Optimal Feature Selection for Subspace Image Matching

Gemma Roig
GTM-Grup de Recerca de Tecnologies Media
La Salle, Universitat Ramon Llull
c/Quatre Camins, 2, 08022, Barcelona
groig@salle.url.edu

Xavier Boix
Computer Vision Center,
Edifici O, Campus UAB,
08193 Bellaterra, Barcelona
xboix@cvc.uab.cat

Fernando De la Torre
Robotics Institute, Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
ftorre@cs.cmu.edu

Abstract

Image matching has been a central research topic in computer vision over the last decades. Typical approaches to correspondence involve matching features between images. In this paper, we present a novel problem for establishing correspondences between a sparse set of image features and a previously learned subspace model. We formulate the matching task as an energy minimization, and jointly optimize over all possible feature assignments and parameters of the subspace model. This problem is in general NP-hard. We propose a convex relaxation approximation, and develop two optimization strategies: naive gradient-descent and quadratic programming. Alternatively, we reformulate the optimization criterion as a sparse eigenvalue problem, and solve it using a recently proposed backward greedy algorithm. Experimental results on facial feature detection show that the quadratic programming solution provides better selection mechanism for relevant features.

1. Introduction

There exists a huge literature that addresses the correspondence problem between images [2, 3, 14, 16, 20, 21, 22]. Most common methods extract some descriptors from images, and solve the matching problem using the Hungarian algorithm, genetic algorithms, Hopfield networks, linear programming, integer programming or several relaxations. Other problems in computer vision involve feature localization in images.

Although establishing correspondences between image features has a long history in computer vision, a relatively unexplored problem is how to find correspondences between a large set of descriptors and a subspace model. There are many applications in computer vision where the problem consist on finding the correspondence between a set of features and a given model, rather than solving correspondence between images. Fig. 1 illustrates the main aim of the paper. Given several image features, (*e.g.* corners detected with Harris detector), the goal is to select a subset of these features that best match a model. (*e.g.* shape and/or appearance). In Fig. 1, the set of ideally selected landmarks would correspond to the corners of the eyes and mouth. The subspace model of shape and appearance has been previously learned from a set of labeled images. In Fig. 1, crosses represent detected points, and the selected landmarks are marked with diamonds (corners of the left

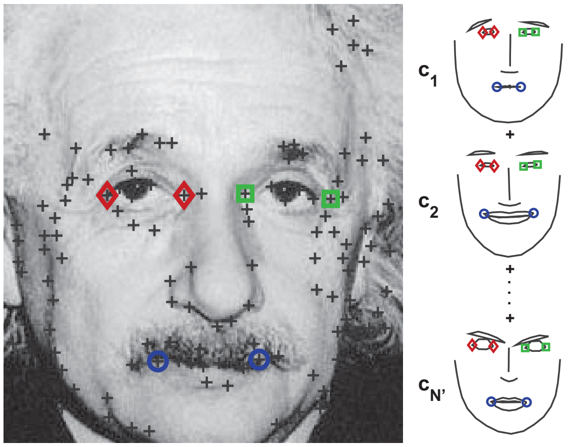


Figure 1. Selection of a subset of image features that best reconstruct a shape and appearance model. Crosses represent detected points, and diamonds, squares and circles are selected landmarks. Right images represent modes of variation in the shape model.

eye), squares (right eye), and circles (mouth).

The method that we propose in this paper jointly optimizes over parameters of the subspace model and the selection of image features. This problem is NP-hard, and a relaxation approximation is proposed. We evaluate two optimization strategies based on gradient-descent and quadratic programming. Moreover, we reformulate the subspace selection problem as a sparse eigenvector computation, and solve it using a backward selection algorithm [17].

