Sparse Representations of Image Gradient Orientations for Visual Recognition and Tracking

Georgios Tzimiropoulos[†] Stefanos Zafeiriou[†]
[†]Dept. of Computing, Imperial College London
180 Queen's Gate, London SW7 2AZ, U.K.
{gt204,s.zafeiriou,m.pantic}@imperial.ac.uk

Maja Pantic [†,*]
[*]EEMCS, University of Twente
Drienerlolaan 5, 7522 NB Enschede,
The Netherlands
PanticM@cs.utwente.nl [*]

Abstract

Recent results [18] have shown that sparse linear representations of a query object with respect to an overcomplete basis formed by the entire gallery of objects of interest can result in powerful image-based object recognition schemes. In this paper, we propose a framework for visual recognition and tracking based on sparse representations of image gradient orientations. We show that minimal ℓ_1 solutions to problems formulated with gradient orientations can be used for fast and robust object recognition even for probe objects corrupted by outliers. These solutions are obtained without the need for solving the extended problem considered in [18]. We further show that low-dimensional embeddings generated from gradient orientations perform equally well even when probe objects are corrupted by outliers, which, in turn, results in huge computational savings. We demonstrate experimentally that, compared to the baseline method in [18], our formulation results in better recognition rates without the need for block processing and even with smaller number of training samples. Finally, based on our results, we also propose a robust and efficient ℓ_1-based "tracking by detection" algorithm. We show experimentally that our tracker outperforms a recently proposed ℓ_1-based tracking algorithm in terms of robustness, accuracy and speed.

1. Introduction

A recent breakthrough in image-based object recognition [18] as well as subsequent work [17, 19, 20] have conclusively shown that this problem can be re-cast as one of finding the sparsest representation of a probe object with respect to an overcomplete dictionary whose elements are the objects in the training set. Given that sufficient training samples are available from each class, such an approach has demonstrated excellent performance for the problem of frontal view face recognition with illumination changes and occlusions in the testing set, while follow-up paper successfully addressed the problems of joint alignment/recognition [16] as well as visual tracking [12].

The basic principles and assumptions of the method for the application of face recognition are as follows. We assume that all images in the training and testing set are aligned images of the same resolution $m = d_1 \times d_2$ written in lexicographic ordering. We form the data matrix $\mathbf{A} = [\mathbf{a}_1|\cdots|\mathbf{a}_n] \in \Re^{m\times n}$, where n is the total number of training samples. We further assume that $m \ll n$. To classify a probe face $\mathbf{y} \in \Re^m$, we look for $\mathbf{x} \in \Re^n$ which solves the following ℓ_1 minimization problem

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1, \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x}. \quad (1)$$

A number of now classical papers [2,3,5,6] have established that, under a number of assumptions, among all $\mathbf{x} : \mathbf{y} = \mathbf{A}\mathbf{x}$, the minimal ℓ_1 solution $\mathbf{x}_o$ is also the sparsest one. This can be used for classification as the largest non-zero coefficients of $\mathbf{x}_o$ indicate the subject's identity.

To deal with small dense noise, the equality constraint in (1) is typically replaced by the inequality $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \epsilon$. For robust face recognition, however, we are particularly interested in the case where a fraction of pixels in the test image is arbitrarily corrupted. That is, we are interested in solving $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$, where $\mathbf{e} \in \Re^m$ is an unknown vector whose nonzero entries correspond to the set of corrupted pixels. To cope with this type of noise, we form the extended data matrix $[\mathbf{A}\ \mathbf{I}] \in \Re^{m\times(n+m)}$, where $\mathbf{I}$ is the identity matrix and look for the sparsest solution $[\mathbf{x}_o^T\ \mathbf{e}_o^T]^T \in \Re^{m+n}$ by minimizing

$$\min_{\mathbf{x},\mathbf{e}} \|\mathbf{x}\|_1 + \|\mathbf{e}\|_1, \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}. \quad (2)$$

The above outlier-robust formulation has been successfully applied for the problems of occlusion-robust face

[*]This work has been funded funded by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB)."

recognition [18] and visual tracking [12]. However, this formulation is applicable only in the pixel domain and achieves striking performance in the presence of real occlusions only when it is applied in a block-based fashion. Both issues, in turn, pose serious computational concerns since the method cannot be combined with dimensionality reduction techniques, while block processing inevitably increases the execution time. Thus, practical implementations of the method are typically based on image down-sampling.

In this paper we show how to efficiently address the above issues. While experimentation reported in [18] suggests that, at least for the outlier-free case, the choice of features (via dimensionality reduction) is of minor importance, we show that the object representation used is highly critical. Motivated by some very recent results on image registration and subspace learning [14, 15], we propose a framework for object recognition based on sparse representations of image gradient orientations. The key idea is to replace pixel intensities with gradient orientations and then define a mapping from the space of gradient orientations into a high-dimensional unit sphere. As illustrated in [14, 15], the key observation is that, in contrast to pixel intensities, representations of this type, when obtained from "visually unrelated" images, are highly incoherent. We show that

1. Minimal ℓ_1 solutions to problems formulated with gradient orientations can be used for fast and robust object recognition even for probe objects corrupted by outliers. These solutions are obtained *without* the need for solving the extended problem in (2).

2. Low-dimensional embeddings generated from gradient orientations perform equally well even when probe objects are corrupted by outliers. This results in huge computational savings. For example, in a Core 2 Duo machine with 8 GB RAM, for a dictionary of $n = 400$ training samples, our scheme requires less than 0.5 seconds to classify of a probe image with $p = 100$ features. For the same setting, the original formulation of [18] with 64×64 images requires about 1 minute.

