Discriminative Local Sparse Representations for Robust Face Recognition

Yi Chen, Umamahesh Srinivas, Thong T. Do, Vishal Monga, and Trac D. Tran

Abstract—A key recent advance in face recognition models a test face image as a *sparse* linear combination of a set of training face images. The resulting sparse representations have been shown to possess robustness against a variety of distortions like random pixel corruption, occlusion and disguise. This approach however makes the restrictive (in many scenarios) assumption that test faces must be perfectly aligned (or registered) to the training data prior to classification. In this paper, we propose a simple yet robust *local block-based sparsity model*, using adaptively-constructed dictionaries from local features in the training data, to overcome this misalignment problem. Our approach is inspired by human perception: we analyze a series of local discriminative features and combine them to arrive at the final classification decision. We propose a *probabilistic graphical model* framework to explicitly mine the conditional dependencies between these distinct sparse local features. In particular, we learn discriminative graphs on sparse representations obtained from distinct local slices of a face. Conditional correlations between these sparse features are first discovered (in the training phase), and subsequently exploited to bring about significant improvements in recognition rates. Experimental results obtained on benchmark face databases demonstrate the effectiveness of the proposed algorithms in the presence of multiple registration errors (such as translation, rotation, and scaling) as well as under variations of pose and illumination.

Index Terms—Face recognition, sparse representation, local sparse features, discriminative graphical models, boosting.

I. INTRODUCTION

The problem of automatically verifying the identity of a certain person using a live face capture and comparing against a stored database of human face images has witnessed considerable research activity over the past two decades. The rich diversity of facial image captures,

due to varying illumination conditions, spatial resolution, pose, facial expressions, occlusion and disguise, offers a major challenge to the success of any automatic human face recognition system. A comprehensive survey of face recognition methods in literature is provided in [1].

In face recognition, indeed any image-based classification problem in general, representative features are first extracted from images typically via projection to a feature space. A classifier is then trained to make class assignment decisions using features obtained from a set of training images. One of the most popular dimensionality-reduction techniques used in computer vision is principal component analysis (PCA). In face recognition, PCA-based approaches have led to the use of eigenpictures [2] and eigenfaces [3] as features. Other approaches have used local facial features [4] like the eyes, nose and mouth, or incorporated geometrical constraints on features through structural matching. An important observation is that different (photographic) versions of the same face approximately lie in a linear subspace of the original image space [5]–[8]. A variety of classifiers have been proposed for face recognition, ranging from template correlation to nearest neighbor and nearest subspace classifiers, neural networks and support vector machines (SVM) [1].

Recently, the merits of exploiting parsimony in signal representation and classification have been demonstrated in [9]–[11]. In their seminal work, Wright *et al.* [9] argue that a test face image approximately lies in a low-dimensional subspace spanned by (lexicographically ordered) training images themselves. If sufficient training is available, a new test face image has a naturally sparse representation in this overcomplete basis. The sparse vector can be obtained via many norm minimization techniques and is then employed directly for recognition by computing a class (face) specific reconstruction error. Note that there is no *offline* training stage in sparsity based face recognition [9], instead the training samples in the dictionary are used directly at the time of testing/recognizing a test face image. The dictionary may be expanded hence as more training (variants of a face image) becomes available. This sparsity-based face recognition algorithm has been shown [9] to yield

markedly improved recognition performance over traditional efforts in face recognition under various conditions, including illumination, disguise, occlusion, and random pixel corruption.

In many real world scenarios, test images for identification obtained by face detection algorithms are not perfectly registered with the training samples in the databases. The sparse subspace assumption in [9], however, requires the test face image to be well aligned to the training data prior to classification. Recent approaches have attempted to address this misalignment issue in sparsity-based face recognition [12]–[14], usually by jointly optimizing the registration parameters and sparse coefficients and thus leading to more complex systems.

It is well known that, compared to global features, local features may contain more crucial information for representation in many signal and image processing applications. One such example is the block-based motion estimation technique which has been successfully employed in multiple popular video compression standards.

Inspired by the success of locality in recognition, our proposal is the development and use of *sparse local features* for face recognition[1]. As our first contribution, we propose a robust yet simpler approach to handle the misalignment problem via a *local block-based sparsity model*. We are motivated by the observation that a block in the test image can be sparsely represented by a linear combination of blocks in the training images within a spatially-neighboring region, and the sparse representation contains the identity information for the block. The final class decision relies on a combination of decisions from multiple local sparse representations (as observed earlier, the more discriminative facial features such as eyes, nose and mouth constitute a good set of local features). This approach exploits the capability of the local sparsity model to capture relatively stationary features under different types of variations and registration errors.

The presence of multiple feature representations (i.e., the distinct local features) naturally leads to the question: how can we combine the decisions based on multiple local features into a global class decision in the best way possible? A variety of heuristic classifier fusion schemes have been proposed in literature (see [16] for example). The outputs of individual classifiers constitute *high-level* features. It is reasonable to expect better classification performance by directly exploring the correlation between *low-level* features. In order to explicitly mine such conditional dependencies between these distinct sparse local features, we propose a probabilistic graphi-

cal model framework as the second main contribution of this paper[2]. In particular, we learn discriminative graphs on sparse representations obtained from distinct local slices of a face. Conditional correlations between these sparse features are first discovered by learning discriminative tree graphs [18] on each distinct feature set. The initial disjoint trees are then thickened, i.e., augmented with more edges to capture newly learnt feature correlations, via boosting [19] on disjoint graphs. Probabilistic graphical models offer additional benefits in terms of robustness to limited training, and reduced computational complexity of inference.

It is informative to contrast our contribution with recent work in robust face recognition that considers registration errors. Huang *et al.* [12] consider the scenario where the test images can be represented in terms of all training images and (linearized versions of) their image plane transformations. The computational cost scales with the complexity of the plane transformation. In [14], the difficult nonconvex problem of simultaneous optimization over sparse coefficients and registration parameters is relaxed via sequential iterative minimization. In addition, a novel projector-based illumination system has been proposed to capture variations in scene lighting. In our proposed approach however, the registration parameters are not explicitly determined. Instead, robustness to misalignment is introduced by augmenting the training with spatially local blocks from each training image. Another significant departure from existing sparse representation-based approaches is our use of a principled strategy via graphical models to explicitly mine feature dependencies, instead of performing classification using only reconstruction residuals.

The rest of this paper is organized as follows. Section II provides a review of sparsity based face recognition, as well as an overview of probabilistic graphical models. The two main contributions of this paper are presented in Section III. An extensive set of experiments is performed on popular face recognition databases to validate the effectiveness of our proposed framework, and results under varying practical settings are provided in Section IV. Section V summarizes the contributions and concludes the paper.

II. Background

A. *Sparse Representation-based Classification*

As mentioned earlier, algorithmic advances in face recognition have been comprehensively surveyed in the

---

[1]Part of this material has been presented in IEEE ICIP 2010 [15].

[2]Part of this material has been accepted to IEEE Asilomar Conf. 2011 [17].

literature [1]. Here, we primarily review recent pioneering work in sparse representation-based face recognition [9], which forms the foundation for our proposed contribution. This method advocates the use of sparse representation in a *discriminative* setting, a novel advance over previous work which exploited sparsity from a *signal recovery* standpoint.

First, let us introduce the standard notation that will be used throughout this paper. Suppose there are K different classes (corresponding to unique faces), labeled $C_1, \ldots, C_K$, and there are N_i training samples (each in $\mathbb{R}^n$) corresponding to class $C_i, i = 1, \ldots, K$. The training samples corresponding to class C_i can be collected in a matrix $\boldsymbol{D}_i \in \mathbb{R}^{n \times N_i}$, and the collection of all training samples is expressed using the matrix:

$$\boldsymbol{D} = [\boldsymbol{D}_1 \ \boldsymbol{D}_2 \ \ldots \ \boldsymbol{D}_K], \tag{1}$$

where $\boldsymbol{D} \in \mathbb{R}^{n \times T}$, with $T = \sum_{k=1}^{K} N_k$. A new test sample $\mathbf{y} \in \mathbb{R}^n$ can be expressed as a sparse linear combination of the training samples:

$$\mathbf{y} = \boldsymbol{D}\boldsymbol{\alpha}, \tag{2}$$

where $\boldsymbol{\alpha}$ is expected to be a sparse vector (i.e., only a few entries in $\boldsymbol{\alpha}$ are nonzero). This is an underdetermined system of linear equations. The classifier seeks the sparsest representation by solving:

$$\hat{\boldsymbol{\alpha}} = \arg\min \|\boldsymbol{\alpha}\|_0 \quad \text{subject to} \quad \|\boldsymbol{D}\boldsymbol{\alpha} - \mathbf{y}\|_2 \leq \varepsilon, \tag{3}$$

where $\|\cdot\|_0$ denotes the number of nonzero entries in the vector. Under a set of sufficient conditions (that hold in general for the above problem set-up), it has been shown theoretically [20] that the non-convex optimization problem represented by (3) can be relaxed to the following convex optimization problem:

$$\hat{\boldsymbol{\alpha}} = \arg\min \|\boldsymbol{\alpha}\|_1 \quad \text{subject to} \quad \|\boldsymbol{D}\boldsymbol{\alpha} - \mathbf{y}\|_2 \leq \varepsilon. \tag{4}$$

Alternatively, the problem in (3) can be solved by greedy pursuit algorithms [21]–[23].