The rest of the paper is organized as follows. Section 2 reviews previous work on image matching and feature selection. Section 3 discusses how to formulate the subspace feature selection problem as an energy minimization. Section 4 describes different methods to compute a relaxed solution based on naive gradient-descent and quadratic programming approaches. In addition, an approximate formulation based on cardinality constraints is proposed. Section 5 compares the performance of the three strategies in facial feature detection, showing that the quadratic programming algorithm is the method that gives more accurate results. Finally, Section 6 discusses conclusion and future work.

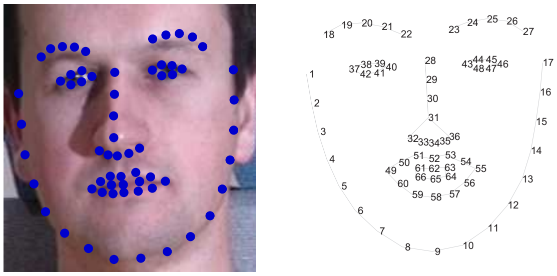
2. Previous work

In this section, we review previous work on two related topics: image matching and feature selection.

2.1. Image matching

Lowe's SIFT descriptor [15] is one of the state-of-the-art methods to construct geometric invariant features to match rigid objects. SIFT has been successfully applied to many problems, *e.g*. in Mikolajczyk and Schmid evaluation [16]. In the context of non-rigid shape registration, Belongie *et al.* [2, 3] designed a shape context histogram that has been shown to provide robustness to image matching. Alternatively, Leordeanu and Herbert [14] formulated visual correspondences as a graph matching problem, building an affinity matrix between pair-wise points, and thresholding the leading eigenvectors. In related work, Sclaroff *et al.* [20] found correspondences, and established canonical descriptors. This method allows computing models' eigenmodes directly from available image information.

Recently, Torresani *et al.* [23] proposed an energy minimization approach to establish correspondences for non-rigid motion. Minimization of an instance of this error function can be solved as a graph matching problem [23]. Other soft assign techniques were developed by Rangarajan *et al.* [19], and the ICP based method by Besl and McKay [4]. In this paper, we extend previous approaches by selecting a subset of features that minimize the distance to a given subspace model.



(a) (b)

Figure 2. (a) 66 hand-labeled landmarks and (b) it's reference.

2.2. Feature selection methods

Feature selection has been extensively studied in machine learning and statistics over the last few years. The problem is defined as finding a subset of features that is sufficient to encode (*e.g*. unsupervised) or predict (*e.g*. supervised) target labels. In this paper, we show that selecting the subset of image features that is better encoded with a model can be posed as a feature selection problem.

Feature selection has been widely explored in supervised learning problems, *e.g*. support vector machines. One popular technique is RELIEF [13] that assigns weights to a particular feature based on the differences between the values of nearest neighbor pairs. Cao *et al.* [5] further developed this method by learning feature weights in kernel space, pruning away unnecessary features. Hermes and Buhmann [11] started by constructing an SVM classifier using all available features, and recursively removing those that have the least impact on the decision function if removed. Similarly, Avidan [1] used a greedy sequential forward selection method to find a subset of features and support vectors that approximate the SVM solution obtained using all available features.

3. Image feature matching as an optimization problem

This section describes the optimization problem to select a subset of features that minimize the distance to a model, and how to build these shape and appearance models.

3.1. Selecting good image features

During the paper, we will illustrate our method in detecting facial features in images, but it can be applied in any other context. This section, describes the process of selecting good features to match.

We use Harris corner detector [10] to detect image features. Fig. 1 shows some examples of the detected features in a face image. To verify that some points obtained with Harris detector correspond to manually labeled landmarks of a face, we have computed statistics on the MultiPIE

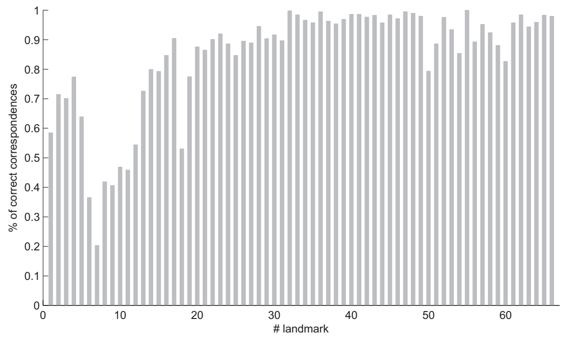


Figure 3. Histogram of percentage of points that are less than 4 pixels from a landmark. It is computed with 2500 images.

