Image and Vision Computing xxx (2008) xxx–xxx

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Distinctive and compact features

Ayelet Akselrod-Ballin *, Shimon Ullman

Department of Computer Science and Applied Math, The Weizmann Institute of Science, Rehovot 76100, Israel

A R T I C L E I N F O

Article history:
Received 16 March 2006
Received in revised form 3 March 2008
Accepted 4 March 2008
Available online xxxx

Keywords:
Feature selection
Object recognition
Distinctive features
Facial features

A B S T R A C T

We consider the problem of extracting features for multi-class recognition problems. The features are required to make fine distinctions between similar classes, combined with tolerance for distortions and missing information. We define and compare two general approaches, both based on maximizing the delivered information for recognition: one divides the problem into multiple binary classification tasks, while the other uses a single multi-class scheme. The two strategies result in markedly different sets of features, which we apply to face identification and detection. We show that the first produces a sparse set of distinctive features that are specific to an individual face, and are highly tolerant to distortions and missing input. The second produces compact features, each shared by about half of the faces, which perform better in general face detection. The results show the advantage of distinctive features for making fine distinctions in a robust manner. They also show that different features are optimal for recognition tasks at different levels of specificity.

1. Introduction

In performing recognition, the visual system, either human or artificial, must cope with the problem of image variability, that is, that an object's appearance is highly variable due to changes in shape, viewing direction, illumination, and occlusion. At the same time, the task often requires making fine distinctions between objects, such as between similar faces. It is particularly surprising given these difficulties that reliable recognition can be obtained on the basis of reduced and distorted representations, such as caricatures and drawings produced by artists, e.g. [1], see examples in Fig. 1. In such images, the faces consist only of a few informative features that are distorted, often represented schematically, and placed in an inaccurate spatial arrangement. This illustrates a fundamental general question: how is it possible to reliably distinguish between multiple similar classes, and yet be tolerant to reduced and distorted information?

To approach this problem, we define and compare two natural strategies for extracting classification features in problems involving multiple similar classes, and apply them to face examples. Both are based on maximizing information for classification, but they produce notable different features. One method divides the problem into multiple binary classification tasks, while the other uses a single multi-class scheme. We show that the first leads to a **sparse** representation based on distinctive features, which is tolerant to large distortions and missing input, and better for robust face identification, requiring only a few distinctive features for reliable identification. The second leads to **compact** coding where each features is shared by about half of the faces, and which performs better in general face detection. The distinctive features are also shown to be similar to the ones selected by an artist specializing in producing reduced face representations [1], and the algorithm is the first to automatically produce such distinctive features. The focus of the study is on feature selection for multi-class recognition, rather than face recognition. Face images are used as a testing domain, for which there are example of distinctive features selected by human experts.

The rest of the paper is organized as follows: Section 2 reviews past relevant approaches to face recognition and detection, with emphasis on the type of features used by these approaches. Section 3 describes the two selection strategies, and automatic extraction of sparse and compact features. Section 4 presents experimental results, comparing sparse and compact features in face recognition and detection. We also compare between the distinctive fragments obtained by the current method and the representations produced by an artist. Section 5 includes a discussion of the results and conclusions.

2. Previous work

The current study considers the problem of extracting features for multi-class recognition problems, and compares two alternative feature selection strategies. Since we evaluate the two schemes in the domain of faces, we briefly review relevant aspects of past approaches for feature extraction and use it in this domain.

A large number of face recognition schemes have been developed in the past, using different families of features and different

* Corresponding author. Tel.: +972 8 9344268; fax: +972 8 934 4122.
E-mail address: ayelet.akselrod-ballin@weizmann.ac.il (A. Akselrod-Ballin).

0262-8856/$ - see front matter © 2008 Published by Elsevier B.V.
doi:10.1016/j.imavis.2008.03.005

classification methods (for recent reviews, see [2–5]). Often, the same type of classifier, for instance, support vector machines (SVM, [6,7]), can be used with different feature types, leading to different classification performance. We focus below on the main approaches and the type of features they selected and used, since this is the most relevant aspect to the current work.

A wide range of features have been used for both face recognition and face detection. Appearances based methods use image examples of face regions for learning models, and typically apply statistical analysis and machine learning techniques for recognition. The image appearance is used directly for recognition, using either global descriptions (e.g. PCA [8], ICA [9]), or the appearance of local face regions such as [10] for face detection. Decision can then be reached using for instance projection distance, [8] or linear discriminant analysis (LDA/FLD) [11].

Structural matching methods based on geometrical constraints use as features measured distances and angles between key points of the face [12,13]. A recent example within this category is the active shape model (ASM) [14], which is a statistical shape model, representing faces with shape and intensity information.

Deformable templates methods use a geometric model of the face, but allow it to deform in a controlled manner during the matching process. For example, in [15], facial features are described by parameterized templates, which are matched to an image by minimizing an energy function.