Once the sparse vector is recovered, the identity of $\mathbf{y}$ is given by the minimal residual

$$\text{identity}(\mathbf{y}) = \arg\min_i \|\mathbf{y} - \boldsymbol{D}\delta_i(\hat{\boldsymbol{\alpha}})\|, \tag{5}$$

where $\delta_i(\boldsymbol{\alpha})$ is a vector whose only nonzero entries are the same as those in $\boldsymbol{\alpha}$ but only associated with class C_i. The particular choice of class-specific residuals makes the task of decision assignment computationally trivial.

Often, it is necessary to check if a particular test image belongs to any of the available classes. The authors develop a sparsity concentration index (SCI) to decide if a test image is valid or not. Given a sparse coefficient vector $\boldsymbol{\alpha} \in \mathbb{R}^T$, the SCI is defined as follows:

$$\text{SCI}(\boldsymbol{\alpha}) = \frac{K \cdot \max_i \|\delta_i(\boldsymbol{\alpha})\|_1 / \|\boldsymbol{\alpha}\|_1 - 1}{K - 1} \in [0, 1]. \tag{6}$$

A high value of SCI indicates a sparse representation corresponding to a valid test image, while a value close to 0 indicates that the feature coefficients are distributed across all classes.

B. Probabilistic Graphical Models

We provide a brief overview of probabilistic graphical models from an inference (hypothesis testing) viewpoint. Discriminative graphs will be used to model the class conditional densities $f(\boldsymbol{\alpha}|C_i)$, i.e., a set of p.d.fs defined on the (sparse) coefficient vector which are employed to make class assignments (each class C_i corresponds to the i-th person in the database).

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined by a collection of nodes $\mathcal{V} = \{v_1, \ldots, v_r\}$ and a set of (undirected) edges $\mathcal{E} \subset \binom{\mathcal{V}}{2}$, i.e., the set of unordered pairs of nodes. A probabilistic graphical model is obtained by defining a random vector on $\mathcal{G}$ such that each node represents one (or more) random variables and the presence of edges indicates conditional dependencies. The graph structure thus enforces a particular factorization of the joint probability distribution in terms of pairwise marginals.

The use of graphical models in applications has been motivated by practical concerns like insufficient training to learn models for high-dimensional data and the need for reduced computational complexity in realtime tasks [24], [25]. Graphical models offer an alternate visualization of a probability distribution from which conditional dependence relations can be easily identified. Graphical models also enable us to draw upon the rich resource of efficient graph-theoretic algorithms to learn complex models and perform inference.

Graphical models can be learnt from data in two different settings: generative and discriminative. In generative learning, a *single* graph is learnt to approximate a given distribution by minimizing a measure of *approximation error*. Generative learning approaches trace their origin to Chow and Liu's [26] idea of learning the optimal tree approximation $\hat{p}$ of a multivariate distribution p using first- and second-order statistics:

$$\hat{p} = \arg\min_{\hat{p} \text{ is a tree}} D(p||\hat{p}), \tag{7}$$

where $D(p||\hat{p}) = E_p[\log(p/\hat{p})]$ denotes the Kullback-Leibler (KL) divergence. This optimization problem is shown to be equivalent to a maximum-weight spanning tree (MWST) problem. In discriminative learning, on

the other hand, a pair of graphs is jointly learnt from a pair of empirical estimates by minimizing the *classification error*. (Note that we consider binary classification problems here to reduce notational clutter. The approach naturally extends to multi-class problems by learning graphs in a one-versus-all manner.)

Recently, Tan *et al.* [18] proposed a graph-based discriminative learning framework, based on maximizing an approximation to the *J*-divergence, which is a symmetric extension of the KL-divergence. Given two probability distributions p and q, their *J*-divergence is defined as: $J(p,q) = D(p||q) + D(q||p)$. The tree-approximate *J*-divergence is then defined as:

$$\hat{J}(\hat{p},\hat{q};p,q) = \int (p(x) - q(x)) \log \left[\frac{\hat{p}(x)}{\hat{q}(x)} \right] dx, \quad (8)$$

which measures the "separation" between tree-structured approximations $\hat{p}$ and $\hat{q}$. Using the key observation that maximizing the *J*-divergence minimizes the upper bound on probability of classification error, the discriminative tree learning problem is then stated (in terms of empirical estimates $\tilde{p}$ and $\tilde{q}$) as follows:

$$(\hat{p},\hat{q}) = \arg \max_{\hat{p},\hat{q} \text{ trees}} \hat{J}(\hat{p},\hat{q};\tilde{p},\tilde{q}). \quad (9)$$

It is shown in [18] that this optimization further decouples into two MWST problems:

$$\hat{p} = \arg \min_{\hat{p} \text{ tree}} D(\tilde{p}||\hat{p}) - D(\tilde{q}||\hat{p}) \quad (10)$$

$$\hat{q} = \arg \min_{\hat{q} \text{ tree}} D(\tilde{q}||\hat{q}) - D(\tilde{p}||\hat{q}). \quad (11)$$

Here, (10) and (11) bring out the distinction (from a classification viewpoint) between: (i) using generative learning techniques to separately learn $\hat{p}$ and $\hat{q}$ and then performing inference, and (ii) simultaneously learning a pair of graphs discriminatively. In (10), the optimal $\hat{p}$ is chosen to minimize its (KL-divergence) distance from $\tilde{p}$ and *simultaneously* maximize its distance from $\tilde{q}$.

The discussion so far mainly considers tree graphs, which are fully connected acyclic graphical structures. The computational burden of learning tree graphs is significantly reduced owing to the sparse connectivity. This feature however imposes a limitation on the complexity of the model so learnt. This inherent trade-off between generalization and performance poses a serious challenge to the application of graphical models in various tasks.

Our contribution as described in the remainder of this paper proposes an extension of discriminative graph learning for the purpose of face recognition, utilizing distinct local features from a block-based sparsity model.

III. FACE RECOGNITION VIA LOCAL DECISIONS FROM LOCALLY ADAPTIVE SPARSE FEATURES

The two main contributions of this paper are presented in Sections III-A and III-B respectively. Section III-A explains the process of obtaining local sparse features. In Section III-B, two different methods of combining class decisions are proposed: (i) based on reconstruction error, and (ii) using graphical models.
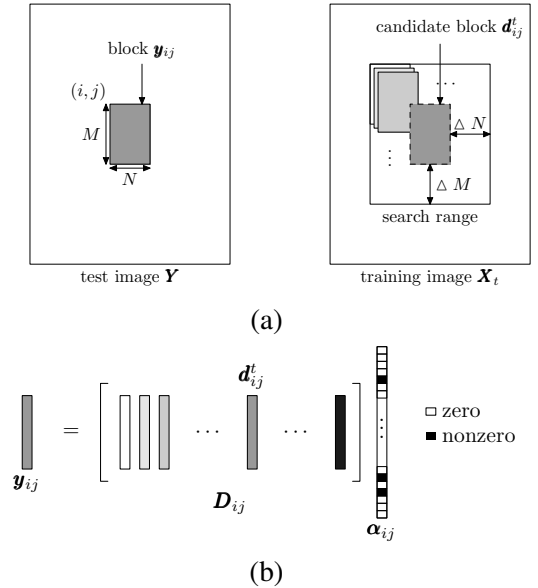

Fig. 1. Representation of a block in the test image from a locally adaptive dictionary. (a) The blocks in the test and training images (only one training sample is displayed). (b) Sparse representation $\boldsymbol{y}_{ij} = \boldsymbol{D}_{ij} \boldsymbol{\alpha}_{ij}$.

A. Locally Adaptive Sparse Representations

The method in [9] advances practical face recognition by enabling significantly enhanced robustness to distortions like occlusion, pixel corruption and disguise. However, as discussed in Section I, the subspace model requires precise registration making their approach vulnerable to alignment errors of rotation, translation and scaling that are natural to face capture processes. To deal specifically with disguise, Wright *et al.* [9] do suggest a block-partitioning scheme which to a first order captures local face image characteristics while still suffering from misalignment. The superior compression ability of local features compared to global representations is also well-known from applications like block-based motion estimation in video coding. In other words, local sparsity is beneficial from the recovery standpoint. In this work, we consummate this intuition by designing adaptive dictionaries for each "local block" such that the resulting (local) sparse representations naturally exhibit robustness to alignment errors.