database [9] over 2500 frontal face images under different expressions (*i.e.* smiling, neutral, surprise, squint, sad). All images are labeled with 66 landmarks (see Fig. 2a), with reference number shown in Fig. 2b. Fig. 3 shows the histogram of percentage of times the points detected by Harris are within 4 pixels from the landmarks. As expected, the landmarks in the outline of the face (1 to 17) are typically dominated by face edges, and hence not very reliable. The best agreement between landmarks and detected points are corners of the eyes, 37, 40, 43, 46, and corners of the mouth, 49 and 55. Our methods will be illustrated using these six ($k = 6$) landmarks to build a shape and appearance subspace model.

3.2. Building a shape and an appearance model

Let $\mathbf{X} \in \mathbb{R}^{kr \times N}$(see notation[1]) be a matrix containing the features, where r denotes the dimension of the descriptor and N the number of training face images. We build two different models: appearance and shape. In the shape model, each column $\mathbf{x}_i$ contains 12 features corresponding to $r = 2$ (x, y) position of the $k = 6$ landmarks. Similarly, we use histogram of oriented gradients [15] as descriptor for the appearance subspace which $r = 128$.

We use 1000 frontal labeled images from MultiPIE [9] to build a generic shape and appearance model of 6 landmarks. The points are previously aligned with Procrustes analysis [6]. For each shape and appearance descriptor, we build a subspace computing the PCA [12] on $\mathbf{X}$, and selecting the N' components that preserve 90% of the energy. Fig. 1 shows an example of the facial features detected using Harris detector, and three modes of variation of the shape model. In testing, we localize and scale-normalize the face using the Viola and Jones face detector [24].

3.3. Image matching as unsupervised subspace feature selection problem

The main aim of this paper is to develop algorithms that optimally select a subset of k landmarks from n image features ($k << n$) that minimize the distance to a subspace model. This can be achieved minimizing:

$$E(\mathbf{P}, \mathbf{c}) = ||vec(\mathbf{PD}) - \boldsymbol{\mu} - \mathbf{Bc}||_2^2 \qquad (1)$$
$$s.t.\ p_{ij} \in \{0, 1\},\ \ \sum_j p_{ij} = 1\ \forall\ i,\ \ \sum_i p_{ij} = \{0, 1\}\ \forall\ j$$

where $\mathbf{D} \in \mathbb{R}^{n \times r}$ is a matrix such that each row contains r descriptors, and n detected points. For instance, when we use the shape model $\mathbf{D} \in \mathbb{R}^{n \times 2}$, and when using appearance model $\mathbf{D} \in \mathbb{R}^{n \times 128}$. $\mathbf{P} \in \mathbb{R}^{k \times n}$ is an indicator matrix that $\sum_j p_{ij} = 1\ \forall i$, $p_{ij} \in \{0, 1\}$, and p_{ij} is 1 if the feature i belongs to the subset of k points that minimize the distance to the model. The sum of the columns of $\mathbf{P}$ can be either 0 or 1, that is: $\sum_i p_{ij} = \{0, 1\}\ \ \forall j$. Besides, $\boldsymbol{\mu} \in \mathbb{R}^{kr \times 1}$, $\mathbf{B} \in \mathbb{R}^{kr \times N'}$ and $\mathbf{c} \in \mathbb{R}^{N' \times 1}$ are respectively the mean, the basis and the coefficients of $vec(\mathbf{PD})$ on the model subspace.