3. Sparse representations of gradient orientations result in better recognition rates *without* the need for block processing and with smaller number of training samples.

Finally, we show how to capitalize on the above results for robust and efficient visual tracking. We propose a tracking algorithm, which although it is also based on ℓ_1 minimization, it is conceptually very different compared to the approach proposed in [12]. In contrast to [12], we use ℓ_1 minimization as a discriminant classifier which separates the object from the background and as such our algorithm is closely related to methods which perform "tracking by detection" [1, 4, 7]. Additionally, as opposed to [12], the proposed tracker is based on sparse representations of image gradient orientations and thus does not rely on the extended problem of (2) to achieve robustness to outliers. This, in turn, results in huge computational savings as we perform tracking in a low dimensional subspace. We show experimentally that our tracking algorithm outperforms the method of [12] in terms of robustness, accuracy and speed.

2. Image Gradient Orientations and Incoherence

Assume that we are given the image-based representations of two objects $\mathbf{I}_i \in \Re^{d_1 \times d_2}$, $i = 1, 2$. At each pixel location, we estimate the image gradients and the corresponding gradient orientation. More specifically, we compute

$$\mathbf{\Phi}_i = \arctan \mathbf{G}_{i,y} / \mathbf{G}_{i,x}, \tag{3}$$

where $\mathbf{G}_{i,x} = \mathbf{F}_x \star \mathbf{I}_i$, $\mathbf{G}_{i,y} = \mathbf{F}_y \star \mathbf{I}_i$ and $\mathbf{F}_x, \mathbf{F}_y$ are the filters used to obtain the image gradients along the horizontal and vertical direction, respectively. Let us denote by $\boldsymbol{\phi}_i$ the $m-$dimensional vector obtained by writing $\mathbf{\Phi}_i$ in lexicographic ordering.

We have difficulty using vectors $\boldsymbol{\phi} \in [0, 2\pi)^m$ directly in optimization problem (1). Clearly, we can neither write such a vector as a linear combination of a dictionary of angle vectors nor use the ℓ_2 norm for measuring the reconstruction error. To use angular data, we use the following mapping onto the $\Re^{2m}$ sphere

$$\mathbf{z}(\boldsymbol{\phi}_i) = \frac{1}{\sqrt{m}} [\cos(\boldsymbol{\phi}_i)^T \sin(\boldsymbol{\phi}_i)^T]^T, \tag{4}$$

where $\cos(\boldsymbol{\phi}_i) = [\cos(\boldsymbol{\phi}_i(1)), \ldots, \cos(\boldsymbol{\phi}_i(m))]^T$ and $\sin(\boldsymbol{\phi}_i) = [\sin(\boldsymbol{\phi}_i(1)), \ldots, \sin(\boldsymbol{\phi}_i(m))]^T$. Using $\mathbf{z}_i \equiv \mathbf{z}(\boldsymbol{\phi}_i)$, we naturally measure correlation from

$$c(\mathbf{z}_1, \mathbf{z}_2) \triangleq \mathbf{z}_1^T \mathbf{z}_2 = \frac{1}{m} \sum_{k=1}^{m} \cos[\Delta \boldsymbol{\phi}(k)], \tag{5}$$

where $\Delta \boldsymbol{\phi} \triangleq \boldsymbol{\phi}_1 - \boldsymbol{\phi}_2$. The distance between $\mathbf{z}_1$ and $\mathbf{z}_2$ is

$$\begin{aligned} d(\mathbf{z}_1, \mathbf{z}_2) &\triangleq \frac{1}{2} ||\mathbf{z}_1 - \mathbf{z}_2||_2 \\ &= \frac{1}{2} (\mathbf{z}_1^T \mathbf{z}_1 - 2\mathbf{z}_1^T \mathbf{z}_2 + \mathbf{z}_2^T \mathbf{z}_2) \\ &= 1 - \frac{1}{m} \sum_{k=1}^{m} \cos[\Delta \boldsymbol{\phi}(k)], \end{aligned} \tag{6}$$

which is simply the chord between $\mathbf{z}_1$ and $\mathbf{z}_2$. From (5) and (6), we observe that if $\mathbf{I}_1 \simeq \mathbf{I}_2$, then $\forall k \ \Delta \boldsymbol{\phi}(k) \simeq 0$, and therefore $c \to 1$ and $d \to 0$.

Let us assume now that the two images are "visually unrelated" (or dissimilar) so that locally do not match. Then,

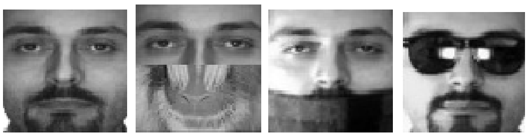


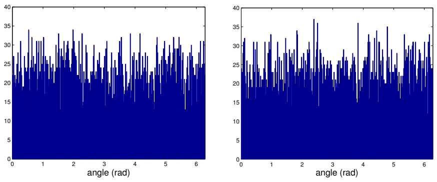
Figure 1. (a) An example of training samples considered in our experiments. (b)-(d) Three examples of testing samples.

it is not unreasonable to assume that for any spatial location k, the difference in gradient orientation $\Delta\phi(k)$ can take any value in the range $[0, 2\pi)$ with equal probability. Thus, we assume that $\Delta\boldsymbol{\phi}$ is a realization of a stationary random process $u(t)$ which $\forall t$ follows a uniform distribution $U(0, 2\pi)$ [14, 15]. Given this, it is not difficult to show that, under some rather mild assumptions, it holds [14, 15]

$$\sum_{k=1}^{m} \cos[\Delta\phi(k)] \simeq 0, \tag{7}$$

and therefore $c \to 0$ and $d \to 1$. Thus, by using (5) as a measure of coherence, "visually unrelated" images are approximately incoherent.