Several recognition systems use constellations of simple local features, including wavelets, Gabor patches, edges, lines and curves, for representing and recognizing faces. In such approaches the face is described by the constellation, sometimes modeled as joint distribution, of the features. The face detection algorithm developed in [16] uses a multi layer network to directly learn input image intensities. The algorithm presented in [17] classifies objects based on a set of rectangular features, where each feature computes the sum and difference of pixel intensities within a number of sub-rectangles. In the Elastic Bunch Graph Matching system of [18], faces are represented as graphs, with nodes positioned at key points on the face (eyes, tip of nose, mouth, etc.), and the features used are based on wavelet responses. Wavelet transforms were used also by Schneiderman and Kanade [19] and applied to the detection of faces and cars. In general, previous methods used the same set of features, often extracted in an ad hoc manner, for all recognition tasks, and did not compare features optimized for a single individual, multi-class recognition, and general face detection.

Psychological studies support the claim that in human vision some type of distinctive features are used for face recognition [20,21]. A recent study [22] showed that in performing recognition, humans focus on restricted regions in the face, and that the selected regions are task-dependent. The study supports the notion that the visual system does not rely on a fixed set of features, but learns for each task to use a small subset of critical features that are the most informative for the task.

The methods described above rely on an accurate geometrical agreement between the face model and the input image. They therefore have severe limitations in their ability to deal with reduced and distorted images. These limitations can be illustrated by comparing real images with artists drawings (as in Fig. 1), which are recognizable by human observers despite the large distortions and features omission in the input images.

In the present work, we compare two alternative strategies to the selection of useful features in multi-class problems in general, and face recognition in particular. We show that one of these strategies produces a representation that relies on the presence of a small number of distinctive features, and can use them for recognition without relying on exact geometric agreement between the model and the input image. These features and their extraction are described in the following section.

3. Feature extraction

3.1. Sparse and compact features

We contrast below two alternative approaches to extracting useful visual features for classifying a novel image, into one of n known classes. For example, the training may consist of face images taken from n different individuals under different viewing conditions (see Fig. 3), and the task is to then classify a novel image of one of the known individuals. One strategy results in sparse, the other in compact representation. Compact coding uses features

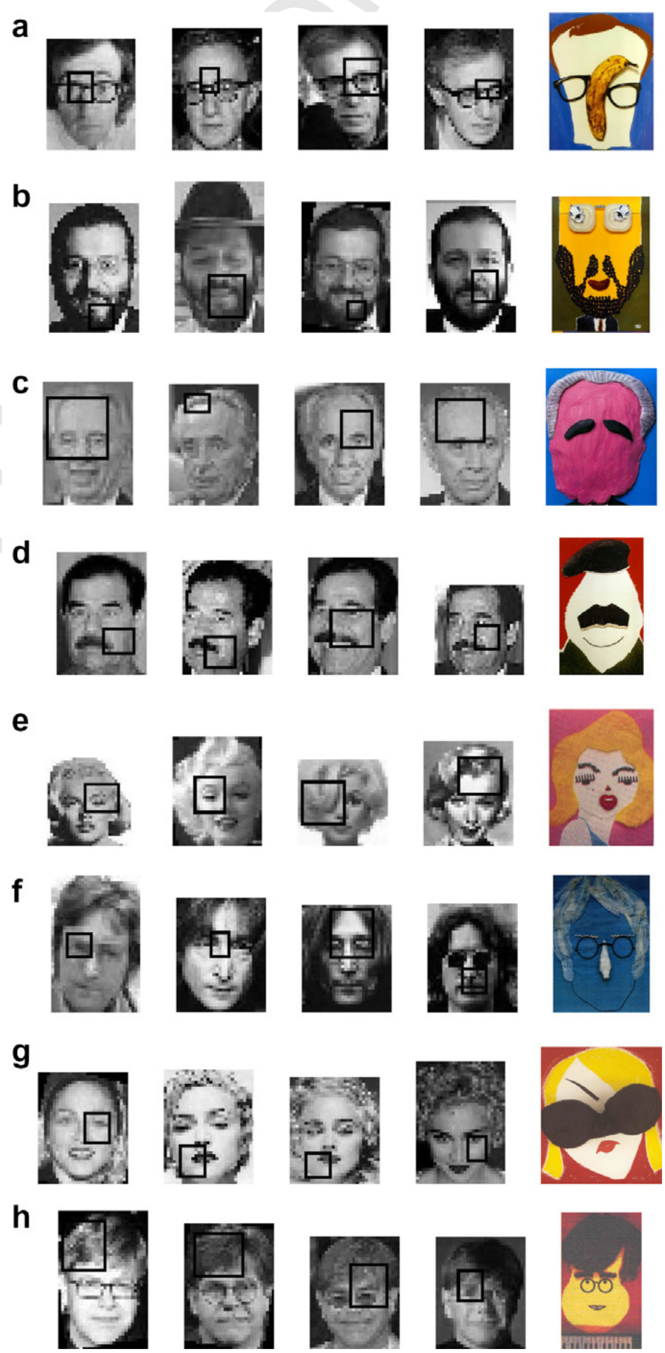

Fig. 1. Sparse fragments extracted for several individuals. The black rectangles displayed on the images illustrate the set of informative extracted fragments (in decreasing order). The corresponding artist's images (by H. Piven) for these individuals is shown on the right column in each panel. (a) Allen. (b) Deri. (c) Peres. (d) Sadam. (e) Monroe. (f) Lennon. (g) Madonna. (h) Elton.

Please cite this article in press as: A. Akselrod-Ballin, S. Ullman, Distinctive and compact features, Image Vis. Comput. (2008), doi:10.1016/j.imavis.2008.03.005

Fig. 2. Compact fragments extracted in decreasing order of mutual information from left to right.