The objective of the optimization is to simultaneously find the subset of k features ($\mathbf{P}$) and the subspace coefficients ($\mathbf{c}$) that minimize the error E in Eq.(1). To reduce the number of parameters, we replace $\mathbf{c}$ to its optimal value $\mathbf{c} = \mathbf{B}^T(vec(\mathbf{PD}) - \boldsymbol{\mu})$. After some algebra, and using the orthonormality property of the basis vectors $\mathbf{B}^T\mathbf{B} = \mathbf{I}$, it can be shown that Eq.(1) can be rewritten as:

$$\begin{aligned} E(\mathbf{P}) &= ||\mathbf{H}(vec(\mathbf{PD}) - \boldsymbol{\mu})||_2^2 \propto \\ &\frac{1}{2}vec(\mathbf{P})^T\mathbf{Q}vec(\mathbf{P}) - \mathbf{b}^T vec(\mathbf{P}) \end{aligned} \qquad (2)$$

where $\mathbf{Q} = (\mathbf{D} \otimes \mathbf{I}_k)\,\mathbf{H}^T\mathbf{H}\,(\mathbf{D} \otimes \mathbf{I}_k)^T \in \mathbb{R}^{kn \times kn}$, $\mathbf{H} = (\mathbf{I}_k - \mathbf{BB}^T) \in \mathbb{R}^{kr \times kr}$, and $\mathbf{b} = (\mathbf{D} \otimes \mathbf{I}_k)\mathbf{H}^T\mathbf{H}\boldsymbol{\mu} \in \mathbb{R}^{kn \times 1}$.

In addition, we combine the error due to the appearance, $E_a(\mathbf{P})$, and the error due to the shape, $E_s(\mathbf{P})$ as:

$$E(\mathbf{P}) = E_a(\mathbf{P}) + \lambda E_s(\mathbf{P}) \qquad (3)$$

where λ weights the contribution of the two functions. Eq.(3) can be re-written as the quadratic form of Eq.(2):

$$\mathbf{Q} = \mathbf{Q}_a + \lambda \mathbf{Q}_s \qquad (4)$$
$$\mathbf{b} = \mathbf{b}_a + \lambda \mathbf{b}_s \qquad (5)$$

4. Optimization strategies

Minimization of Eq.(1) subject to $p_{ij} \in \{0, 1\}$ and $\mathbf{P1}_n = \mathbf{1}_k$ is a binary NP-hard optimization problem. This section explores several algorithms to solve an approximated problem. In two of our proposed algorithms, $\mathbf{P}$ is relaxed in order to use naive gradient-descent and a quadratic programming algorithm. In the other method, we constrain the cardinality of $\mathbf{P}$ and use a backward greedy strategy.

[1] Bold capital letters denote a matrix $\mathbf{D}$, bold lower-case letters a column vector $\mathbf{d}$. $\mathbf{d}_j$ are the jth column of the matrix $\mathbf{D}$. Non-bold letters denote scalar variables. d_{ij} the scalar in the row i and column j of the matrix $\mathbf{D}$. $\mathbf{1}_k \in \mathbb{R}^{k \times 1}$ is a vector of ones. $\mathbf{I}_k \in \mathbb{R}^{k \times k}$ is the identity matrix. $||\mathbf{d}||_2^2$ is the norm squared of the vector $\mathbf{d}$. vec($\mathbf{D}$) is an operator that vectorizes a matrix $\mathbf{D}$ into a vector. $\circ$ is the Hadamard product, $\otimes$ the Kronecker product, and $*$ the convolution.