As an example, we consider three examples of "visually unrelated" image patches. We assume that the face region in Fig. 1 (a) and the "baboon" patch in Fig. 1 (b) are visually dissimilar. Similarly, the face region in Fig. 1 (a) is "visually unrelated" with the image regions corresponding to the scarf and the glasses in Fig. 1 (c) and (d) respectively. Fig. 2 (a) shows the distribution of $\Delta\boldsymbol{\phi}$ for the scarf case, while Fig. 2 (b) shows the distribution of uniformly distributed samples drawn from Matlab's random number generator. In the following section, we show how to exploit this incoherence property for fast and robust object recognition.



(a) (b)

Figure 2. (a) The distribution of $\Delta\boldsymbol{\phi}$ for the face region in Fig. 1 (a) and the region corresponding to the scarf in Fig. 1 (c). (b) The distribution of samples (uniformly distributed) drawn from Matlab's random number generator.

3. Recognition with Sparse Representations of Image Gradient Orientations

Given a set of n training samples of m pixels, we obtain our dictionary as follows. We compute the orientation images $\mathbf{\Phi}_i$, $i = 1, \ldots, n$ from (3), obtain $\boldsymbol{\phi}_i$ by writing $\mathbf{\Phi}_i$ in lexicographic ordering, compute $\mathbf{z}_i$ from (4) and form the matrix $\mathbf{Z} = [\mathbf{z}_1 | \cdots | \mathbf{z}_n] \in \Re^{2m \times n}$. Given a probe object $\mathbf{y}$, we follow the same procedure for $\mathbf{q} \in \Re^{2m}$. Next, we solve

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1, \text{ subject to } \|\mathbf{q} - \mathbf{Z}\mathbf{x}\|_2 \le \epsilon. \tag{8}$$

Within our framework of image gradient orientations, the solution to typical ℓ_1 minimization problems (such as the one in (8)) for probe objects corrupted by outliers can be efficiently used for fast and robust object recognition, *without* the need of formulating and solving the equivalent of the extended problem of (2).

To show this, we start by noting that our aim is object classification and not precise object reconstruction. We assume K object classes with L_k objects per class. As in [18], given a probe object $\mathbf{q}$ and the solution $\mathbf{x}^*$, we perform classification as follows. For any class k, we form $\mathbf{c}_k \in \Re^n$, where $\mathbf{c}_{k,l} = \mathbf{x}_l^*$ for all indices l corresponding to class k and $\mathbf{c}_{k,l} = 0$ otherwise. We reconstruct from $\tilde{\mathbf{q}}_k = \mathbf{Z}\mathbf{c}_k$ and classify using the minimum of the reconstruction error

$$\text{identity}(\mathbf{q}) = \min_k \|\mathbf{q} - \tilde{\mathbf{q}}_k\|_2. \tag{9}$$

Similarly to [18], we assume that the training samples of each subject do span a subspace. Therefore, we can write each probe object belonging to the k^{th} class as

$$\mathbf{q} = \sum_{l=1}^{L_k} w_{k,l} \mathbf{z}_{k,l}, \tag{10}$$

where $w_{k,l} \in \Re$ and $\mathbf{z}_{k,l} \in \Re^{2m}$, $l = 1, \ldots, L_k$ are the weights and bases corresponding to the L_k samples of the k^{th} class. Without loss of generality, we assume that $\mathbf{z}_{k,l}$ are the eigenvectors obtained from the eigen-analysis of the k^{th} class' covariance matrix, sorted in decreasing order according to their eigenvalues. We retrieve $w_{k,l}$ from the non-zero elements of the solution $\mathbf{x}^*$ of the following problem

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1, \text{ subject to } \mathbf{q} = \mathbf{Z}\mathbf{x}, \tag{11}$$

that is $\mathbf{x}^* = [0, \ldots, 0, w_{1,1}, \ldots, w_{1,L_k}, 0, \ldots, 0]^T \in \Re^n$.

Let us now assume that a part of the probe object is corrupted by outliers. We will assume that this part of the object and the corresponding parts of all objects in our dictionary are "visually unrelated". According to the previous section, for any object in our dictionary, we have

$$\begin{aligned} \mathbf{q}^T \mathbf{z}_i &= \frac{m_u}{m} \mathbf{q}_u^T \mathbf{z}_{u,i} + \frac{m - m_u}{m} \mathbf{q}_o^T \mathbf{z}_{o,i} \\ &\approx \frac{m_u}{m} \mathbf{q}_u^T \mathbf{z}_{u,i}, \end{aligned} \tag{12}$$

where $\mathbf{q}_u, \mathbf{z}_{u,i} \in \Re^{2m_u}$ and $\mathbf{q}_o, \mathbf{z}_{o,i} \in \Re^{2(m-m_u)}$ are the object's parts corresponding to the un-occluded and occluded regions respectively, m_u is the number of pixels in the un-occluded region and $\mathbf{q}_o^T \mathbf{z}_{o,i} \approx 0$ according to (7). We also have $\mathbf{q}^T \mathbf{Z} \approx \frac{m_u}{m} \mathbf{q}_u^T \mathbf{Z}_u$ and write

$$\mathbf{Z}^T \mathbf{Z} = \frac{m_u}{m} \mathbf{Z}_u^T \mathbf{Z}_u + \frac{m - m_u}{m} \mathbf{Z}_o^T \mathbf{Z}_o, \qquad (13)$$

where $\mathbf{Z}_u = [\mathbf{z}_{u,1}| \ldots |\mathbf{z}_{u,n}]$ and $\mathbf{Z}_o = [\mathbf{z}_{o,1}| \ldots |\mathbf{z}_{o,n}]$.