To achieve this, in the proposed local sparse representation model for face recognition, we adopt the inter-frame sparsity model in [27], where a block in a video frame is expressed as a sparse linear combination of a few spatially-adjacent blocks from the reference frames. An illustration of the proposed model in shown in Fig. 1, where a block in the (possibly misaligned) test image $\boldsymbol{Y}$ is sparsely represented by a locally adaptive dictionary consisting of blocks in the training images $\{\boldsymbol{X}_t\}_{t=1,...,T}$ within the same spatial neighborhood. Note that for illustration, only one training image is shown in Fig. 1(a). Specifically, let $\boldsymbol{y}_{ij} \in \mathbb{R}^{MN}$ be the vectorized $M \times N$ block in the test image $\boldsymbol{Y}$ with the upper left pixel located at (i, j). The search region $\boldsymbol{S}_{ij}^t$ in the t-th training image $\boldsymbol{X}_t$ is an $(M+2\triangle M) \times (N+2\triangle N)$ image region:

$$\boldsymbol{S}_{ij}^t = \begin{bmatrix} x_{i-\triangle M,j-\triangle N}^t & \cdots & x_{i-\triangle M,j+N-1+\triangle N}^t \\ \vdots & \ddots & \vdots \\ x_{i+M-1+\triangle M,j-\triangle N}^t & \cdots & x_{i+M-1+\triangle M,j+N-1+\triangle N}^t \end{bmatrix}.$$

The local dictionary $\boldsymbol{D}_{ij}$ for the block $\boldsymbol{y}_{ij}$ is then constructed by all $M \times N$ blocks within the search regions $\{\boldsymbol{S}_{ij}^t\}_{t=1,2,...,T}$ in the T training images:

$$\boldsymbol{D}_{ij} = \begin{bmatrix} \boldsymbol{D}_{ij}^1 & \boldsymbol{D}_{ij}^2 & \cdots & \boldsymbol{D}_{ij}^T \end{bmatrix},$$

where each

$$\boldsymbol{D}_{ij}^t = \begin{bmatrix} \boldsymbol{d}_{i-\triangle M,j-\triangle N}^t & \boldsymbol{d}_{i-\triangle M,j-\triangle N+1}^t & \cdots & \boldsymbol{d}_{i+\triangle M,j+\triangle N}^t \end{bmatrix}$$

is an $(MN) \times \big((2\triangle M+1)(2\triangle N+1)\big)$ sub-dictionary whose columns are the vectorized blocks in the t-th training image defined in the same way as $\boldsymbol{y}_{ij}$.

In this way, a locally-adaptive dictionary $\boldsymbol{D}_{ij}$ is constructed for every block of interest in the test image. The size of the dictionary depends on the non-stationary behavior of the data as well as the level of computational complexity we can afford. For significant registration errors, the local dictionaries can be augmented by distorted versions of the local blocks in the training data for better performance at the cost of higher computational intensity. Compared to the original global approach, the dictionary $\boldsymbol{D}_{ij}$ captures local characteristics better and yields a reasonable approximation of the training image at the block level. Our approach is different from patch-based dictionary learning [28] in multiple aspects: (i) we emphasize the local adaptivity of the dictionaries, and (ii) our dictionaries are constructed by simply taking blocks from training data without any sophisticated learning process.

We propose that the block $\boldsymbol{y}_{ij}$ in the misaligned image $\boldsymbol{Y}$ can be approximated by a linear combination of only a few atoms in the dictionary $\boldsymbol{D}_{ij}$:

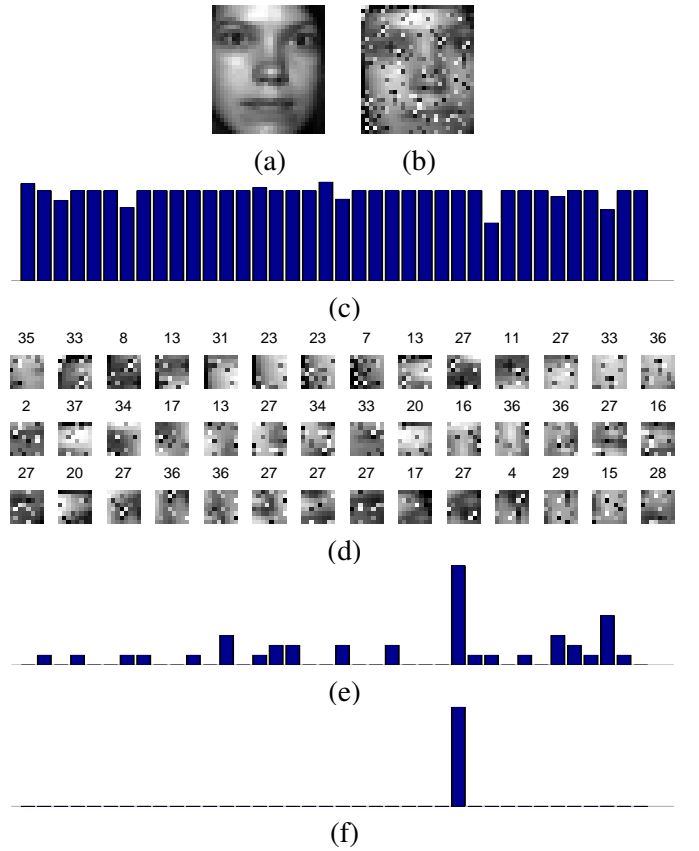$$\boldsymbol{y}_{ij} = \boldsymbol{D}_{ij} \boldsymbol{\alpha}_{ij}, \tag{12}$$

Fig. 2. Example of the proposed sparsity-based approach using multiple test blocks. (a) Original image (Class 27). (b) Distorted test image $\boldsymbol{Y}$. (c) Residuals using the original global approach: identity$(\boldsymbol{Y}) = 29$. (d) Classification results for each of the 42 blocks $\{\boldsymbol{y}_l\}_{l=1,...,42}$. (e) Number of votes for the kth class, $k = 1,...,38$. Identity$(\boldsymbol{Y}) = 27$. (f) Probability of (identity$(\boldsymbol{Y}) = k$), $k = 1,...,38$. Identity$(\boldsymbol{Y}) = 27$.

where $\boldsymbol{\alpha}_{ij}$ is a sparse vector, as illustrated in Fig. 1(b). Similar to the global case, the sparse vector is recovered by solving the following optimization problem:

$$\hat{\boldsymbol{\alpha}}_{ij} = \arg\min \|\boldsymbol{\alpha}_{ij}\|_0 \text{ subject to } \|\boldsymbol{D}_{ij}\boldsymbol{\alpha}_{ij} - \boldsymbol{y}_{ij}\|_2 \le \varepsilon. \tag{13}$$

Note that the resulting complexity of the overall algorithm is still manageable since the above sparse recovery is performed on a small block with a dictionary of modest size. After the sparse vector $\hat{\boldsymbol{\alpha}}_{ij}$ is obtained, the error residual with respect to the k-th class sub-dictionary is calculated by

$$r^k(\boldsymbol{y}_{ij}) = \|\boldsymbol{y}_{ij} - \boldsymbol{D}_{ij}\boldsymbol{\delta}_k(\hat{\boldsymbol{\alpha}}_{ij})\|_2, \tag{14}$$

where $\boldsymbol{\delta}_k(\hat{\boldsymbol{\alpha}}_{ij})$ is as defined in (5). Then, the identity of the test block can be determined by the minimal residual as follows:

$$\text{identity}(\boldsymbol{y}_{ij}) = \arg\min_{k=1,...,K} r^k(\boldsymbol{y}_{ij}). \tag{15}$$

The usage of a single block certainly cannot produce outstanding classification performance. To improve

the algorithm's robustness, we employ multiple blocks: solving the sparse recovery problem for each block individually, and then combining the results for all of the blocks. Blocks may be chosen manually in the areas with discriminative features (such as eyes, nose, and mouth), or areas with high SNR/more variations, or uniformly in the entire test image in non-overlapped or overlapped fashion. It should be noted that the blocks can be processed independently in parallel. Moreover, since blocks can be overlapped, our approach is computationally scalable - more computation simply delivers better recognition performance - a feature that will be illustrated by experimental results in Section IV.

Finally, we would like to remark that our locally adaptive sparse representation is a more general and more powerful framework comparing to the global sparse representation proposed in [9]. In other words, if we set the parameters $\triangle M$ and $\triangle N$ to zero, and further force the local sparse vectors $\boldsymbol{\alpha}_{ij}$ to be the same for all non-overlapped test block $\boldsymbol{y}_{ij}$, then what we get back is essentially the global sparse representation.