Fig. 3. Examples of face images used for training and testing in the multi-class recognition tasks. (a) Allen. (b) Monroe. (b) Peres.

Fig. 4. Examples of occluded face images for task 5.

that are common to many faces, where, for binary features, each face is represented on average by about half of the features being active. The entire set of faces can then be represented with a relatively small number of features. In contrast, in the sparse coding each feature is present only in a small subset of the faces [23] (Fig. 4).

In terms of the features used in this work, we follow a number of recent recognition schemes that have successfully used as basic classification features a set of selected image regions, called fragments, or patches [10,24–26]. We describe briefly below the general scheme used in the current work, and then show how the feature extraction process can be used to extract either sparse or compact features. For more details on the general scheme see [24,25].

The features extracted during training are image patches, selected by maximizing the information they deliver about the objects to be recognized. The use of mutual information for feature selection is motivated by both theoretical and experimental results, and was shown to produce highly effective features [27,28]. During a training phase, informative features are extracted automatically from a set of labeled training images. First, a large number of candidate fragments are extracted from the training images, at different positions and scales. They provide an initial pool of possible classification features, from which a subset of non-redundant features is selected (see Sections 3.1.1 and 3.1.2). Second, for each fragment, the amount of information it supplies for classification is evaluated, based on the frequency of detecting the fragment within and outside the class.

In the simplest case, features are extracted to make a binary distinction between class and non-class images. To classify a novel image into one of n different classes (i.e. different individuals), two extensions of the simpler binary classification are possible, leading to two families of visual features. One is to treat the extended classification as n different classification problems, between a single class and all remaining images (**multiple classification approach**). The other is to consider all the classes jointly in a single classification task (**joint classification approach**). Both approaches use information maximization, but feature information is evaluated in a different manner, as discussed below. Table 1 summarizes the fragment based classification algorithm.

3.1.1. *Extraction of candidate fragments*

In both strategies, the first stage of candidate feature extraction generates a set of candidate fragments, computes their similarity to all images in the database, and selects the best fragments in the sense of mutual information between class C and fragment F. Potential fragments, up to several tens of thousands, are generated by extracting rectangular sub-images of different sizes and locations from class images. The fragment extraction of sub-images is proportional to the number of training images and the size of the images. Each potential fragment is compared to all training images, by searching over a restricted range of locations (steps of two pixels), and the location with the highest similarity measure sets a score for the fragment and the image. The similarity measure used in our comparisons was based on the absolute value of normalized cross-correlation (NCC) which is given by:

$$\mathrm{NCC} = \frac{\sum_{i=1}^{m}(I^i - \bar{I})(f^i - \bar{f})}{\sqrt{\sum_{i=1}^{m}(I^i - \bar{I})^2}\sqrt{\sum_{i=1}^{m}(f^i - \bar{f})^2}} \tag{1}$$

where f, I stand for the fragment and image patch of the same size, $(\bar{f}, \bar{I})$ correspond to their gray level mean, and m is the number of pixels in the fragment. Other similarity measures, such as SIFT, can also be used to allow some invariance to changes in scale and orientation [26]. A threshold is used so that the fragments may be considered as a binary random variable. Thus, a fragment f is considered present in the image and its value is set to 1 if its similarity measure score is higher than a predefined threshold, and to 0 otherwise. The joint probability distribution of the class label and fragment variables $P(C = c, F = f)$ is estimated to calculate the mutual information $I(C; F)$ between the fragment F and the class C of images, defined as:

$$I(C; F) = \sum_{c,f} P(C = c, F = f) \log \frac{P(C = c, F = f)}{P(C = c)P(F = f)} \tag{2}$$

IMAVIS 2692
11 April 2008 Disk Used
ARTICLE IN PRESS
No. of Pages 8, Model 5G

Table 1
Outline of the Algorithm

- **Training Stage:** Given a set of training set images extract a set of fragments for subsequent classification and model parameters
  - T1. Extract a candidate set of fragments: cut rectangular sub-images of different sizes and locations from the class images and compare these fragments to all training images based on the normalized cross-correlation (NCC) measure (see Eq. (1))
  - T2. For each candidate fragment, calculate the optimal detection threshold (θ), the mutual information (MI), relative position, and weight (W) according to Eqs. (2) and (3). The crucial difference between the sparse and compact feature families, is in the way the mutual information is evaluated. For distinctive features (multiple classification): compute MI using a binary class variable. For compact features (joint classification): compute MI using an n-value class variable
  - T3. Select a subset of non-redundant fragments by the max–min iterative optimization scheme (Eq. (4))
- **Recognition Stage:**
  - R1. Given a novel image, find all the fragments F_i within a search window with NCC(F_i) $\geqslant \theta_i$ (normalized cross-correlation exceeds detection threshold see Section 3.1.1)
  - R2. Combine detected fragments by summing their pre-determined weights and compare to threshold. By using different thresholds construct a receiver operating characteristic (ROC) curve (see Figs. 5–7)

The amount of mutual information depends on the fragment detection threshold. The detection threshold (θ) is therefore determined automatically to maximize the delivered mutual information [29,30]. In a similar manner, an optimal search window is selected for each feature, by searching for the window position and size that will maximize the mutual information of the feature [29]. At the end of this process each fragment has a detection threshold (θ), a mutual information (I), a weight (W), and an approximate location within the window of analysis. The fragment's weight is used later for classification, and it is defined as the log likelihood ratio:

$$W_i(F) = \log\frac{P(F|C_i)}{P(F|\overline{C})} \quad (3)$$

where $P(F|C_i)$ and $P(F|\overline{C})$ represent the detection frequency of the fragment in the class and non-class images and W_i is the weight of the fragment for class i.