Since a perfect probe object reconstruction is infeasible, we consider the equivalent form of (8) given by [10]

$$\min_{\mathbf{x}} ||\mathbf{q} - \mathbf{Z}\mathbf{x}||_2 + \lambda ||\mathbf{x}||_1. \qquad (14)$$

The question of interest now is the following: How well the solution of (14) approximates the solution of (11)? Using the above, we can write

$$\begin{aligned} ||\mathbf{q} - \mathbf{Z}\mathbf{x}||_2 &= \mathbf{q}^T \mathbf{q} - 2\mathbf{q}^T \mathbf{Z}\mathbf{x} + \mathbf{x}^T \mathbf{Z}^T \mathbf{Z}\mathbf{x} \\ &= 1 - 2\frac{m_u}{m} \mathbf{q}_u^T \mathbf{Z}_u \mathbf{x} + \frac{m_u}{m} \mathbf{x}^T \mathbf{Z}_u^T \mathbf{Z}_u \mathbf{x} \\ &+ \frac{m - m_u}{m} \mathbf{x}^T \mathbf{Z}_o^T \mathbf{Z}_o \mathbf{x} \\ &= \frac{m_u}{m} ||\mathbf{q}_u - \mathbf{Z}_u \mathbf{x}||_2 + \frac{m - m_u}{m} ||\mathbf{Z}_o \mathbf{x}||_2 \\ &+ \frac{m - m_u}{m} \qquad (15) \end{aligned}$$

Now, since $\mathbf{q}_u = \sum_{l=1}^{L_k} w_{k,l} \mathbf{z}_{u,k,l}$, where $\mathbf{z}_{u,k,l}$ are the un-occluded parts of the bases $\mathbf{z}_{k,l}$, it is trivial to see that $\mathbf{x}^*$ is the only possible $\mathbf{x}$ such that $||\mathbf{q}_u - \mathbf{Z}_u \mathbf{x}||_2 = 0$. Therefore, any other $\mathbf{x}$ can minimize (14) if and only if

$$\frac{m_u}{m} ||\mathbf{q}_u - \mathbf{Z}_u \mathbf{x}||_2 + \frac{m - m_u}{m} ||\mathbf{Z}_o \mathbf{x}||_2 < \frac{m - m_u}{m} ||\mathbf{Z}_o \mathbf{x}^*||_2. \qquad (16)$$

While it is not unlikely that there exists $\mathbf{x}$ such that (16) is satisfied, it is highly unlikely that the elements of $\mathbf{x}$ corresponding to the most dominant eigenvectors $\mathbf{z}_{k,l}$ are zero. This is as in this case a large increase in the first term of the right hand side of (16) will be incurred. Finally, this further suggests that the reconstruction error of (9) for the correct class is very likely to be the minimum one. Therefore, robust classification can be performed without formulating and solving the equivalent of the extended problem of (2).

As an example, we considered a training set of $K = 100$ subjects with one sample per class ($L_k = 1, \forall k$) taken from the AR database. Fig. 1 (a) shows an example of the training images. We then considered three different testing sets. For the first set, we directly obtained the testing samples from the training samples after placing a baboon patch which occluded approximately 60% of the original image. Fig. 1 (b) shows an example of the testing samples. For the second and third testing set, the testing samples are faces occluded by a scarf and glasses respectively. Figs. 1 (c) and (d) show examples of the testing images in these cases.

Note that, for this single-sample-per-class experiment, we can assume that the single training sample of each class does span a subspace only for artificially created testing sets such as our first testing set prior to the corruption induced by the baboon patch. Fig. 3 (a)-(c) show the minimal ℓ_1 solutions of (14) (i.e. with a dictionary built from gradient orientations) for the training samples in Fig. 1 (b)-(d) respectively. For *all* cases, the solution is sparse and corresponds to the correct class. On the other hand, Fig. 4 (a)-(c) shows the ℓ_1 solution (more specifically, the first $K = 100$ elements of the solution) obtained by solving the extended problem of (2) (i.e. with a dictionary built from pixel intensities). As we may observe, the solution is sparse, nevertheless, for the case of Fig. 1 (b) and (c), this solution does not indicate the subject's identity. This suggests that sparsity for these cases are mainly due to the inclusion of the identity matrix. This is further illustrated by measuring the efficacy of the solution using the sparsity concentration index [18]. This index takes values in $[0, 1]$ with large values indicating that the probe object is represented by samples of the correct subject only. Table 1 summarizes our results. As we may observe, for pixel intensities, the sparsity concentration index is large only when the testing samples are obtained directly from the training samples (Testing set 1).

	Image Gradient Orientations	Pixel Intensities
Testing set 1	0.160	0.495
Testing set 2	0.164	0.013
Testing set 3	0.158	0.117

Table 1. Average sparsity concentration index for the three testing sets considered in our experiment.

