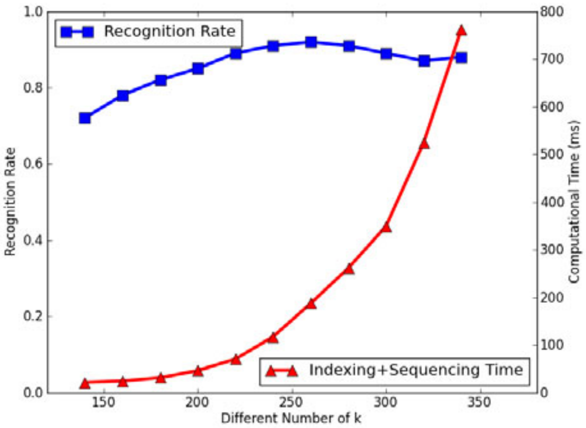

Figure 7. Computational efficiency. Computation time and recognition rates are measured under different settings of k. The time is the summation of indexing and sequencing steps for all exemplars.

5.2. Face Detection

In the multi-view face detection experiment, we test our system on the FDDB data set. We use the same fold of the data set that is used in pose recognition experiment and separate the rest as testing images. During the testing stage, a 96×72 sliding window with 25% overlapping is applied to scan the input image in scale-space for possible face locations. A 320×240 testing image returns with 800 detection windows in 10 scales. Since VARIS has large tolerance to spatial variations, the number of scanning windows is significantly reduced. In the end, the standard non-maxima-suppression approach is applied to remove overlapping detections. We set $\sigma = 0.02$ (Eq. 3 & 4), $t_w = 0.1$ (Eq. 4), $\alpha = -0.25$ (Eq. 5) and $t_c = 0.2$ for all exemplars. We also set k equal to 220 for a higher computational efficiency. To further improve the detection speed, we prune exemplars that have few matching features after the indexing stage. Some detection results are displayed in Figure 9.

To validate the performance of VARIS with exemplar compounding, we compare it with the ESVM approach and the LSVM [9] approach in the context of multi-view face detection. The LSVM is trained with thousands of positive images from different sources, including the training data set of VARIS, and a negative data set. The ESVM uses the same data set from the pose recognition section. We implement these two systems both in their authors' default configurations. We also add a few results from some previous work as baselines[1], which includes Viola-Jones face detector [24], Mikolajczyk *et al.*'s face detector [20] and Subburaman and Marcel's face detector [22].

We adopt the same evaluation criterion as in [13] that represents the degree of matching between a detection bounding-box (bb_i) and ground truth (gt_j) by using the ratio of intersected regions to joined regions:

$$M(bb_i, gt_j) = \frac{region(bb_i) \bigcap region(gt_j)}{region(bb_i) \bigcup region(gt_j)} \quad (7)$$

Based on our detection results and ground truth, a discrete Receiver Operating Characteristic (ROC) curve is generated by the FDDB evaluation toolkit. The estimated bounding-box is considered as a correct detection if the overlapping is more than 50%, otherwise a false positive detection is granted. Figure 8 shows the ROCs of the approaches mentioned above as well as two different versions of VARIS, one with compound exemplars and the other without. The curves show that VARIS achieves better detection results than the compared approaches. We also note that the exemplar compounding scheme is instrumental to VARIS in achieving high detection rates.

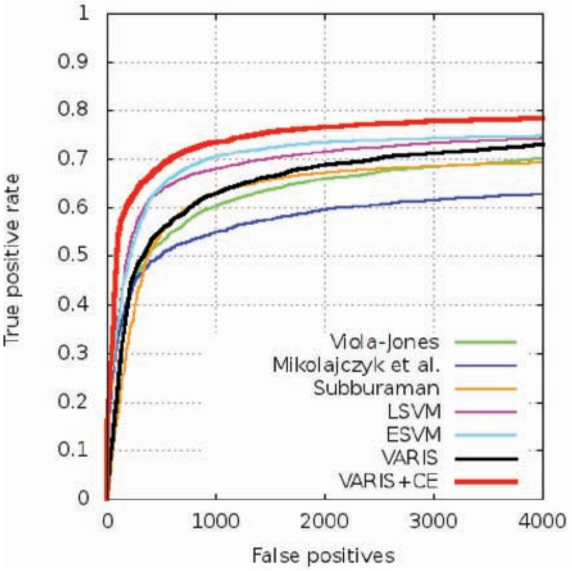

Figure 8. Discrete ROC curve. VARIS+CE is VARIS with compound exemplars. It outperforms all other methods.

6. Conclusion

In this work we presented a new approach that addresses the multi-view face detection problem. The proposed framework is based on maximizing the aggregate similarity score between the input image and a compound face exemplar which is dynamically assembled from classified exemplar patches. We have shown that VARIS with compounding outperforms the current state-of-the-art object detection methods in both face detection and pose recognition tasks. Although more research work is needed to optimize the indexing and sequencing steps, we confirm our initial hypothesis that object detection can be done by comparing similarities between the input image and class exemplars with topological constraints. We believe our exemplar compounding scheme is a promising avenue to further improve object detection performance.

Acknowledgements. This research was supported by NSF grant IIS-0905593. We thank Simone Franzini for sharing source code of RISq.