In performing multiple individual classifications, optimal features are extracted for each class in turn. For class C_i, features are selected to maximize the mutual information between the set of features F and the class C_i. Here $C_i = 1$ if the image belongs to class i, and 0 otherwise. In this case, the class C_i contains different face images of the same individual (Fig. 3), whereas the non-class includes faces of all other individuals. The process is then repeated for all the different classes. In contrast, in extracting joint features, the feature F is sought to maximize the measure $I(C;F)$ for a multi-class variable C. As before, Eq. (2) is used to evaluate $I(C;F)$. However, in this case the class variable C has n rather than just two values. A vector of $W(F)$ is also computed for each fragment, computing a particular weight for each individual.

Intuitively, the multiple classification approach is expected to produce a sparse feature representation and the joint classification approach a compact representation, for the following reason. The first approach seeks for each class a subset of distinctive features that separate this particular class from all other classes, such as one particular face from all others. The resulting representation is sparse since such features would ideally be activated by a single individual, and different individuals would require different features. The joint approach seeks features that can make a useful separation between sub-classes. Ideally, each feature will separate the n classes into two equal subgroups. The resulting representation is compact in the sense that each feature will be activated by many different classes, but the joint activation of a small number of features will be sufficient for unambiguous classification.

3.1.2. Selecting a subset of non-redundant features

The second stage of the automatic fragment selection algorithm is based on a greedy iterative optimization scheme to select a subset of non-redundant features. The algorithm is a max–min iterative scheme [30] which was shown in comparative evaluations to produce a highly effective selection [31]. The algorithm goes over the initial pool of candidates denoted by P, in several steps. Each step moves the fragment that adds the largest amount of information from P, to the selected set of fragments constructed by the previous steps, denoted by S. The set S is initialized by selecting from P the fragment F_1 with the highest mutual information. At iteration step $k + 1$, the fragment F_{k+1} added to $S_k = \{F_1, \ldots, F_k\}$, can be formulated as:

$$F_{k+1} = \arg\max_{F_i \in P_k}\{\min_{F_j \in S_k}[I(C;F_j \cup F_i) - I(C;F_j)]\} \quad (4)$$

The idea behind this pairwise selection criterion is simple. For a fixed new fragment F_i, the term above measures how much information is added by F_i to that of a previous fragment F_j. For example, if F_i is very similar to a previous feature F_j, this addition will be small. The minimization over all the already selected fragments F_j guarantees that F_i is sufficiently different from all previous fragments. Finally, the maximization stage selects the new fragment i with maximal additional contribution. This max–min algorithm ends when the increase in information added by new fragments falls beyond a selected threshold (0.05) or when a maximum number of iterations is reached (1500). The final selected set of fragments S will be the output of the training stage, serving as the fragments for classification (Figs. 1 and 2). A very similar selection procedure is applied to both the sparse and compact feature families, and both use information for classification as a selection criterion. There is a basic difference, however, in the way the mutual information is evaluated (Eq. (2)), using the binary class variable for individual classification and the n-value class variable for the joint classification. This leads to the selection of different feature sets with different classification properties as discussed next.

The computational complexity of the learning stage is determined primarily by the number of images (T) in the training database, and their size (N) in pixels. The number of candidates in the initial fragment pool is proportional to TN. Each fragment is searched in the database by convolution, requiring time also proportional to TN. We assume that the maximal fragment size is $K \ll N$. The max–min computation can be performed efficiently [31], its computation time is small compared with the first selection stage, as is the selection of optimal thresholds. The overall complexity is therefore $O((TN)^2)$. In practice, we have used up to several tens of thousands candidates in the fragment pool and this computation is required once only during the learning of a new object class.

3.2. Performing classification

Both strategies perform classification of a new input image based on the fragments detected in the image. For a given fragment, its maximal NCC with the image is computed, and if it exceeds the fragment's detection thresholds (θ) within its detection window, then the fragment is considered to be detected in the image. The final decision is obtained by summing the weights (W_i) of all the detected fragments in the image. In the case of the compact

Please cite this article in press as: A. Akselrod-Ballin, S. Ullman, Distinctive and compact features, Image Vis. Comput. (2008), doi:10.1016/j.imavis.2008.03.005

strategy, the weight of the specific individual is taken from the weight vector. The computed sum is compared to a detection threshold. By using different thresholds we generate a receiver operating characteristic (ROC) curve, which presents both the hits (images detected correctly from the class) and false alarms (images detected incorrectly in non-class examples) of the classification (Figs. 5–7). The classification scheme was identical for the sparse and compact families, the only difference accounting for changes in the ROC was in the selected features. Other classification schemes, such as SVM [6,7] can also be used based on the selected features, but we found that the differences in classification performance between SVM and our scheme was small. By using a fixed classification scheme, but using different features, our comparison focuses on the main issue of interest – comparing the usefulness of the sparse and compact families for classification.