We can derive similar results for low-dimensional embeddings generated from gradient orientations. Let us reformulate (14) in a low dimensional subspace as follows

$$\min_{\mathbf{x}} ||\tilde{\mathbf{q}} - \tilde{\mathbf{Z}}\mathbf{x}||_2 + \lambda ||\mathbf{x}||_1, \qquad (17)$$

where $\tilde{\mathbf{q}} = \mathbf{B}^T \mathbf{q} \in \Re^p$ and $\tilde{\mathbf{Z}} = \mathbf{B}^T \mathbf{Z} \in \Re^{p \times n}$ and $\mathbf{B} \in \Re^{2m \times p}$ are the projection bases.

For most subspace learning methods of interest (etc. PCA, LDA), we can write the projection bases as a linear combination of the data, $\mathbf{B} = \mathbf{Z}\mathbf{V}$, where $\mathbf{V} \in \Re^{n \times p}$. Using this last equation and based on (12) and (15), we have

$$\begin{aligned} ||\tilde{\mathbf{q}} - \tilde{\mathbf{Z}}\mathbf{x}||_2 &= (\mathbf{q} - \mathbf{Z}\mathbf{x})^T \mathbf{Z}\mathbf{V}\mathbf{V}^T \mathbf{Z}^T (\mathbf{q} - \mathbf{Z}\mathbf{x}) \\ &= \left(\frac{m_u}{m}\right)^2 ||\mathbf{V}^T \mathbf{Z}_u^T (\mathbf{q}_u - \mathbf{Z}_u \mathbf{x})||_2 \\ &+ \left(\frac{m - m_u}{m}\right)^2 ||\mathbf{V}^T \mathbf{Z}_o^T \mathbf{Z}_o \mathbf{x}||_2 + \delta, \end{aligned} \qquad (18)$$

where

$$\begin{aligned} \delta &= 2\frac{m_u(m - m_u)}{m^2} \mathbf{x}^T \mathbf{Z}_o^T \mathbf{Z}_o \mathbf{V}\mathbf{V}^T \mathbf{Z}_u^T (\mathbf{q}_u - \mathbf{Z}_u \mathbf{x}) \\ &= 2\frac{m_u(m - m_u)}{m^2} [\mathbf{B}^T \mathbf{Z}_o \mathbf{x}]^T [\mathbf{B}^T (\mathbf{q}_u - \mathbf{Z}_u \mathbf{x})]. (19) \end{aligned}$$

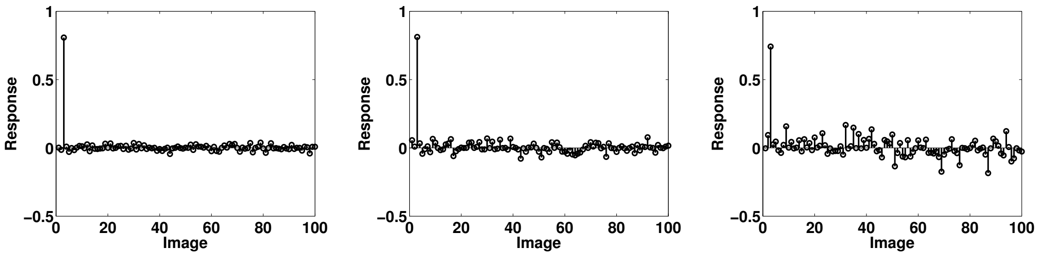


Figure 3. Sparse representations of image gradient orientations. (a)-(c) The minimal ℓ_1 solutions obtained by solving (14) for the probe images of Fig. 1 (b)-(d).

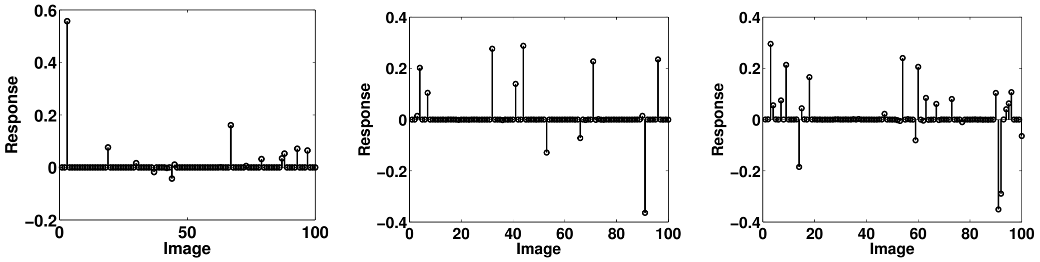


Figure 4. a)-(c) The minimal ℓ_1 solutions (more specifically, the first $K = 100$ elements of the solution) obtained by solving the extended problem (2) for the probe images of Fig. 1 (b)-(d).