B. Recognition Decisions from Local Sparse Features

1) Classifiers based on reconstruction error: We first present two simple schemes that combine the individual recognition results from the blocks. Let $\{\boldsymbol{y}_l\}_{l=1,\ldots,L}$ be the L blocks in the test image $\boldsymbol{Y}$. (Note that in Section III-A, we have identified each block with the location (i,j) of its upper left pixel. Here, each block identifier l is implied to have unique correspondence with one such pixel location, and we will use $\boldsymbol{y}_l$ instead of $\boldsymbol{y}_{ij}$ henceforth.)

a) Majority voting:

$$\text{identity}(\boldsymbol{Y}) = \max_{k=1,\ldots,K} |\{l = 1,\ldots,L : \text{identity}(\boldsymbol{y}_l) = k\}|,$$
(16)

where $|S|$ denotes the cardinality of a set S and identity$(\boldsymbol{y}_l)$ is determined by (15).

b) Maximum likelihood: This is another intuitive approach of fusing classification results from multiple blocks. Let $\hat{\boldsymbol{\alpha}}_l$ be the recovered sparse representation vector of the block $\boldsymbol{y}_l$ and the local dictionary $\boldsymbol{D}_l$. We define the probability of $\boldsymbol{y}_l$ belonging to the k-th class to be inversely proportional to the residual associated with the dictionary atoms in the k-th class:

$$p_l^k = P(\text{identity}(\boldsymbol{y}_l) = k) = \frac{1/r_l^k}{\sum_{k=1}^{K}(1/r_l^k)},$$
(17)

where $r_l^k = \|\boldsymbol{y}_l - \boldsymbol{D}_l\boldsymbol{\delta}_k(\hat{\boldsymbol{\alpha}}_l)\|_2$ is the residual associated with the k-th class as in (14). The identity of the test

image $\boldsymbol{Y}$ is then given by

$$\text{identity}(\boldsymbol{Y}) = \arg\max_{k=1,\ldots,K} \log\left(\prod_{l=1}^{L} p_l^k\right).$$
(18)

The likelihood measure can also be used as a criterion for outlier rejection, since the probability of an outlier belonging to a particular class tends to be uniformly distributed among all training classes.

An example of the proposed approach fusing results of multiple local blocks is illustrated in Fig. 2 using the Extended Yale B Database [29], which consists of facial images of 38 individuals. More details about experiments will be discussed in Section IV. Fig. 2(a) shows an image belonging to the 27th class, and Fig. 2(b) shows the test image to be classified, which is the image in (a) distorted by rotation, scaling, and random pixel corruption. The distortion causes the failure of the original global approach in [9] in this case, as seen by the error residuals in Fig. 2(c) where the 29th class turns out to yield the minimal residual. For the proposed local approach, we use 42 blocks of size 8×8 chosen uniformly from the distorted test image. The blocks and class labels for each individual block are displayed in Fig. 2(d). Figs. 2(e) and (f) show the number of votes and the probability defined in (17), respectively. It is obvious that in both cases, the local approach yields the correct class label (i.e., the 27th class has the highest number of votes and the maximal probability). This example also highlights the robustness of local sparse representations under reduced feature dimensions, although the individual blocks are chosen uniformly instead of selectively corresponding to representative facial features.

2) Graphical models to mine feature correlations: The two schemes discussed above, albeit intuitively motivated, are essentially heuristic ways of fusing classifier outputs. We now present a two-stage probabilistic graphical model framework to directly exploit conditional correlations between features from local regions themselves. The overall framework is shown in Fig. 3.

We introduce some additional notation. Let $C_i, i = 1, 2, \ldots, K$ denote the i-th class of face images (as defined earlier), and let $\tilde{C}_i$ denote the class of face images complementary to class C_i, i.e., $\tilde{C}_i = \bigcup_{k=1,\ldots,K,k\neq i} C_k$. Let $\mathcal{B}_i$ denote the i-th binary classification problem of classifying a query face image (or corresponding feature) into C_i or $\tilde{C}_i$ ($i = 1,\ldots,K$). As will be clear shortly, defining K such binary problems is necessary for application of the discriminative graphical framework. The graphical model-based algorithm is summarized in Algorithm 1, and it consists of an *offline* stage to learn the discriminative graphs (Steps 1-4) followed by an
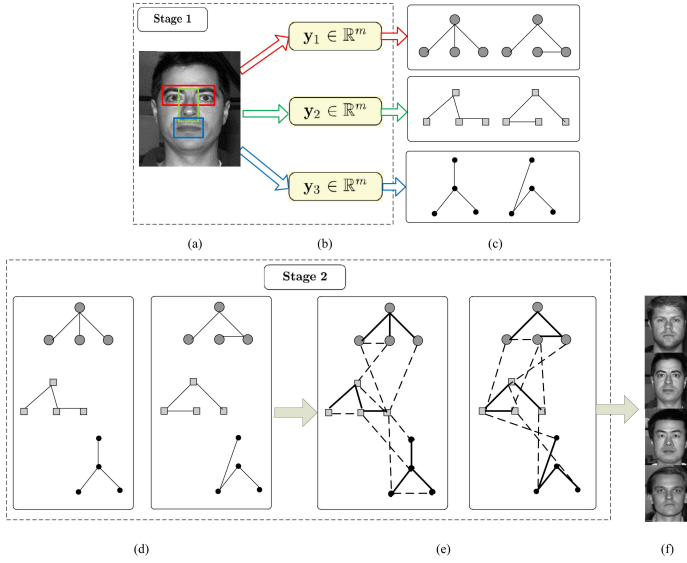
Fig. 3. Proposed framework for face recognition: (a) Target face image, (b) Local regions for extracting sparse features, (c) Initial pairs of *tree* graphs for each feature set, (d) Initial sparse graph formed by tree concatenation, (e) Final pair of thickened graphs; newly learned edges represented by dashed lines, (f) Graph-based inference. In (c)-(e), the graphs on the left and right correspond to distributions p (class C_i) and q (class $\tilde{C}_i$) respectively.

Algorithm 1 Discriminative graphical models for face recognition (Steps 1-4 offline)

1: **Feature extraction (training):** Obtain sparse representations $\boldsymbol{\alpha}_l, l = 1, \ldots, P$ in $\mathbb{R}^m$ from facial features, using adaptive locally block-sparsity model (19)
2: **Initial disjoint graphs:**
   For $l = 1, \ldots, P$
   Discriminatively learn pairs of m-node tree graphs $\mathcal{G}_l^p$ and $\mathcal{G}_l^q$ on $\{\boldsymbol{\alpha}_l\}$ obtained from training data
3: Separately concatenate nodes corresponding to p and q respectively, to generate initial graphs
4: **Boosting on disjoint graphs**: Iteratively thicken initial disjoint graphs via boosting to obtain final graphs $\mathcal{G}^p$ and $\mathcal{G}^q$
   {**Online process**}
5: **Feature extraction (test):** Obtain sparse representations $\boldsymbol{\alpha}_l, l = 1, \ldots, P$ in $\mathbb{R}^m$ from test image
6: **Inference**: Classify based on output of the resulting classifier using (20).

online stage (Steps 5-6) where a new test image is classified.

The offline stage involves extraction of features from training images, which comprise the empirical estimates from which approximate p.d.fs for each class are learnt after the graph thickening procedure. The individual steps in this stage are explained next.

a) Feature extraction: Let us first consider one of the local spatial regions in the face, say corresponding to the eyes. For the binary classification problem $\mathcal{B}_i$, dictionaries $\boldsymbol{D}_i$ and $\tilde{\boldsymbol{D}}_i$ are constructed according to the procedure in Section III-A, using samples from C_i and $\tilde{C}_i$ respectively. (Subscripts are dropped while denoting the dictionaries to avoid confusion, and they can be inferred from context.) Features in $\mathbb{R}^m$ are now extracted for any block $\boldsymbol{z}$ (spatially corresponding to eyes) by solving the sparse recovery problem:

$$\hat{\boldsymbol{\beta}} = \arg\min \|\boldsymbol{\beta}\|_0 \text{ subject to } \|\boldsymbol{D}\boldsymbol{\beta} - \boldsymbol{z}\|_2 \leq \varepsilon, \quad (19)$$

where $\boldsymbol{D} := [\boldsymbol{D}_i, \ \tilde{\boldsymbol{D}}_i]$. Features corresponding to other local regions are generated analogously. Training features (that form the overcomplete dictionary) for C_i are obtained by using training faces that are *known* to belong to class C_i, while features for $\tilde{C}_i$ are obtained by choosing representative training from $\tilde{C}_i$ as input to the feature extraction process.

b) Initial disjoint pairs of trees: The extraction of different sets of features from input face images

is performed offline. Each such representation may be viewed as a projection $\mathcal{P}_l : \mathbb{R}^n \mapsto \mathbb{R}^m$. In our framework we consider, in all generality, P distinct projections $\mathcal{P}_l, l = 1, 2, \ldots, P$. For every input image $\boldsymbol{y} \in \mathbb{R}^n$, P different features $\boldsymbol{\alpha}_l \in \mathbb{R}^m, l = 1, 2, \ldots, P$ are obtained. Fig. 3(b) depicts this process for the particular case $P = 3$, i.e., using eyes, nose and mouth as features. The different projections lead to local features that have complementary yet correlated information, since they arise from the same original face image.