4.1. Naive gradient-descent

Using relaxation techniques, $\mathbf{P}$ can be optimized with a naive gradient-descent algorithm. The discrete constraint on p_{ij} is relaxed allowing values in the range $(0, 1)$. Following previous work [7], $\mathbf{P}$ is parameterized as the Hadamard product of two matrices $\mathbf{P} = \mathbf{V} \circ \mathbf{V}$. This ensures positiveness of the search. The gradient-descent updates $\mathbf{V}$ as:

$$\mathbf{V}^{n+1} = \mathbf{V}^{n} - \eta \tfrac{\partial E}{\partial \mathbf{V}} \tag{6}$$
$$\tfrac{\partial E}{\partial \mathbf{V}} = 4 \left(\mathbf{H} \left(vec(\mathbf{PD}) - \boldsymbol{\mu} \right) vec(\mathbf{D})^{T} \right) \circ \mathbf{V}$$

The increment of the gradient, η, is determined with a line search strategy [8]. To impose $\mathbf{P1}_n = \mathbf{1}_k$, $\mathbf{V}$ is normalized after each iteration to satisfy the constraints. After convergence, the result is obtained by selecting the points that correspond to the maximum of each row of $\mathbf{P}$. Because Eq.(6) is prone to local minima, the method is started from several initial random points, and selects the solution with smallest error.

4.2. Quadratic programming

Alternatively, quadratic programming approaches can be used for solving Eq.(2). Eq.(1) can be reformulated as a quadratic programming problem. That is,

$$\min \;\; \tfrac{1}{2} vec(\mathbf{P})^{T} \mathbf{Q} vec(\mathbf{P}) - \mathbf{b}^{T} vec(\mathbf{P}) \tag{7}$$
$$s.t.\; p_{ij} \in \{0, 1\}, \;\; \textstyle\sum_j p_{ij} = 1 \;\forall\, i, \;\; \sum_i p_{ij} = \{0, 1\} \;\forall\, j$$

Observe that if $\mathbf{Q}$ is positive definite, the problem is convex. Following previous work by Billionnet [18], we subtract the minimum eigenvalue of $\mathbf{Q}$, $\lambda_{min}(\mathbf{Q})$, to make the new matrix $\mathbf{Q}'$ positive definite, that is:

$$\mathbf{Q}' = \mathbf{Q} - \lambda_{min}(\mathbf{Q}) \mathbf{I}_k \tag{8}$$

Therefore, the optimization problem translates to:

$$\frac{1}{2} vec(\mathbf{P})^{T} \mathbf{Q}' vec(\mathbf{P}) - (\mathbf{b}')^{T} vec(\mathbf{P}) \tag{9}$$

where $\mathbf{b}' = \mathbf{b} + \lambda_{min}(\mathbf{Q})$.

As in the case of the gradient, the rounding is done by selecting the maximum elements for each row of $\mathbf{P}$.

4.3. Greedy

The QP problem formulated in Section 3 can be approximated as a sparse quadratic programming if a cardinality constraint is added, *i.e.* card$(vec(\mathbf{P})) = k$. Recently, Moghaddam *et al.* [17] proposed a greedy method that uses backward elimination to optimize sparse QP with cardinality constraints. The connection between [17] and Eq.(7) becomes clearer if we rewrite Eq.(7) as the maximization of a Generalized Rayleigh Quotient with an sparsity constraint:

$$\frac{vec(\mathbf{P})^{T} \mathbf{b}\mathbf{b}^{T} vec(\mathbf{P})}{vec(\mathbf{P})^{T} \mathbf{Q} vec(\mathbf{P})} \quad s.t.\; \text{card}(vec(\mathbf{P})) = k \tag{10}$$

Observe that with the cardinality constraint there is no guarantee that the solution $\mathbf{P}$ satisfies the constraints of Eq.(7). However, we will show that the sparse QP problem provides a good solution.

The method described by Moghaddam *et al.* [17] starts by selecting all possible detected points, *i.e.* $vec(\mathbf{P}) = \mathbf{1}_{kn}$. Then, it iteratively assigns 0 to the p_{ij} that leads to the maximum eigenvalue of $(\mathbf{Q}_{k'})^{-\frac{1}{2}} (\mathbf{b}\mathbf{b}^{T})_{k'} (\mathbf{Q}_{k'})^{-\frac{1}{2}}$. The subindex k' denotes the $k' \times k'$ submatrix obtained by deleting the rows and columns where p_{ij} is selected (substituted by 0), and k' is the card$(vec(\mathbf{P}))$ in each iteration. At the first step $k' = kn$, and at the end $k' = k$. See [17] for more details on the greedy approach.