Clearly, δ represents the correlation between the embedding of $\mathbf{Z}_o\mathbf{x}$ and $\mathbf{q}_u - \mathbf{Z}_u\mathbf{x}$. Since these terms correspond to different object parts, we can assume that this correlation is negligible, so in a similar fashion to (16), $\mathbf{x}$ will be a minimizer of (17) other than $\mathbf{x}^*$ if and only if

$$\left(\frac{m_u}{m-m_u}\right)^2 ||\mathbf{V}^T\mathbf{Z}_u^T(\mathbf{q}_u - \mathbf{Z}_u\mathbf{x})||_2 + ||\mathbf{V}^T\mathbf{Z}_o^T\mathbf{Z}_o\mathbf{x}||_2 < ||\mathbf{V}^T\mathbf{Z}_o^T\mathbf{Z}_o\mathbf{x}^*||_2. \qquad (20)$$

Overall, the minimal ℓ_1 solution of (17) can be efficiently used for robust object recognition. Additionally, since this solution is obtained without the need for solving the extended problem of (2) and, typically, $p \ll 2m$ our formulation is far more computationally efficient. For example, in a Core 2 Duo machine with 8 GB RAM, for a dictionary of $n = 400$ training samples, classification of a probe image with $p = 100$ features requires about 0.3 seconds. For the same setting, the original formulation of [18] with 64×64 images requires about 1 minute.

4. Visual Tracking with Sparse Representations of Image Gradient Orientations

Visual tracking aims at finding the position of a predefined target object at each frame of a given video sequence. Most existing methods are capable of tracking an object in well-controlled environments. However, tracking in unconstrained environments is still an unsolved problem. For example, in real-word face analysis, appearance changes caused by illumination changes, occlusions, non-rigid deformations, abrupt head movements, and pose variations make most methods fail. In this section, we show how to capitalize on the results of the previous section for robust and efficient visual tracking.

4.1. Related Work

Our algorithm is somewhat related to the approach of [12], where the authors proposed to reformulate visual tracking as a sparse approximation problem. The basic principles and assumptions of this method are as follows. At time instance t, an affine motion model A_t and a particle filter is used to generate a set of target candidates $\mathbf{y}_i$ [8,13]. Let us assume that the columns of the data matrix $\mathbf{A}$ span a linear subspace which models the appearance of the object to be tracked. Then, it is assumed that a target candidate $\mathbf{y}_i$ models the appearance of the tracked object at time t successfully, if it can be written as a linear combination of the bases in $\mathbf{A}$ and few elements of the identity matrix $\mathbf{I}$. This leads to a sparse coefficient vector [12]. On the contrary, a target candidate which models the object appearance poorly will result in a dense representation. As in [18], the columns of $\mathbf{I}$ are used to compensate for possible deviations from the subspace $\mathbf{A}$, for example, due to possible occlusions. Let us denote by $[\mathbf{x}_{o,i}^T \; \mathbf{e}_{o,i}^T]^T$ the solution to

$$\min_{\mathbf{x}_i,\mathbf{e}_i} \|\mathbf{x}_i\|_1 + \|\mathbf{e}_i\|_1 \text{, subject to } \mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{e}. \qquad (21)$$

Then, in [12], the authors proposed to track from

$$\min_i \|\mathbf{y}_i - \mathbf{A}\mathbf{x}_i\|_2 . \qquad (22)$$

4.2. The Proposed Tracker

We propose a visual tracking algorithm, also based on ℓ_1 minimization, however, conceptually very different compared to the approach proposed in [12]. In particular, we reformulate tracking as two-class recognition problem as follows. We form our data matrix as the concatenation of two linear subspaces $\mathbf{Z} = [\mathbf{Z}_{\text{pos}} \;\; \mathbf{Z}_{\text{neg}}] \in \Re^{2m \times (n_{\text{pos}}+n_{\text{neg}})}$. The subspace $\mathbf{Z}_{\text{pos}} \in \Re^{2m \times n_{\text{pos}}}$ models the appearance of the object to be tracked. This is built from the eigen-space of image gradient orientations of the object in previously tracked frames. The matrix $\mathbf{Z}_{\text{neg}} \in \Re^{2m \times n_{\text{neg}}}$ is the eigenspace learned from the gradient orientations of *misaligned* examples also obtained from previously examined frames. Given a set of target candidates at time t, we assume that a candidate $\mathbf{q}_i$ models the appearance of the tracked object successfully, if it can be written as a linear combination of the bases in $\mathbf{Z}_{\text{pos}}$, while, possibly shifted and misaligned candidates are efficiently represented as a linear combination of the bases in $\mathbf{Z}_{\text{neg}}$. Thus, we solve

$$\min_{\mathbf{x}_i} \|\mathbf{x}_i\|_1 \,, \text{ subject to } \mathbf{q}_i = \mathbf{Z}\mathbf{x}_i, \tag{23}$$

and track from

$$\min_{i} \|\mathbf{q}_i - \mathbf{Z}_{\text{pos}}\mathbf{x}_i\|_2 \,. \tag{24}$$

Notice that, in contrast to [12], we use ℓ_1 minimization as a discriminant classifier which separates the object from the background and as such our algorithm is closely related to methods which perform “tracking by detection” [1, 4, 7].

As opposed to [12], the proposed tracker is based on sparse representations of image gradient orientations and does not rely on the extended problem of (21) to achieve robustness to outliers. This results in huge computational savings as (23) is solved in a low dimensional subspace.

We perform dimensionality reduction in a similar fashion to “Randomfaces” [18]. Let us denote by $\mathbf{\Theta} = [\boldsymbol{\theta}_1 | \cdots | \boldsymbol{\theta}_p]$ the $m \times p$ matrix whose columns $\boldsymbol{\theta}_j$ are samples from a uniform distribution $U(0, 2\pi)$. We define the projection bases as the columns of $\mathbf{B} \in \Re^{2m \times p}$ which is obtained by mapping $\mathbf{\Theta}$ onto the $\Re^{2m}$ sphere using (4). Notice that $\mathbf{B}^T\mathbf{B} \approx \mathbf{I}$ [15]. Finally, using $\mathbf{B}$ we can solve (23) using only p features.