3.3. Details of implementation and testing

We implemented the sparse and compact feature extraction methods described above and compared them in the task of individual face recognition. The database included 500 faces of 25 individuals. The 20 images for each individual were divided to training and testing sets. Both the sparse and compact features were extracted from the same training images for all faces. The database was selected to include individuals corresponding to the artist's drawings. The images were taken from different internet sites, and were often of low quality. The images were cropped to exclude most of the background, and were normalized in size to 30 columns. This is above the minimal size (18 pixels in the horizontal dimension) required for reliable recognition by human observers [32]. The database described has significant variability in orientation, pose, facial expression, illumination, age, and artificial features (e.g. it includes faces with moustaches, beards, changes in hair styles, glasses, makeup, and the like).

For each individual, a set of 50 sparse features were extracted, yielding a total of 1250. The same total number of informative compact features was extracted as well. These features served as the total set of features to use for classification. In practice, as described below, classification performance often reached an asymptote with a smaller number of fragments.

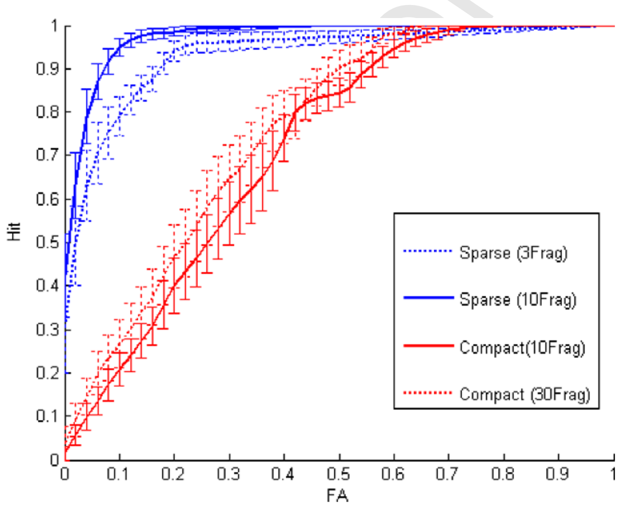

Fig. 5. Face recognition by sparse representation with either 10 (upper solid blue trace) or 3 fragments per face (lower dotted blue trace) and compact, with either 10, (lower solid red trace) or 30 fragments (upper dotted red trace). The ROC curves were averaged across faces for each strategy. Adding up to 1000 fragments to the compact family had a minor effect on the results.

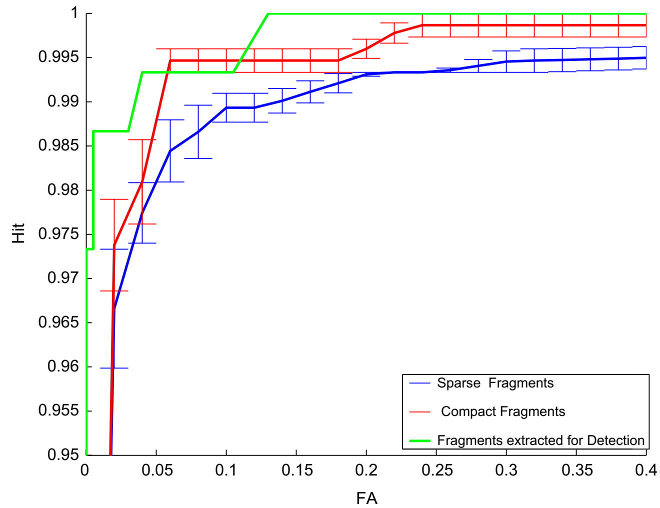

Fig. 6. Face Detection by sparse (the lowest blue curve), compact (red curve), and fragments extracted for face detection (green curve). ROC curves averaged over 5 experiments.

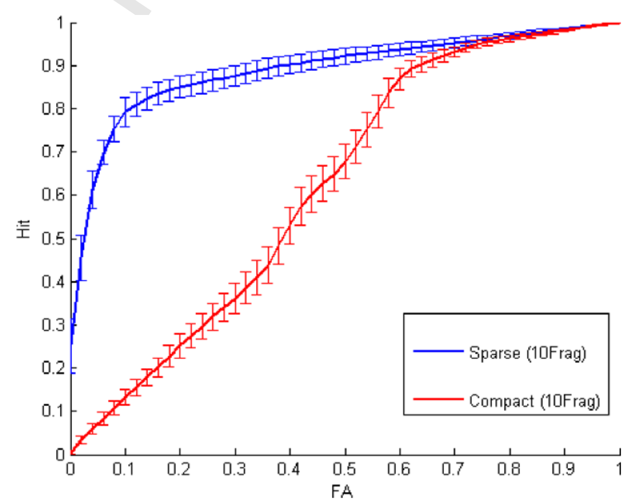

Fig. 7. Occluded images recognition by sparse and compact fragments. The ROC curves were averaged across 10 individuals for each strategy.

After the sparse and compact features have been extracted, we compared them in the following way. First, we compared the two families, to test whether the two approaches produce similar or different sets of features. Second, we tested the performance of each family on individual face recognition. Third, the features were compared in the task of general face detection, to test how they generalize from one task to a somewhat different one. Fourth, we compared the sparseness of the representations: ideal compact features are expected to be detected for about half of the faces in the new testing images, sparse features in only a small fraction. Fifth, we tested the recognition of occluded images by both strategies. Finally, we compared the sparse features detected automatically with distinctive features selected by an artist specializing in reduced face images. We summarize briefly below the method and parameters used in each of these tests.