Figs. 3(c)-(f) represent the graph learning process. We denote the class distributions corresponding to C_i and $\tilde{C}_i$ by p and q respectively, i.e., $f_p^i(\boldsymbol{\alpha}_l) = f(\boldsymbol{\alpha}_l | C_i)$ and $f_q^i(\boldsymbol{\alpha}_l) = f(\boldsymbol{\alpha}_l | \tilde{C}_i)$. A pair of m-node discriminative tree graphs $\mathcal{G}_l^p$ and $\mathcal{G}_l^q$ is learnt for each projection $\mathcal{P}_l, l = 1, 2, \ldots, P$, by solving (10) and (11). The local sparse features $\boldsymbol{\alpha}_l$ obtained from the P local blocks are used as empirical estimates to train the tree graphs[3]. By concatenating the nodes of the graphs $\mathcal{G}_l^p, l = 1, \ldots, P$, we have one initial sparse graph structure with Pm nodes (Fig. 3(d)). Similarly, we obtain another initial graph by concatenating the nodes of the graphs $\mathcal{G}_l^q, l = 1, \ldots, P$. We have now learnt (graphical) p.d.fs $\hat{f}_p^i(\boldsymbol{\alpha}_l)$ and $\hat{f}_q^i(\boldsymbol{\alpha}_l)$, where $\boldsymbol{\alpha}_l$ is the sparse feature vector obtained from the l-th local block ($l = 1, \ldots, P$), and i refers to the i-th binary classification problem $\mathcal{B}_i$. Inference based on these disjoint graphs can be interpreted as feature fusion assuming statistical independence of the individual target

---

[3]The same training faces present in the overcomplete dictionary are used to generate the sparse features to train the graphs.

image representations.

c) Discriminative graphs for classification: Although simple tree graphs can be learnt efficiently, their ability to model general distributions is limited. However, learning graphs with arbitrarily complex structure is known to be an NP-hard problem [30]. To overcome this trade-off, we learn different pairs of discriminative graphs over the same sets of nodes (but weighted differently) in different iterations via boosting and obtain a "thicker" graph by augmenting the original trees with the newly-learned edges [31]. Boosting [19] iteratively improves the performance of weak learners into a classification algorithm with arbitrarily accurate performance.

For each binary classification problem, the P pairs of tree graphs in Fig. 3(c) are discriminatively learnt [18] from distinct local regions of the face image using empirical estimates of distributions available from corresponding training samples of locally sparse features. In Fig. 3(c), an example instantiation is shown for $P = 3$ where the local regions correspond to eyes, nose and mouth respectively. They are subsequently thickened by the process of boosting [19], [31]. This process of learning new edges is tantamount to discovering new conditional correlations between distinct sets of local features, as illustrated by the dashed edges in Fig. 3(e). The thickened graphs $\hat{f}_p^i(\boldsymbol{\alpha})$ and $\hat{f}_q^i(\boldsymbol{\alpha})$ are therefore estimates of the true (but unknown) class conditional p.d.fs $f_p^i(\boldsymbol{\alpha}) = f(\boldsymbol{\alpha}|C_i)$ and $f_q^i(\boldsymbol{\alpha}) = f(\boldsymbol{\alpha}|\tilde{C}_i)$, where $\boldsymbol{\alpha}$ is the concatenated feature vector from all P blocks.

The graph learning procedure described so far is performed offline. The actual classification of a new test image is performed in an online process, explained next.

d) Feature extraction: The feature extraction is identical to the process described in the offline stage. Corresponding to each test image, local features $\boldsymbol{\alpha}_l, l = 1,\ldots,P$ are obtained by solving the individual sparse recovery problems.

e) Inference: Classification is performed in a *one-versus-all* manner by solving K separate binary classification problems $\mathcal{B}_i$. If $\hat{f}_p^i$ and $\hat{f}_q^i$ denote the final probabilistic graphical models learnt for C_i and $\tilde{C}_i$ ($i = 1, 2, ...K$) respectively, then the face image feature vector comprising of sparse coefficients from all the local blocks, i.e., $\boldsymbol{\alpha}$ is assigned to a class i^* according to the following decision rule:

$$i^* = \arg \max_{i\in\{1,...,K\}} \log\left(\frac{\hat{f}_p^i(\boldsymbol{\alpha})}{\hat{f}_q^i(\boldsymbol{\alpha})}\right). \qquad (20)$$

IV. EXPERIMENTS AND RESULTS

We test the proposed algorithm(s) on popular face databases. Experiments performed in [9] reveal the ro-

TABLE I
OVERALL RECOGNITION RATES USING CALIBRATED TEST IMAGES
FROM THE EXTENDED YALE B DATABASE (SECTION IV-A).

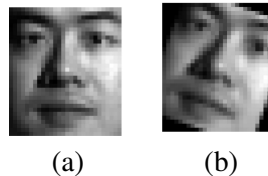| Method	Recognition rate (%)
LSGM	97.3
SRC	97.1
Eigen-NS	89.5
Eigen-SVM	91.9
Fisher-NS	84.7
Fisher-SVM	92.6

(a) (b)

Fig. 4. An example of rotated test images. (a) Original image and (b) the image rotated by 20 degrees clockwise.

bustness of the approach to distortions, under the assumption that the test images are well-calibrated. As a first result, we show in Section IV-A that our proposed approach produces equally competitive results on calibrated test images (with no registration errors) from the Extended Yale B database [29]. Subsequently, via experiments in Section IV-B, we establish the robustness of our approach to registration errors and a variety of other distortions. The ability to reject invalid images is tested in Section IV-C. Finally, we discuss different flavors of classifier fusion (to combine the local recognition decisions) in Section IV-D. MATLAB code corresponding to all the experiments and algorithms reported in this paper is available at: `http://signal.ee.psu.edu/FaceRec-LSGM.htm`.

A. Calibrated Test Images: No Alignment Errors

For this experiment we use the Extended Yale B database, which consists of 2414 perfectly-aligned frontal face images of size 192×168 of 38 individuals, 64 images per individual, under various conditions of illumination. In our experiments, for each subject we randomly choose 32 images in Subsets 1 and 2, which were taken under less extreme lighting conditions, as the training data. The remaining images are used as test data. All training and test samples are downsampled to size 32×28.

In the following experiments, our face recognition algorithm comprises the extraction of local sparse features along with graphical model decisions (as described in Section III-B, part 2) which we term as Local-Sparse-GM abbreviated to LSGM. We compare our LSGM technique against five popular face recognition
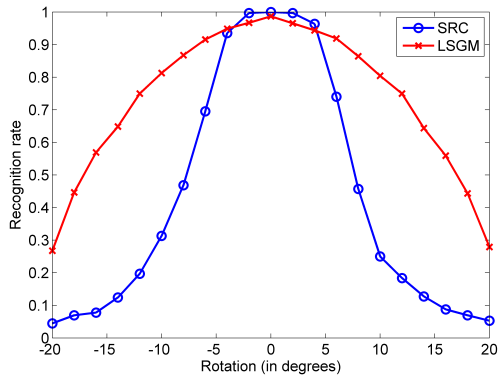
Fig. 5. Recognition rate for rotated test images (Section IV-B).
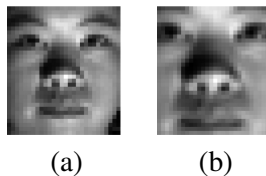

(a)   (b)

Fig. 6. An example of scaled test images. (a) Original image and (b) the image scaled by 1.313 vertically and 1.357 horizontally.

algorithms: (i) sparse representation-based classification (SRC) [9], (ii) Eigenfaces [3] as features with nearest subspace [32] classifier (Eigen-NS), (iii) Eigenfaces with support vector machine [33] classifier (Eigen-SVM), (iv) Fisherfaces [6] as features with nearest subspace classifier (Fisher-NS), and (v) Fisherfaces with SVM classifier (Fisher-SVM). Overall recognition rates - ratio of the total number of correctly classified images to the total number of test images, expressed as a percentage - are reported in Table I. The results reveal that the choice of local sparse features over global features does not significantly affect the overall recognition performance in the scenario of no registration errors.

B. Recognition Under Distortions and Misalignment

1) Presence of registration errors: The primary motivation for our contribution in this paper is to achieve robust recognition under misalignment of test images. We create distorted test images in several ways and keep the training images unchanged, again using images from the Extended Yale B database. Robustness to image translation is ensured by simply choosing an appropriate search region for each block such that the corresponding blocks in the training images are included in the dictionary.

Next, we show experimental results for test images under rotation and scaling operations. In the first set of experiments, the test images are randomly rotated by an angle between -20 and 20 degrees, as illustrated by

TABLE II
RECOGNITION RATE (IN PERCENTAGE) FOR SCALED TEST IMAGES
USING SRC [9] UNDER VARIOUS SCALING FACTORS (SF).