5. Experiments

This section reports experimental results comparing the performance of the three proposed algorithms on the problem of facial feature detection. The test set is composed by 1500 face images of 640×480, different from the training, of MultiPIE database [9].

5.1. Synthetic experiments

This section evaluates the robustness of the proposed solutions. We artificially add noisy points around the ground truth data. The spatial distribution of noise follows a two dimensional Gaussian of zero-mean and standard deviation of 10 pixels. Fig. 5 shows few examples of an image with different noise levels.

Fig. 4 shows the results of running the three algorithms proposed in Section 4 with appearance, and appearance and shape models. The λ that weights the contribution of shape and appearance in Eq.(3) is empirically set to 10^{-5}.

The same random initialization is used for the different noise levels and methods. Fig. 4a shows the error between the selected points and the ground truth. The error is the mean Euclidean distance between the labeled landmarks and the selected points by the algorithm. It shows the mean, and the error contained between the first and third quartile for different noise levels. Fig. 4b shows the final value obtained by the energy function, Eq.(2). Observe that there is consistency between having lower energy and better results.

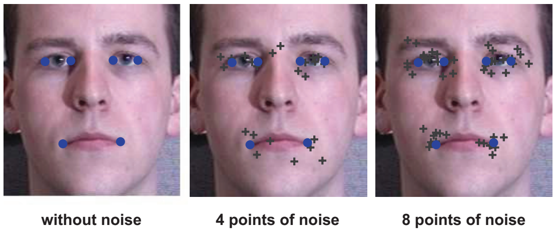
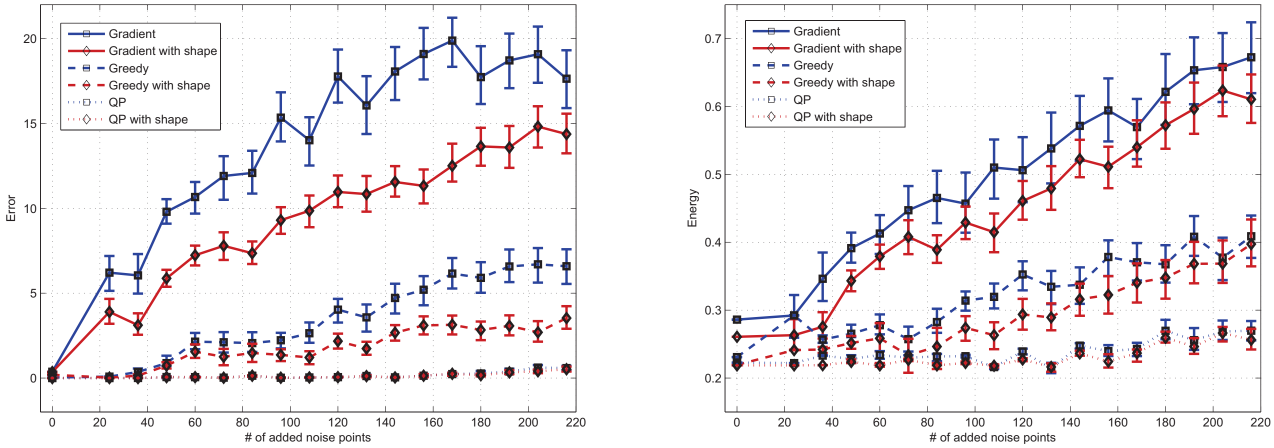


Figure 5. Blue dots denote the 6 hand-labeled landmarks. Crosses represents artificial noisy landmarks.



(a)

(b)

Figure 4. (a) Mean Euclidean distance between the selected points and manually labeled landmarks versus noise levels. (b) Final energy values for Eq.(2) versus different noise levels.