5. Experimental Results

We evaluated the performance of the proposed framework for the application of face recognition, facial expression recognition and face tracking.

5.1. Face Recognition

Similarly to [18], we used the popular AR database [11] in order to compare the performance of our framework with the original formulation described in [18]. For all experiments, we used manually aligned cropped images of resolution 64×64. With the exception of the occlusion experiments, we used Linear Discriminant Analysis (LDA) for feature extraction and dimensionality reduction for both formulations. For the occlusion experiments, we used LDA only for feature extraction from image gradient orientations, while for pixel intensities, we solved the extended problem in (2). Finally, compared to the experiments described in [18], we considered a significantly smaller number of training samples for each subject, thus making our experimental setting noticeably more realistic.

The AR database [11] consists of more than 4,000 frontal view face images of 126 subjects. Each subject has up to 26 images taken in two sessions. Both sessions contains 13 images, numbered from 1 to 13, including different facial expressions (1-4), illumination changes (5-7), and occlusions under different illumination changes (8-13). We randomly selected a subset with 100 subjects and investigated the robustness of our scheme for the case of facial expressions, illumination variations and occlusions as follows

1. In experiment 1, we used images 1-4 of session 1 for training and images 2-4 of session 2 for testing.
2. In experiment 2, we used images 1-4 of session 1 for training and images 5-7 of session 2 for testing.
3. In experiment 3, we used images 1-4 of session 1 for training and images 8-13 of session 2 for testing.

Table 2 and Fig. 5 summarize our results. Note that for experiment 3, we did not apply feature extraction for the case of pixel intensities and solved the extended problem in (2). As we can see, sparse representations of gradient orientations performed better than the original formulation based on pixel intensities in all experiments. More specifically, our formulation achieves 100% recognition rate for the case of facial expressions and illumination changes (experiments 1 and 2), while the performance improvement over the original intensity-based formulation for the case of occlusions (experiment 3) goes up to 30%. Notice that this last result is significantly better that the one reported in [18] which was obtained with block processing and used twice as many training samples taken from both sessions.

	Image Gradient Orientations	Pixel Intensities
Experiment 1	100.0 %	96.00 %
Experiment 2	100.0 %	92.30 %
Experiment 3	97.55 %	66.00 %

Table 2. Recognition rates on the AR database

5.2. Facial Expression Recognition

We carried out facial expression recognition experiments on the CohnKanade database [9]. This database is annotated with Facial Action Units (FAUs). The combinations of

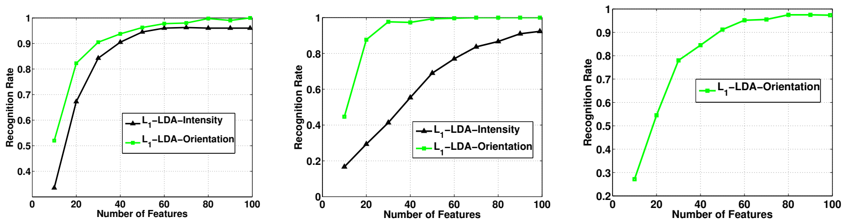


Figure 5. Face recognition experiments on the AR database. (a) Experiment 1 (facial expressions), (b) Experiment 2 (lighting conditions) and (c) Experiment 3 (occlusions).

FAUs were translated into six basic facial expression classes (anger, disgust, fear, happiness, sadness, and surprise). We considered all subjects in order to form the database for our experiments. Our database consists of a total of 352 of manually aligned cropped faces (of resolution 64 × 64) at the apex of each emotion. As in the previous subsection, we used LDA for feature extraction and dimensionality reduction for the proposed method and the baseline method of [18]. For each method, Tables 3 and 4 present the confusion matrices obtained using the "one subject out" procedure. Our algorithm resulted in a total recognition rate of 75% as opposed to a rate of 64% achieved in [18].

	anger	disgust	fear	happiness	sadness	surprise
anger	20	5	1	2	7	0
disgust	2	32	0	2	0	1
fear	4	1	21	19	5	4
happiness	2	0	6	78	2	1
sadness	3	0	2	5	50	6
surprise	2	0	3	1	2	63

Table 3. Confusion matrix for emotion recognition on the Cohn-Kanade database using the proposed scheme.

	anger	disgust	fear	happiness	sadness	surprise
anger	24	2	5	3	1	0
disgust	9	20	5	1	2	0
fear	4	3	29	16	1	1
happiness	3	3	14	68	1	0
sadness	9	4	14	6	30	3
surprise	3	6	3	4	1	54

Table 4. Confusion matrix for emotion recognition on the Cohn-Kanade database using the method in [18].

5.3. Face Tracking

We evaluated the performance of the proposed ℓ_1-based "tracking by detection" algorithm on two very popular video sequences, "Dudek" and "Trellis", available from http://www.cs.toronto.edu/dross/ivt/. The target was to assess the proposed algorithm's performance for face tracking under pose variation, occlusions and non-uniform illumination. 'Dudek" is provided along with seven annotated points which are used as ground truth. We also annotated seven fiducial points for "Trellis". As usual, quantitative performance evaluation is based on the RMS errors between the true and the estimated locations of these seven points [13]. The performance of our tracker is compared with that of [12]. No attempt to optimize the performance of both methods was attempted. For both methods, we used the same particle filter parameters (taken from [13]) and the same number of particles (600). For our method, we used $p = 100$ features after performing dimensionality reduction on 32 × 32 orientation images as described in Section 4. Finally, to make the method of [12] run in reasonable time, the region of interest was down-sampled to size 16 × 16.