SF	1	1.071	1.143	1.214	1.286	1.357
1	100	100	98.0	88.2	76.5	58.8
1.063	99.7	96.5	86.1	68.5	50.3	37.6
1.125	83.8	70.2	49.8	33.6	26.2	17.9
1.188	54.5	43.7	26.8	20.0	18.0	12.6
1.25	36.1	27.2	20.9	16.6	12.3	11.3
1.313	31.5	24.3	16.7	13.9	10.6	9.8

TABLE III
RECOGNITION RATE (IN PERCENTAGE) FOR SCALED TEST IMAGES
USING PROPOSED BLOCK-BASED APPROACH UNDER VARIOUS SF.

SF	1	1.071	1.143	1.214	1.286	1.357
1	98.8	98.2	98.5	97.5	97.5	97.2
1.063	97.5	96.7	96.0	96.0	93.5	93.4
1.125	97.4	96.5	96.2	95.2	93.2	91.1
1.188	94.9	92.9	91.6	89.4	87.1	83.3
1.25	94.9	93.0	92.2	87.9	82.0	77.8
1.313	90.7	90.4	84.1	81.0	75.5	64.2

the example in Fig. 4. We compare the SRC approach with the proposed LSGM framework. Fig. 5 shows the recognition rate (*y*-axis) for each rotation degree (*x*-axis). We see that when the misalignment becomes more severe, the LSGM algorithm outperforms the SRC approach by a significant margin.

For the second set of experiments, the test images are stretched in both directions by scaling factors up to 1.313 vertically and 1.357 horizontally. An example of an aligned image in the database and its distorted version to be tested are shown in Fig. 6. Tables II and III show the percentage of correct identification with various scaling factors. The first row and the first column in the tables indicate the scaling factors in the horizontal and vertical directions respectively. We again see that when there are large registration errors, the block-based algorithm leads to a better identification performance than the original algorithm. We observe similar behaviors when the scaling factors are in the range of 0.8 to 1 (that is, the test image is shrunk comparing to the training images in the dictionary).

We now compare the performance of our LSGM approach with five other algorithms: SRC, Eigen-NS, Eigen-SVM, Fisher-NS and Fisher-SVM, for the particular scenario where the test images have been scaled by a horizontal factor of 1.214 and a vertical factor of 1.063. The per-face recognition rates are displayed for each approach in Fig. 7, and the overall recognition rates are shown in Table IV.

Next, we repeat the experiment on the Georgia Tech face database [34], wherein the test face captures are
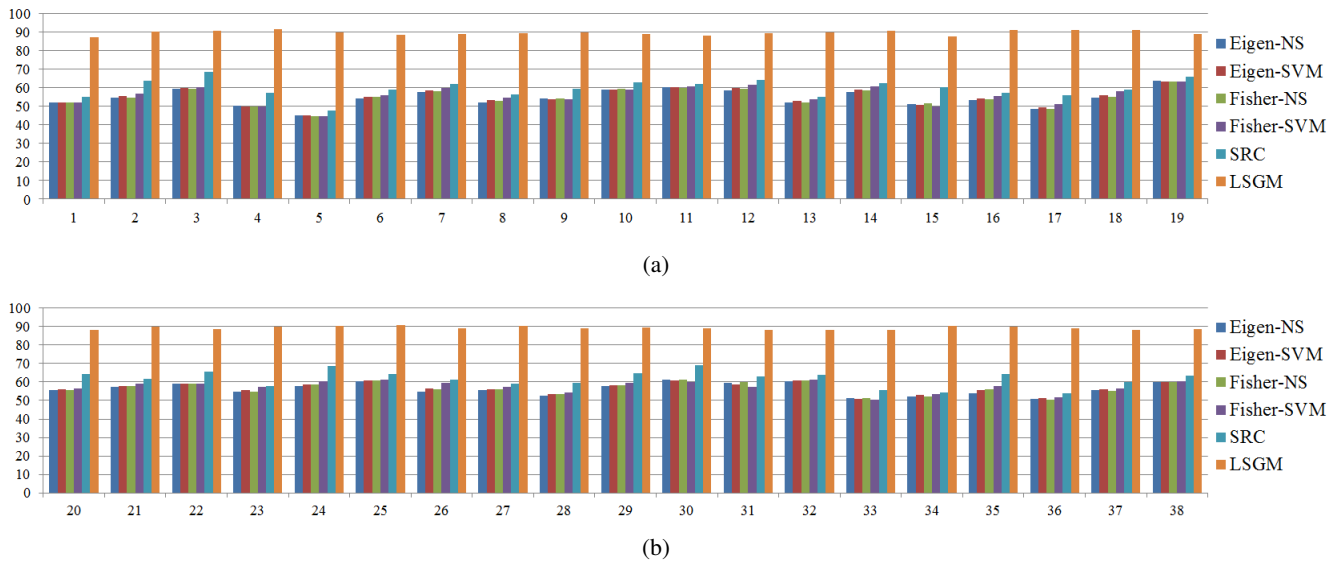
Fig. 7. Face-specific recognition rates using the Extended Yale B database, with registration errors introduced in test images. (a) Results shown for faces numbered 1 through 19. (b) Results shown for faces numbered 20 through 38.

TABLE IV
OVERALL RECOGNITION RATE (AS A PERCENTAGE) FOR THE SCENARIO OF SCALING BY HORIZONTAL AND VERTICAL FACTORS OF 1.214 AND 1.063 RESPECTIVELY.

Method	Recognition rate (%)
LSGM	89.4
SRC	60.8
Eigen-NS	55.5
Eigen-SVM	56.7
Fisher-NS	54.1
Fisher-SVM	57.1

TABLE VI
OVERALL RECOGNITION RATE (AS A PERCENTAGE) FOR THE SCENARIO WHERE TEST IMAGES ARE SCALED AND SUBJECTED TO RANDOM PIXEL CORRUPTION (SECTION IV-B2).

Method	Recognition rate (%)
LSGM	96.3
SRC	93.2
Eigen-NS	54.3
Eigen-SVM	58.5
Fisher-NS	56.2
Fisher-SVM	59.9

TABLE V
OVERALL RECOGNITION RATE (AS A PERCENTAGE) UNDER REGISTRATION ERRORS, FOR IMAGES OBTAINED FROM [34].

Method	Recognition rate (%)
LSGM	87.6
SRC	61.3
Eigen-NS	47.4
Eigen-SVM	50.5
Fisher-NS	45.3
Fisher-SVM	51.8

naturally frontal and/or tilted with different facial expressions, lighting conditions and scale. This database contains 15 faces each of 50 different individuals. For convenience, we restrict the data set to 38 classes of faces (chosen with no particular preference). We use five images from each class for training and the rest for testing. Here too, we provide a comparison of the per-face recognition rates for the LSGM method, and compare it with the five other approaches. The overall rates in Table V confirm once again the robustness of the LSGM to misalignments in test images.

2) Recognition despite random pixel corruption: We return to the Extended Yale database for this experiment, where we randomly corrupt 50% of the image pixels in each test image. In addition, each test image is scaled by a horizontal factor of 1.071 and a vertical factor of 1.063. Local sparse features are extracted using the robust form of the ℓ_1-minimization similar to the approach in [9]. The overall recognition rates are shown in Table VI. These results reveal that under the mild scaling distortion scenario, our LSGM approach retains the robustness characteristic of the global sparsity approach (SRC), while the other competitive algorithms suffer drastic degradation in performance.

3) Recognition despite disguise: We test the robustness of our proposed LSGM approach to disguise (representative of real-life scenarios) using the AR Face Database [35]. We choose a subset of the database containing 50 male and 50 female subjects chosen randomly. For training, we consider 8 clean (with no occlusions) images each per subject. These images may however capture different facial expressions. Faces with two different

TABLE VII
OVERALL RECOGNITION RATE (AS A PERCENTAGE) FOR THE
SCENARIO WHERE TEST IMAGES ARE SCALED AND SUBJECTS
WEAR DISGUISE (SECTION IV-B3).

Method	Recognition rate (%) Sunglasses	Recognition rate (%) Scarves
LSGM	96.0	92.9
SRC	93.5	90.1
Eigen-NS	47.2	29.6
Eigen-SVM	53.5	34.5
Fisher-NS	57.9	41.7
Fisher-SVM	61.7	43.6

Fig. 8. ROC curves for outlier rejection (Section IV-C).

types of disguise are used for testing purposes: subjects wearing sunglasses and subjects partially covering their face with a scarf. Accordingly, we present two sets of results. In each scenario, we use 6 images per subject for testing, leading to a total of 600 test images each for sunglasses and scarves. Consistent with our other experiments, we also introduce mild misalignment in the test images, in the form of scaling by horizontal and vertical factors of 1.071 and 1.063 respectively.

To enable robustness against disguise, in [9] the authors also suggest block partitioning to improve the results, by aggregating results from individual blocks using voting. It is useful to point out two key differences between this strategy and our proposed approach: (i) we use an adaptive local block-based model to build the training dictionary to incorporate robustness to mis-alignment, and (ii) we use a principled classification framework using graphical models to combine results from the individual blocks rather than simple voting.