There are several conclusions that can be extracted. First, as expected, the combination of shape and appearance models always outperforms the only appearance model. Typically, adding the shape model avoids the confusion between the appearance of the corners of the eyes due to the symmetry of the face. Moreover, it can be seen that the gradient-descent method often gets stuck into local minima. This behavior gets worse with increasing noise levels. The technique that gives better performance is quadratic programming, followed by the greedy approach. This last one is much less myopic than gradient-descent, and it has better performance, *e.g*. a mean of the error of 1.1 pixels when 108 wrong points are added.

One reason for the good performance of QP is the fact that it is reformulated as a convex problem with Eq.(8). Moreover, in the case of QP adding the shape model does not increase its performance significantly . In terms of computational complexity, the backward greedy is the most computationally demanding, followed by QP and gradient descent.

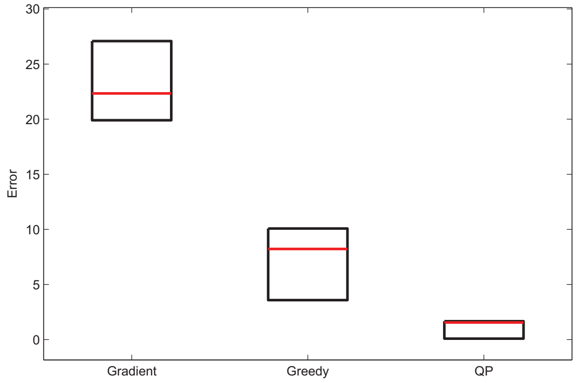


Figure 6. Error obtained with different methods to select a subset of 6 points detected with Harris. The mean is represented in red, and first and third quartile in black.

5.2. Locating facial landmarks in images

This experiment tests the ability of the three methods to locate facial features in frontal faces in untrained images. Fig. 7 shows several faces where we have run Harris corner detector. Diamonds indicate the selected points for the left eye corners, squares denote right eye corners, and circles the corners of the mouth. In the results shown in Fig. 7, we used an appearance and a shape model. As we can observe, landmarks of the mouth are easily detected, and most of the errors are in the eyes because of the distribution and amount of natural noise surrounding in the eye region.

Fig. 6 shows the numerical values for the error. As in the previous experiment, the QP method outperforms the other methods. In comparison with the synthetic experiment, the error has increased in general because Harris detector does not always detect the exact landmark position.

6. Conclusion and future work

We have presented a method to find correspondences between a set of image features and a learned subspace model. We have proposed an energy function to select the subset of points that minimize the distance to a subspace. The problem has been posed as a joint optimization over subspace parameters and matrix assignment. We have proposed two algorithms to solve a relaxation of the original problem, and another approximation with a cardinality constraint. Synthetic and real experiments show that QP has better performance than greedy and gradient-descent. Merging appearance and shape models improve all optimization methods.