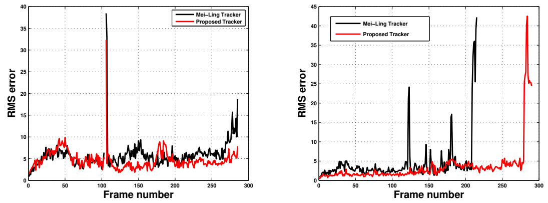


Figure 6. RMS error vs frame number for 'Dudek" (left figure) and "Trellis" (right figure) sequences.

Fig. 6 and Table 5 summarize our results. The proposed tracker outperforms the method of [12] in three important aspects. First, it is more robust. This is illustrated by the total number of frames successfully tracked before the tracking algorithm goes completely off. For "Dudek", both trackers went off after the 285^{th} frame. For "Trellis" however, our tracker appeared to be significantly more robust. More specifically, the proposed tracker failed to track after 280 frames, while the tracker of [12] after 200 frames. Second, the proposed scheme is more accurate. This is illustrated by the RMS error computed for frames where the face region was successfully tracked. For both methods and sequences, Fig. 6 plots the RMS error as a function of the frame number, while Table 5 gives the mean and median RMS error over the first 285 and 200 frames where the face region was tracked by both methods. Third, as our method does not rely on the extended problem of (21) to achieve robustness to outliers, it is significantly faster. Finally, Fig. 7 illustrates the performance of the proposed tracker for some cumbersome tracking conditions.

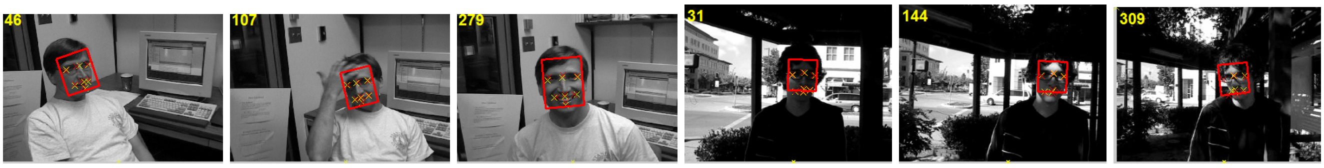


Figure 7. Face tracking using the proposed tracker. First three examples from “Dudek” and last three from “Trellis”.

	Proposed Tracker	Tracker of [12]
‘Dudek”	4.65 (4.22)	5.76 (5.36)
“Trellis”	2.15 (1.69)	3.37 (2.77)

Table 5. Mean (Median) RMS error for ‘Dudek” and “Trellis” sequences. The errors are computed for the first 285 and 200 frames.

6. Conclusions

We presented a framework for appearance-based visual recognition and tracking using sparse representations of gradient orientations. Our framework can handle outliers *without* the need for solving the extended problem considered in [18], can be combined with dimensionality reduction schemes and results in better recognition rates. Thus, it is not only significantly faster but also more robust.

References

[1] B. Babenko, M.-H. Yang, and S. Belongie. Visual Tracking with Online Multiple Instance Learning. In *CVPR*, 2009.

[2] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2007.

[3] E. Candes and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52:33–61, 2006.

[4] R. Collins, Y. Liu, and M. Leordeanu. Online selection of discriminative tracking features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1631–1643, 2005.

[5] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[6] D. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.

[7] H. Grabner, M. Grabner, and H. Bischof. Real-time tracking via on-line boosting. In *BMVC*, volume 1, pages 47–56. Citeseer, 2006.

[8] M. Isard and A. Blake. Condensationconditional density propagation for visual tracking. *International journal of computer vision*, 29(1):5–28, 1998.

[9] T. Kanade, Y. Tian, and J. Cohn. Comprehensive database for facial expression analysis. *fg*, page 46, 2000.

[10] K. Koh, S. Kim, and S. Boyd. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine learning research*, 8(8):1519–1555, 2007.

[11] A. Martinez and R. Benavente. The AR face database. Technical report, CVC Technical report, 1998.

[12] X. Mei and H. Ling. Robust visual tracking using l1 minimization. In *Proc. of Int. Conf. on Computer Vision (ICCV)*, pages 1–8, 2009.

[13] D. Ross, J. Lim, R. Lin, and M. Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77(1):125–141, 2008.

[14] G. Tzimiropoulos, V. Argyriou, S. Zafeiriou, and T. Stathaki. Robust fft-based scale-invariant image registration with image gradients. *IEEE transactions on pattern analysis and machine intelligence*, pages 1899–1906, 2010.

[15] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Principal Component Analysis of Image Gradient Orientations for Face Recognition. *Face and Gesture*, 2011.

[16] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma. Towards a practical face recognition system: Robust registration and illumination via sparse representation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, 2009.

[17] J. Wright and Y. Ma. Dense error correction via l1-minimization. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3033–3036. IEEE, 2009.

[18] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and M. Yi. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[19] A. Yang, A. Ganesh, Z. Zhou, S. Sastry, and Y. Ma. A Review of Fast l1-Minimization Algorithms for Robust Face Recognition. *Arxiv preprint arXiv:1007.3753*, 2010.

[20] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma. Face recognition with contiguous occlusion using Markov random fields. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1050–1057. IEEE, 2010.