The results of our proposed approach (using three representative local regions) are compared with five other competitive approaches in Table VII. The LSGM and SRC approaches significantly outperform the other methods. Further, the improvements of LSGM over SRC reveal the benefits of the graphical model framework for classification over the voting scheme. For additional improvements in recognition rate, we can use a larger number of local spatial blocks.

C. Outlier Rejection

In this experiment, samples from 19 of the 38 classes in the Yale database are included in the training set, and faces from the other 19 classes are considered outliers. For training, 15 samples per class from Subsets 1 and 2 are used ($19 \times 15 = 285$ samples in total), while 500 samples are randomly chosen for testing, among which 250 are inliers and the other 250 are outliers. All test samples are rotated by five degrees.

The five different competing approaches are compared with our proposed LSGM method. For the LSGM ap-
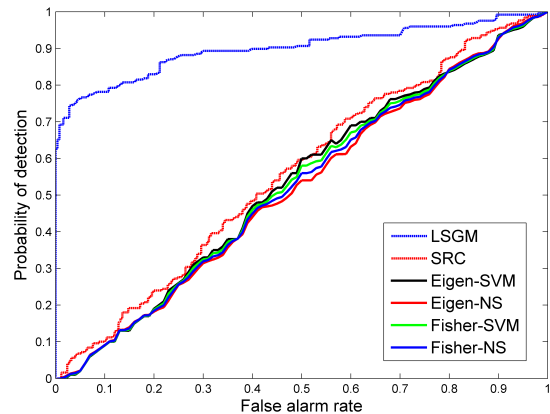
proach, we use a minimum threshold δ on the quantity described in (20). If the maximum value of the log-likelihood ratio does not exceed δ, the corresponding test sample is labeled an outlier. In the SRC approach, the Sparsity Concentration Index (6) is used as the criterion for outlier rejection. For the other approaches under comparison which use the nearest subspace and SVM classifiers, reconstruction residuals are compared to a threshold to decide outlier rejection. The receiver operating characteristic (ROC) curves for all the approaches are shown in Fig. 8, where the probability of detection is the ratio between the number of detected inliers and the total number of inliers, and the false alarm rate is computed by the number of outliers which are detected as inliers divided by the total number of outliers. Under scaling distortion, we see that LSGM offers the best performance, while some of the approaches are actually worse than random guessing.

D. Classifier Fusion: Variants of Proposed Method

We now compare the performance of the different proposed ways of combining the local classifier decisions from Section III-B: (i) majority voting (Voting), (ii) heuristic maximum likelihood (ML)-type fusion using reconstruction residuals (LHML), and (iii) the discriminative graphical model framework (LSGM). The images are taken from the Extended Yale B Database. We introduce mild misalignment in the test images in the form of scaling by a horizontal factor of 1.214 and a vertical factor of 1.063. We use 15 training samples per class, and a total of 1844 samples for testing.

Although the LSGM approach has superior overall recognition performance in comparison to the Voting and LHML techniques, we see from Fig. 10 that for some of the classes, the LHML approach in fact offers slightly better recognition rates. So, we propose a *principled* meta-classification framework to further exploit
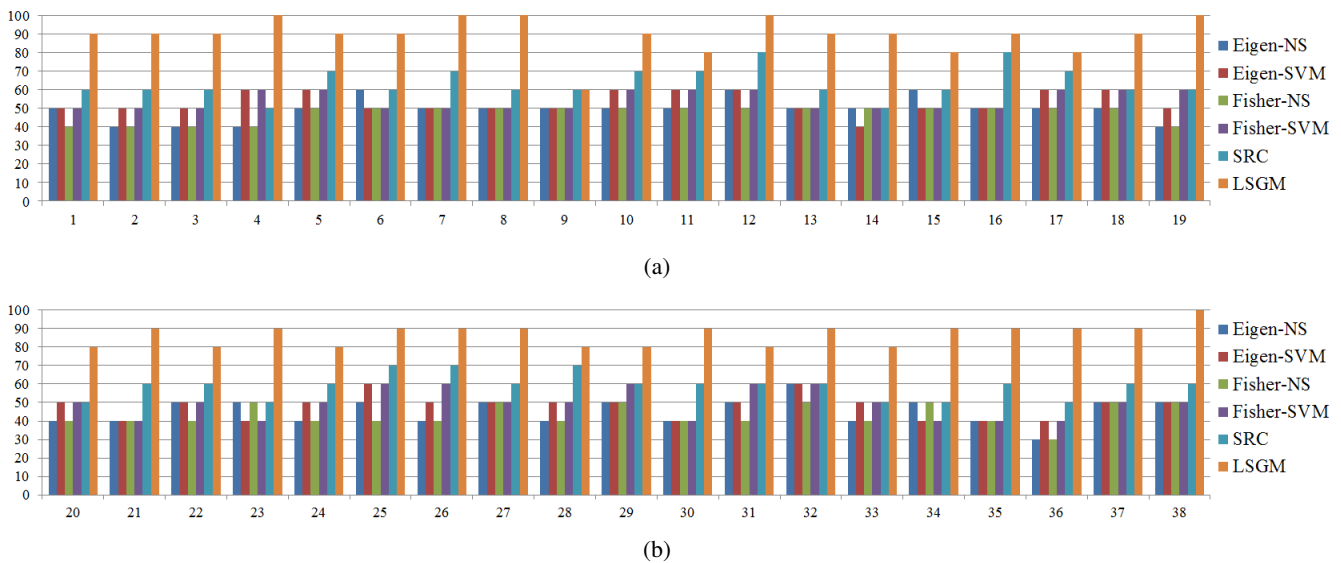
Fig. 9. Face-specific recognition rates using the Georgia Tech face database, with registration errors introduced in test images. (a) Results shown for faces numbered 1 through 19. (b) Results shown for faces numbered 20 through 38.

these complementary benefits offered by the individual classifiers. From each type of classifier, we obtain "soft" outputs, that estimate the posterior probability of a face belonging to a particular class. These soft outputs may also be interpreted as indicating the degree of confidence in the decision. These outputs may then be treated as meta-feature vectors to be fed into a support vector machine for meta-classification. We train the SVM using the soft outputs obtained from the training samples. A radial basis function (RBF) kernel is used in the SVM.

For perfectly calibrated test images, voting presents a computationally simple way of benefitting from the classification results from individual local blocks. However, in the presence of registration errors, voting performs poorly, leading to reduced overall performance of the meta-classifier. So, we present results using only two classifiers: LHML and LSGM. The per-class rates for the individual schemes as well as the meta-classifier are presented in Fig. 10. Meta-classification shows that the complementary benefits of different classifiers can be mined to improve recognition performance.

E. Influence of number of local blocks on recognition performance

So far, we have used three preceptively meaningful local blocks for the LSGM approach, while proposing the use of 42 uniformly sampled blocks of size 8×8 for the LHML method. Unsurprisingly, the presence of more local blocks can improve recognition by offering more robustness to distortions. So, in this section, we evaluate the performance of our proposed algorithms as a function of number of blocks. Specifically, we use 3, 5,
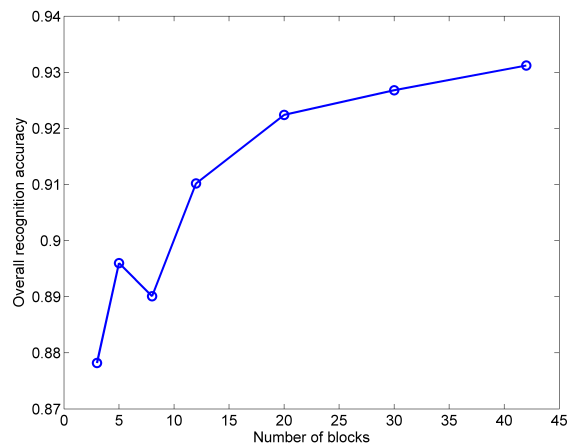

Fig. 11. Recognition performance of LSGM as a function of number of local blocks. Experiments are performed on the Georgia Tech database.

8, 12, 20, 30 or 42 blocks in different experiments. For the case of 5 blocks, we pick the five (perceptually most meaningful) regions to be the block of two eyes, nose, mouth, and the two eyes taken individually. For larger number of blocks, the blocks are chosen uniformly from the entire image and the size of the blocks is either 8×12 or 8×8.