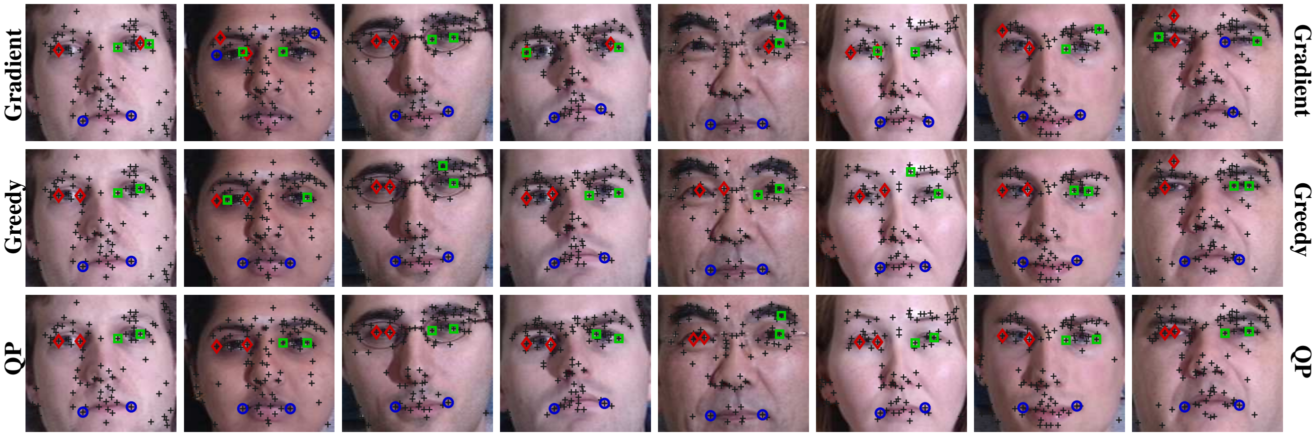


Figure 7. Example of representative images with detected facial features using gradient-descent, backward greedy and quadratic programming. Detected points are obtained with Harris (crosses). See Fig. 1 for the expected location of circles, squares and diamonds.

Acknowledgment

This work was supported by University, Research and Information Society Department of Catalonia Government, and by Spanish Ministry of Education and Science under project TRA2007-62526/AUT, Research Program Consolider Ingenio 2010: MIPRCV (CSD2007-00018).

References

[1] S. Avidan. Joint feature-basis subset selection. In *CVPR*, 2004.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. In *Advances in Neural Infor. Processing Systems*, 2000.

[3] S. Belongie, J. Malik, and J. Puzicha. Matching shapes. In *ICCV*, pages 454–461, 2001.

[4] P. J. Besl and H. D. Mckay. A method for registration of 3-d shapes. *PAMI*, 14(2):239–256, 1992.

[5] B. Cao, D. Shen, J.-T. Sun, Q. Yang, and Z. Chen. Feature selection in a kernel space. In *International Conference on Machine Learning*, 2007.

[6] T. Cootes and C. Taylor. Statistical models of appearance for computer vision. In *Tech. Report. U. of Manchester*, 2001.

[7] F. de la Torre and T. Kanade. Discriminative cluster analysis. In *International Conference on Machine Learning*, 2006.

[8] R. Fletcher. *Practical methods of optimization*. Wiley, 1987.

[9] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker. The cmu multi-pose, illumination, and expression (multipie) face database. Technical report, Carnegie Mellon University Robotics Institute.TR-07-08, 2007.

[10] C. Harris and M. Stephens. A combined corner and edge detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[11] L. Hermes and J. Buhmann. Feature selection for support vector machines. In *ICPR*, 2000.

[12] I. T. Jolliffe. *Principal Component Analysis*. New York: Springer-Verlag, 1986.

[13] K. Kira and L. A. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *AAAI*, pages 129–134, 1992.

[14] M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pairwise constraints. In *ICCV*, pages 1482–1489, 2005.

[15] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[16] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.

[17] B. Moghaddam, A. Gruber, Y. Weiss, and S. Avidan. Sparse regression as a sparse eigenvalue problem. *Information Theory and Applications Workshop*, pages 121–127, February 2008.

[18] M. C. Plateau. Quadratic convex reformulations for quadratic 0-1 programming. *4OR: A Quarterly Journal of Operations Research*, 6(2):187–190, June 2008.

[19] A. Rangarajan, F. Chui, and H. Bookstein. The softassign procrustes matching algorithm. *Information Processing in Medical Imaging*, pages 29–42, 1997.

[20] S. Sclaroff and A. Pentl. Modal matching for correspondence and recognition. *PAMI*, 17:545–561, 1995.

[21] G. L. Scott and H. C. Longuet-Higgins. An algorithm for associating the features of two images. In *Proc. Royal Society London*, volume B244, pages 21–26, 1991.

[22] P. H. S. Torr. Geometric motion segmentation and model selection. *Phil. Trans. Royal Society of London*, 1998.

[23] L. Torresani, V. Kolmogorov, and C. Rother. Feature correspondence via graph matching: Models and global optimization. *ECCV*, pages 596–609, 2008.

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