3.3.1. Test 1: similarity of the two feature families Q3

We used for this testing the 100 most informative fragments from each family. As explained, features are detected in the image using normalized cross-correlation (NCC), and each feature has its own detection threshold. In comparing a compact feature FCi with a sparse feature FSj we therefore computed their NCC, and compared the result with the corresponding two thresholds, (θ_i, θ_j). If the NCC exceeded both thresholds, the features were considered as equivalent.

3.3.2. Test 2: individual face recognition

Both sets of extracted fragments, sparse and compact, were tested for face recognition of the individuals they were trained on. We used the test sets with novel images for all individuals and performed classification as described in Section 3.2, obtaining two sets of ROC curves, one for each individual and each strategy. The points along the ROC curve represent pairs of false alarms (FA) and hit rates. To infer statistical significance in comparing the schemes, the ROC's were averaged by dividing the FA into discrete bins and computing the average and standard deviation of the hit rates in these bins over the individual curves of each strategy. The recognition task was challenging compared with most past tests [5] for two reasons. First, multi-class recognition of similar classes is known from previous studies to be difficult [33–35]. Second, the images were highly variable, and only a small number of images were used for training.

3.3.3. Test 3: face detection

Both sets of extracted fragments were trained for the task of general face detection in images. We used for this task a new set of 300 face images and 400 non-face images, half of each set was used for training, the other half for testing. During training, new weights for the detection task were determined for each fragment according to the training results, as explained in Section 3.2 (Eq. (3)). From each set (sparse and compact fragments) we selected the 10 best features (highest information computed after training for detection) and compared the performance using ROC curves. To estimate statistical significance we repeated the experiment five times with different random division of the images into training and testing. The ROC's were averaged and the error bars represent the standard deviation obtained by the repeated experiments. The detection tests were always performed using a new set of test images (see Section 3.1).

3.3.4. Test 4: sparseness of the representation

The test compared the fraction of fragments from each family being activated on average by a face image. We used in the comparison the 60 best sparse fragments, compared with the 60 best compact fragments. The experiment evaluated the fragments on the individual test sets. The fragments were considered detected in the test database if their NCC exceeded the fragments threshold.

3.3.5. Test 5: individual recognition in occluded images

A set of masked images for 10 individuals was generated based on the original test sets. Square regions of different sizes (width ranging from 7 to 13 pixels) at different locations were masked in all the images of the test sets. The same classification procedure described above for the sparse and compact fragments was then applied as in test 2 to the masked images.

3.3.6. Test 6: comparing with artist's distinctive features

To compare the two representations, faces were divided into 8 regions. We then tested whether the artist and the automatic extraction method selected features in the same face sub-regions. The 10 best features (highest information for recognition) for each individual were selected from both the sparse and compact representation, for 10 different individuals. For each individual, a human observer made a binary decision of whether a feature was present in the artist representation. So, for example Lenon's mouth, Elton's nose and Sadam's eyes were considered missing in the image and their values were set to 0. We then tested the consistency between the artist's selection of features and the two extraction methods. A region was considered'inconsistent' if it contained a feature either in the automatic extraction or in the artist's image but not in both. This comparison was performed for the sparse and compact feature families, and the consistency fractions were averaged for each of the two representations.

4. Results

We tested by simulations the two families of visual features, and found that they produce different features with different classification properties.

4.1. Test 1: similarity of the two families

The features produced by the two strategies were significantly different. Comparing the 100 most informative features for the 25 individuals, only 35.5% of the sparse features were also included in the set of compact features. In general, the compact features are more similar to features commonly used in other face recognition schemes, with high proportion of the eyes region. The distinctive features are more variable, and extract idiosyncratic aspects of different faces. Figs. 1 and 2, illustrates several examples of these sparse and compact fragment areas. For example, see Sadam's hairline as opposed to his distinctive moustache fragment, W. Allen's compact fragment including hair,nose,cheeks and eyes as opposed to his very specific spectacles and nose distinctive fragment, S. Peres's nose and eyes as opposed to his distinctive forehead fragment, etc.

4.2. Test 2: individual face recognition

The sparse representation proved to be significantly better for face recognition compared with the compact representation (Fig. 5). Reliable recognition was obtained from 10 sparse fragments for each individual, and good results were obtained with as few as 3 fragments, showing that reliable identification of highly variable examples can be obtained using a small number of distinctive features. The graph in Fig. 5 shows the recognition performance for the sparse and compact features plotted as ROC curves. The sparse representation produced a significantly higher curve than the compact representation. Significant difference in all the following tests means ($p \leqslant 0.05$). Adding more compact features, up to 1000, has minor effect on the classification results. The distinctive features showed a similar advantage when performing n-class recognition, namely, when an input image is classified into one of n given classes.

4.3. Test 3: face detection

The compact features proved better than the sparse features in face detection, a related recognition task. Even better performance was achieved by a new set of features that were selected specifically for the general detection task (Fig. 6) using an identical training procedure with the class variable $C = 1$ for all face images. The sparse and compact ROC curves in this graph were obtained by averaging five repetitions of the experiments.