We choose two specific experiments to illustrate the dependence on number of blocks. First, we consider images from the Georgia Tech database, where the test images are naturally misaligned (Section IV-B1). The performance of the LSGM approach is shown in Fig. 11. There is a dip in recognition performance for the case of 8 blocks compared to the case of 5 blocks, since the 8
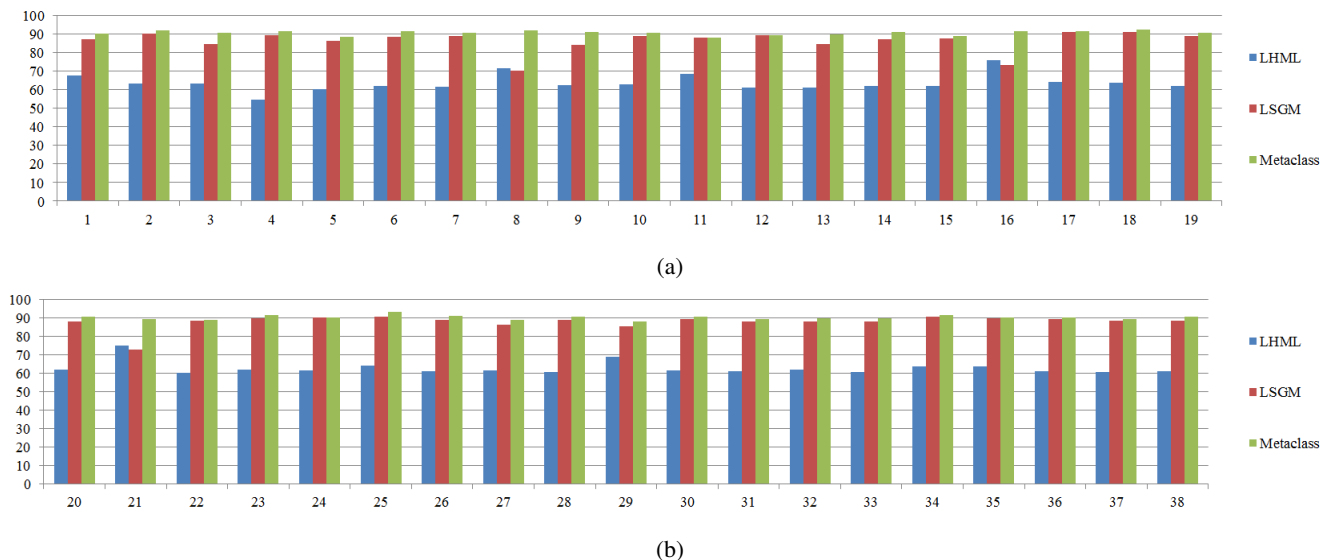
Fig. 10. Meta-classification: Face-specific recognition rates using the Extended Yale B face database, with scaling registration errors introduced in test images. (a) Results shown for faces numbered 1 through 19. (b) Results shown for faces numbered 20 through 38.
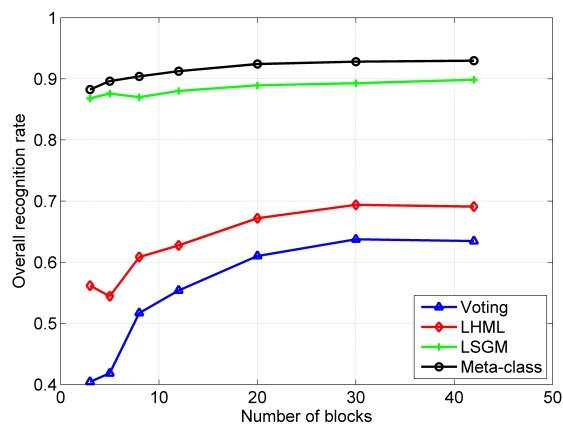

Fig. 12. Recognition performance of proposed classifiers and meta-classifier, as a function of number of local blocks.

blocks are chosen uniformly from the image and need not necessarily carry perceptual meaning, while the 5 blocks are chosen in a particular meaningful manner. However, with the increase in the number of blocks, the particular choice of blocks seemingly becomes less relevant.

For the second experiment, we consider the meta-classification scenario described earlier in this section. The resulting plot is showed in Fig. 12. The voting approach performs very poorly in comparison with the LHML and LSGM approaches. As expected, the meta-classifier improves upon the performance of all the methods. More significantly, Fig. 12 reveals that the LSGM approach is less sensitive to variations in the number of blocks and particular choice of blocks, while the performance of other proposed local approaches is

contingent on the availability of sufficient number of local blocks.

V. CONCLUSION

We developed a local block-based sparsity model to realize a practical face recognition algorithm which exhibits robustness to alignment errors and a host of distortions such as noise, occlusion, disguise and illumination changes. Unlike other competing techniques, no explicit registration step is required - which makes our approach computationally simpler. Inspired by human perception, our sparse local features are extracted via projections onto adaptive dictionaries built from informative regions of the face image such as eyes, nose and mouth. Instead of using class specific reconstruction error (which does not capture inter-class variation), we present a probabilistic graphical model framework to explicitly capture the conditional correlations between these sets of local features. Experiments on benchmark face databases and comparisons against state-of-the-art face recognition techniques under numerous practical testing environments reveal the merits of our proposal.

REFERENCES

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, Dec. 2003.

[2] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Optical Soc. of Am. A*, vol. 4, no. 3, pp. 519–524, Mar. 1987.

[3] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, Winter 1991.

[4] J. Zou, Q. Ji, and G. Nagy, "A comparative study of local matching approach for face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 10, pp. 2617–2628, Oct. 2007.

[5] A. Shashua, "Geometry and photometry in 3D visual recognition," Ph.D. dissertation, MIT, 1992.

[6] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[7] C. Liu and H. Wechsler, "A shape- and texture-based enhanced fisher classifier for face recognition," *IEEE Trans. Image Process.*, vol. 10, no. 4, pp. 598–608, Apr. 2001.

[8] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 2, pp. 218–233, Feb. 2003.

[9] J. Wright, A. Y. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[10] J. K. Pillai, V. M. Patel, R. Chellappa, and N. Ratha, "Secure and robust iris recognition using sparse representations and random projections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1877–1893, Sep. 2011.

[11] X. Hang and F.-X. Wu, "Sparse representation for classification of tumors using gene expression data," *Journal of Biomedicine and Biotechnology*, vol. 2009, 2009, doi:10.1155/2009/403689.

[12] J. Huang, X. Huang, and D. Metaxas, "Simultaneous image transformation and sparse representation recovery," in *Proc. of IEEE Conf. Comput. Vision Pattern Recognition*, Anchorage, AK, Jun. 2008, pp. 1–8.

[13] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma, "Towards a practical face recognition system: Robust registration and illumination by sparse representation," in *Proc. of IEEE Conf. Comput. Vision Pattern Recognition*, Miami, FL, Jun. 2009, pp. 597–604.

[14] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Towards a practical face recognition system: Robust alignment and illumination by sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, to appear.

[15] Y. Chen, T. T. Do, and T. D. Tran, "Robust face recognition using locally adaptive sparse representation," in *Proc. IEEE Intl. Conf. Image Processing*, Hong Kong, Sep. 2011, pp. 1657–1660.

[16] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[17] U. Srinivas, V. Monga, Y. Chen, and T. D. Tran, "Sparsity-based face recognition using discriminative graphical models," in *Proc. IEEE Asilomar Conf. on Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2011.

[18] V. Y. F. Tan, S. Sanghavi, J. W. F. III, and A. S. Willsky, "Learning graphical models for hypothesis testing and classification," *IEEE Trans. Signal Processing*, vol. 58, no. 11, pp. 5481–5495, Nov. 2010.

[19] Y. Freund and R. E. Schapire, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 14, no. 5, pp. 771–780, Sep. 1999.

[20] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[21] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[22] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2230–2249, May 2009.

[23] T. T. Do, L. Gan, N. H. Nguyen, and T. D. Tran, "Sparsity adaptive matching pursuit algorithm for practical compressed sensing," in *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, CA, Oct. 2008, pp. 581–587.

[24] S. L. Lauritzen, *Graphical Models*. Oxford University Press, NY, 1996.

[25] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[26] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 462–467, Mar. 1968.

[27] T. T. Do, Y. Chen, D. T. Nguyen, N. H. Nguyen, L. Gan, and T. D. Tran, "Distributed compressed video sensing," in *Proc. IEEE Int. Conf. Image Process.*, Cairo, Egypt, Nov. 2009, pp. 1393–1396.

[28] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[29] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 643–660, Jun. 2001.

[30] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, pp. 131–163, Nov. 1997.

[31] U. Srinivas, V. Monga, and R. G. Raj, "Automatic target recognition using discriminative graphical models," in *Proc. IEEE Intl. Conf. Image Processing*, Brussels, Belgium, Sep. 2011, pp. 33–36.

[32] J. Ho, M. Yang, J. Lim, K. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *Proc. of IEEE Conf. Comput. Vision Pattern Recognition*, Madison, WI, Jun. 2003, pp. 11–18.

[33] V. N. Vapnik, *The nature of statistical learning theory*. New York, USA: Springer, 1995.

[34] "*The Georgia Tech Face Database*." [Online]. Available: http://www.anefian.com/research/face_reco.htm

[35] A. M. Martinez and R. Benavente, "The AR face database," *CVC Tech. Report*, no. 24, 1998.