References

[1] M. Bar. The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11(7):280 – 289, 2007. 2

[2] J. S. Beis and D. G. Lowe. Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In *CVPR*, 1997. 3

[3] J. Ben-Arie. Method of Recognition of Human Motion, Vector Sequences and Speech. US Patent 7,366,645, April 2008. 1

[1] The results have been released with the FDDB data set and are available at vis-www.cs.umass.edu/fddb/results.html

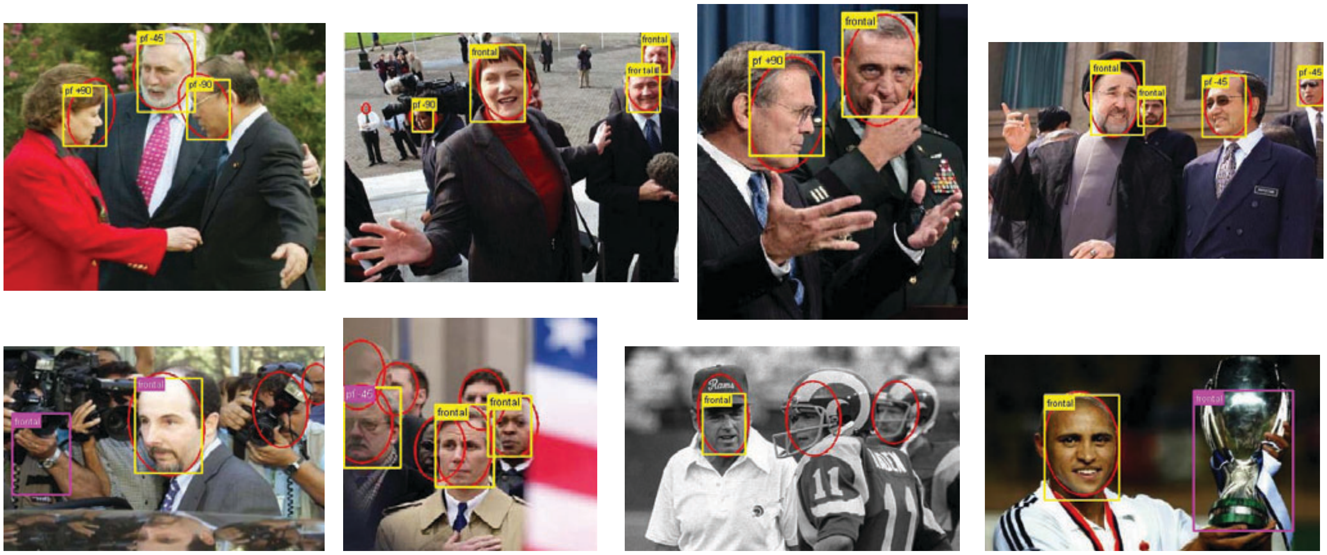

Figure 9. Detection results. The ground truth is labeled by red ellipses and our detection is labeled by squares. First row shows the correct detections where both face locations and face poses are correctly estimated. Second row shows the cases in which VARIS doesn't work well. Faces labeled only by red ellipses are the missed cases. The magenta texts label the faces that are correctly detected but the poses are not correct. The magenta text and square denote the false detection cases.

[4] J. L. Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 1975. 3

[5] M. Bergtholdt, J. Kappes, S. Schmidt, and C. Schnörr. A Study of Parts-Based Object Class Detection Using Complete Graphs. *IJCV*, 2010. 2

[6] H. Cai, F. Yan, and K. Mikolajczyk. Learning weights for codebook in image classification and retrieval. In *CVPR*, 2010. 4

[7] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *CVPR*, 2007. 2

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1, 3

[9] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010. 1, 2, 3, 6, 7

[10] S. Franzini and J. Ben-Arie. Speech Recognition by Indexing and Sequencing. *International Journal of Computer Information Systems and Industrial Management Applications*, 2012. 1, 4

[11] A. Frome, F. Sha, Y. Singer, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *ICCV*, 2007. 4

[12] A. Frome, Y. Singer, and J. Malik. Image retrieval and classification using local distance functions. In *NIPS*, 2006. 2

[13] 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. 1, 5, 7

[14] N. Kumar, L. Zhang, and S. K. Nayar. What is a good nearest neighbors algorithm for finding similar patches in images? In *ECCV*, 2008. 3

[15] 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. 2

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 3

[17] S. Maji and A. Berg. Max-margin additive classifiers for detection. In *ICCV*, 2009. 4

[18] T. Malisiewicz and A. A. Efros. Recognition by association via learning per-exemplar distances. In *CVPR*, 2008. 2

[19] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 1, 2, 4, 6

[20] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. *ECCV*, 2004. 7

[21] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing*, 1998. 5

[22] V. B. Subburaman and S. Marcel. Fast bounding box estimation based face detection. In *ECCV, Workshop on Face Detection: Where we are, and what next?*, 2010. 7

[23] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. J. V. Gool. Towards multi-view object class detection. In *CVPR*, 2006. 2

[24] M. Viola, M. J. Jones, and P. Viola. Fast multi-view face detection. In *CVPR*, 2003. 1, 2, 7

[25] P. Wang and Q. Ji. Multi-view face detection under complex scene based on combined svms. In *ICPR*, 2004. 1, 2

[26] C. Zhang and Z. Zhang. A survey of recent advances in face detection. Technical Report MSR-TR-2010-66, Microsoft Research, 2010. 1