4.4. Test 4: sparseness of the representation

The representation produced by the multiple classification approach is significantly sparser: a face view activated on average 14.5% (s.d. 10.2) of the overall set of sparse features, compared with 50.0% (s.d. 7.1) of the features produced by joint classification,

Please cite this article in press as: A. Akselrod-Ballin, S. Ullman, Distinctive and compact features, Image Vis. Comput. (2008), doi:10.1016/j.imavis.2008.03.005

which is in agreement with the expected optimal probability of 0.5 for compact representation. A face view activates with high probability (74.12%, s.d. 20.7) the features in the subset of sparse features extracted for identifying this particular individual. The results illustrate that sparse fragments have a high detection probability in images of the face they were trained for, and low probability of appearing in other faces. The compact fragments have no systematic preferences.

4.5. Test 5: individual recognition in occluded images

The ROC curves shown in Fig. 7 were averaged over the individuals faces. As can be seen, the distinctive features are highly efficient in dealing with such occlusions compared with the compact features.

4.6. Test 6: comparing with artist's distinctive features

The features selected by the sparse representation method were similar to the reduced representation produced by an artist ([1], Fig. 1). The fraction of consistent regions in the sparse and artist representations was 0.88 (s.d. 0.10, 100 total fragments), significantly higher than the compact representation 0.65 (s.d. 0.14, 100 total fragments). Consistency for the compact features is not significantly higher than chance, but for the sparse features, although not perfect, it is highly significant. The comparison to the artist's features shows that the features selected automatically by the sparse method are similar to the distinctive features selected by the artist. The recognizability of the artist's renditions illustrate that these features are useful for robust identification from reduced and distorted input images.

5. Discussion

The present study compared two alternative feature selection strategies for the recognition of multiple similar classes: one uses multiple binary classifications, the other a single joint classification. The methods produce markedly different sets of features, one set is compact, with features shared by about half of the classes, the other extracts distinctive individual features. When applied to face images, our results show that the distinctive features are better for robustly recognizing a specific individual, and can compensate for distortions and missing information in the input images. This advantage of the sparse features is not expected a priori: although they are more informative individually, in the compact coding each class activates on average significantly more features, which may jointly perform better classification. The current testing focused on face features. A recent study showed the usefulness of distinctive features in the domain of cars as well as individual faces compared with other methods [36]. It will be of interest to compare in the future compact and distinctive features in other domains as well, particularly in tasks requiring distinctions between multiple similar classes.

The fact that individual distinctive features are superior for face identification is consistent with a large body of psychological research on face recognition [20,21]. Our method is the first to automatically extract distinctive face features for recognition. A previous method for defining what is distinctive in a face was based on deviations of the face contours from the average face [5,21] which does not capture the distinctive features extracted by our method. The selected distinctive features showed good agreement with the representation produced by an artist, and in both cases reliable identification could be obtained with a small number of features. We also found in comparisons that using the sparse features with different classifiers (e.g. SVM) produced only small changes in performance, probably because the distinctive features are highly informative by themselves and produce good separation between classes.

The results also show that different face features are better for different recognition tasks. Compact features performed better in general face detection. Features selected specifically for face detection achieved higher performance than either the sparse or compact set. Models of visual classification often assume the use of generic features, namely, a fixed set of features which are used for different classification tasks. Our results show that optimal classification features depend on the class as well as the specificity of the recognition task. Even within a single class such as faces, different feature types are required for generalizations at different levels, as opposed to general "face features" that are used for all tasks.

A major difficulty faced by any recognition approach has to do with image variability due to viewing conditions, noise, occlusion and the like. In the current study, the distinctive features were shown to be particularly useful in dealing with difficult variations caused by large distortions, highly reduced information, occlusion, aging effects and added artificial features. These advantages of distinct features were studied here in the context of face identification, but we expect that they will be applicable in other domains as well, such as identifying different cars, airplanes dogs, and the like.

Along with these advantages, it is important to note that the use of distinctive features also has limitations compared with the compact features, which are shared by multiple classes. A number of studies [35,37–39] have shown how the extraction and use of shared features can be useful for generalization and the fast learning of new object classes. The distinctive features proved relatively insensitive to illumination to changes and some rotation in space, but more complex features were shown in previous studies to allow larger changes in viewing angle [37] and scale [38].

An intriguing question for future study is therefore the optimal combination of different feature types within an overall recognition scheme. Such a combined scheme, which has not been developed so far, could use the relative merits of distinctive and compact features, to obtain high discrimination and robustness together with broad generalization and the fast learning of new object classes from limited data.

Acknowledgments

We thank H. Piven for his help, and the excitement he added to this work. This work was supported by ISF Grant 7-0369 and by EU IST Grant FP6-2005-015803.

References

[1] H. Piven, Piven in America, Am Oved Publishers Ltd., Tel-aviv, 2002.
[2] A. Samal, P. Iyenger, Automatic recognition and analysis of human faces and facial expression: a survey, Pattern Recognit. 25 (1992) 65–77.
[3] R. Chellapa, C. Wilson, S. Sirohey, Human and machine recognition of faces: a survey, Proc. IEEE 83 (1995) 705–741.
[4] M.H. Yang, D.J. Kreigman, N. Ahuja, Detecting faces in images: a survey, IEEE PAMI 24 (2002) 34–58.
[5] W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld, Face recognition: a literature survey, ACM Comput. Surv. (2003) 399–458.
[6] E. Osuna, R. Freund, F. Girosi, Training support vector machines: an application to face detection, CVPR (1997) 130–136.
[7] P.J. Phillips, Support vector machines applied to face recognition, Adv. Neural Inform. Process. Syst. 11 (1998) 803–809.
[8] M. Turk, A. Pentland, Face recognition using eigenfaces, CVPR (1991) 586–591.
[9] M.S. Bartlett, T.J. Sejnowski, Viewpoint invariant face recognition using independent component analysis and attractor networks, Advances in Neural Information Processing Systems, vol. 9, The MIT press, 1997. p. 817.
[10] R. Fergus, P. Perona, A. Zisserman, Object class recognition by unsupervised scale-invariant learning, CVPR (2003).
[11] P. Belhumeur, J.P. Hespanha, D. Kriegman, Eigenfaces vs Fisherfaces: recognition using class specific linear projection, IEEE PAMI 19 (1997) 711–720.

Please cite this article in press as: A. Akselrod-Ballin, S. Ullman, Distinctive and compact features, Image Vis. Comput. (2008), doi:10.1016/j.imavis.2008.03.005

[12] T. Kanade, Computer recognition of human faces, Bazel, Birkhauser, Switzerland, and Stuttgart, Germany, 1973.
[13] R. Brunelli, T. Poggio, Face recognition: features vs. templates, IEEE PAMI 15 (1993) 1042–1052.
[14] A. Lanitis, C.J. Taylor, T.F. Cootes, Automatic interpretation and coding of face images using flexible models, IEEE PAMI 19 (1997) 743–756.
[15] A.L. Yuille, D.S. Cohen, P.W. Hallinan, Feature extraction from faces using deformable templates, Intl. J. Comput. Vis. 8 (1992) 99–112.
[16] H. Rowley, S. Baluja, T. Kanade, Neural network based face detection, IEEE PAMI 20 (1998) 23–38.
[17] P. Viola, M. Jones, Robust real-time object detection, Intl. J. Comput. Vis. 57 (2) (2004) 137–154.
[18] L. Wiskott, J.M. Fellous, N. Kruger, C. von der Malsburg, Face recognition by elastic graph matching, IEEE PAMI 19 (1997) 775–779.
[19] H. Schneiderman, T. Kanade, A statistical method for 3d object detection applied to faces and cars, CVPR (2000) 746–751.
[20] V. Bruce, A.M. Burton, N. Dench, What is distinctive about a distinctive face?, Quart J. Exp. Psych. 47 (A) (1994) 119–141.
[21] G. Rhodes, Superportrait Caricatures and Recognition, Hove Psychology Press, 1997.
[22] P.G. Schyns, F. Gosselin, Bubbles: a technique to reveal the use of information in recognition tasks, Vis. Res. 41 (2001) 2261–2271.
[23] D.J. Field, What is the goal of sensory coding?, Neural Comput 6 (1994) 559–601.
[24] E. Sali, S. Ullman, Combining class-specific fragments for object recognition, BMVC (1999) 203–213.
[25] S. Ullman, E. Sali, M. Vidal-Naquet, A fragment-based approach to object representation and classification, IWVF4 (2001).
[26] D. Lowe, Distinctive image features from scale-invariant key-points, Intl. J. Comput. Vis. 60 (2004) 91–110.
[27] M. Vidal-Naquet, S. Ullman, Object recognition with informative features and linear classification, ICCV (2003).
[28] N. Vasconcelos, M. Vasconcelos, Scalable discriminant feature selection for image retrieval and recognition, CVPR (2004).
[29] B. Epshtein, S. Ullman, Feature hierarchies for object classification, ICCV (2005).
[30] S. Ullman, M. Vidal-Naquet, E. Sali, Visual features of intermediate complexity and their use in classification, Nat. Neurosci. 5 (2002) 682–687.
[31] F. Fleuret, Fast binary feature selection with conditional mutual information, J. Mach. Learn. Res. 5 (2004) 1531–1555.
[32] T. Bachman, Identification of spatially quantized tachistoscopic images of faces: how many pixels does it take to carry identity?, Eur J. Cogn. Psychol. 3 (1991) 87–103. Q1
[33] A.C. Berg, T.L. Berg, J. Malik, Shape matching and object recognition using low distortion correspondences, CVPR (2005).
[34] A. Holub, P. Perona, A discriminative framework for modelling object classes, CVPR (2005).
[35] A. Torralba, K.P. Murphy, W.T. Freeman, Sharing features: efficient boosting procedures for multiclass object detection, CVPR (2004) 762–769.
[36] B. Epstein, S. Ullman, Satellite features for the classification of visually similar classes, in: Proceedings of the IEEE CVPR, 2006, pp. 2079–2086.
[37] E. Bart, E. Byvatov, S. Ullman, View-invariant recognition using corresponding object fragments, Proc. ECCV 2 (2004) 152–165.
[38] T. Serre, L. Wolf, T. Poggio, Object recognition with features inspired by visual cortex, CVPR (2005).
[39] L. Fei-Fei, R. Fergus, P. Perona, A bayesian approach to unsupervised one-shot learning of object categories, ICCV 2 (2003) 1134–1141.