ABSTRACT

Work presented suggests a combined informational space and decision rule for recognition of 3-D objects. The informational space consists of heterogeneous sets of features (i.e. belonging to different spaces), that are object images, images of certain object features and 3-D object surface representation. Decision rule for recognition in this combined space is proposed. The method was tested on a database of human face stereo-images and gave a significant improvement of reliability of automatic recognition system.

Keywords: computer vision, image recognition, principal component analysis.

1. INTRODUCTION

The problem of automatic real time recognition of 3-D objects is urgent for high performance computer vision systems. One of the fields of recognition is face recognition. Principal components analysis is a widely used method. Usually, this method is implemented to informational space constructed on the basis of image pixel values (signal or brightness). Meanwhile, fast and reliable stereo-reconstruction techniques have been developed. Hence the problem is to implement common-use algorithms of recognition to the informational space that combines 2-D and 3-D data.

Classic problem of image recognition may be defined as follows. Objects of recognition are given as images and image is a set of features or vector in a multidimensional space. Training set of images where for every image it’s object is known is inputted in the algorithm. In other words, if all objects are divided into classes, one will know which class every image belongs to. On the basis of this information, algorithm of recognition is generated. Algorithm should define which class (or classes) new input image belongs to.

Images usually consist of raw and bulk sets of features. It can be raster photo-image, digitized audio signal, etc. Usually this data is inconvenient for direct use in recognition algorithm. Informative features optimization problem occurs on a stage of image preprocessing, therefore before recognition process itself. Preprocessing is intended to solve two main problems: clustering and features selection. Clusterization is a grouping of vectors representing objects of the same class. The results of clusterization are minimizing inner-class distances and maximizing inter-class distances. Effective features selection may dramatically reduce number of dimensions of image space. Obviously this results in informativity loss, however optimal selection of features minimizes these losses and even may improve recognition results in the case of limited time and computational cost. The point is while reduction of number of features and accordingly reduction of processing time are substantial, informativity loss is very low. Optimal features selection should maximize reduction of number of features and minimize informativity loss.

Possible optimization methods can be based on signal filtering, using a priori knowledge of objects of recognition, statistical analysis, etc. Principal Components Analysis (PCA) is an effective statistical method of informative features optimization. It is widely applied in image recognition, particularly in face recognition.3,4,5 This method has a characteristics of clusterization and optimal feature selection, because it minimizes original image reconstruction error. In this work an extension of PCA is proposed. This extension can be useful for compound images consisting of images.
of different nature. The method is applied to normalized images, i.e. images after preprocessing. Current system uses stereo-images and elevation maps (3-D surface images) reconstructed from stereo-images (Figure 1) Hereinafter we use the term 'stereo-pair' to denote pairs of photo- images taken at the same moment in time from a slightly different points of view.

2. IMAGE SPACE

Image recognition problem deals with visual, audio or other type of information. In the case of visual information input image is a vector function of two coordinates \( B(x, y) \). \( B \) is a vector that may consist of brightness, hue, saturation, relief in a 3-D case and the like. Current system works with stereo-images and reliefs (elevation maps), reconstructed from these pairs. Also, bound the problem to discrete case, that is most important for automatic recognition systems. Thus, \( x \) and \( y \) take integer values in a limited range and image is a matrix which elements also take discrete or integer values in a limited range:

\[
I = \begin{bmatrix}
B_{11} & B_{21} & \cdots & B_{m1} \\
B_{12} & B_{22} & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
B_{1n} & \cdots & \cdots & B_{mn}
\end{bmatrix}
\]

That 2-D matrix may be presented as 1-D vector:

\[
\tilde{I} = \begin{bmatrix}
B_1 \\
B_2 \\
\vdots \\
B_N
\end{bmatrix}
\]

where \( N = m \times n \). Image space has \( N \) dimensions. Variety of images of all objects \( I_F \) is a subset of a set of all possible images or image space \( I \) (Figure 2).
Figure 2. Set of objects’ images \( I_f \) in the image space \( I \).

Objects of recognition in our concern are human faces. Let us consider possible variations of features relating to one object of recognition. They are caused by different light conditions, different position of object, including shifting along \( x, y, z \) axes (\( z \)-shifting or zooming is equivalent to scaling, if size of object is small relatively to distance from camera) and rotating around this axes (tilt, turn, rotation), different facial expression, some features, changing with time (hairs, closing part of face, glasses, mustache and beard, etc.). Some of them may be and actually are compensated by special techniques. For example finding face border and particular face features can fully compensate shifting and scaling, and even rotation. However, tilt and turn are harder to compensate. Different light conditions can be partly compensated by special filters: normalizing brightness and/or contrast, equalizing and so on. Selection of algorithms applied depends on implementation. Detailed discussion of these methods is beyond the scope of this article. We should note however, that in terms of optimization these methods clearly serve as clusterization, feature selection or both. For example, “finding face” procedure excludes features not related to object of recognition, and compensates differences between images, in which object is positioned differently. Hence the procedure serves as clusterization and feature selection. But the result of these procedures is still a raster image that is a bulk data and this image is still subject to further preprocessing (Figure 1).

3. RECOGNITION IN A SINGLE IMAGE SPACE

Usually similar objects (for instance, faces) have many generic features and their images differs weakly if compared to differences among all possible objects. In this case set of objects’ images \( I_f \) is a very narrow subset of \( I \) and it can be concluded that image space is far from optimal for describing objects in terms of their images. One can present input raster images (vectors of a huge number of dimensions) using vectors of lesser dimensionality. This is possible because: a) highly correlated features duplicate information, thus some of the features may be omitted; b) features that do not change significantly while shifting from one object to another also may be omitted as they don’t yield substantial information. While dimensionality of image space reduces, image processing time and quantity of saved data decreases but informativity does not fall significantly.

PCA, based on Karhunen-Loeve expansion, is an approach to reducing the dimensionality of image space in such a way that a basis in a new space reflects properties of a variety of recognized classes in optimal way. It has following optimal properties: a) it minimizes error of approximation, thus working as optimal features selection (this property assures that the error of reconstruction by any fixed number of components is the smallest possible among any reconstructions made by the same number of components.) and b) it shows behavior typical to clusterization. Karhunen-Loeve expansion yields statistically uncorrelated components. These components are calculated as eigenvectors of autocorrelation matrix. That is why they are called further just as ‘eigenvectors’. Eigenvectors corresponding to maximal dispersion of training image set are called principal components (PC). Choosing principal components for representation of faces provides first optimal property of PCA.

Let us illustrate how PCA works for object recognition. Optimal implementation of PCA presumes, that mean of all images is zero vector, hence mean vector is calculated for training set and then subtracted from all images treated in this set. \textit{A priori} one knows only images, constituting training set, so it is reasonable hypothesis to suppose, that mean
of all images is equal (or close enough) to mean of training set images (Figure 3). Principal components calculated in image spaces of photo-images and elevation maps are shown on Figure 4 and Figure 5 respectively. Every image (photo-image or elevation map) is represented in the basis of corresponding principal components. Vectors in this presentation are input data for decision rule. Performance of decision rule algorithm is very high, since dimensionality of vectors is very small by proposed procedure and distances in principal components spaces are calculated very fast.

Figure 3. Mean vectors of training sets of photo-images and elevation maps respectively.

Figure 4. First eigenvectors of photo-images – principal components.

Figure 5. First eigenvectors of elevation maps – principal components

Every processed image is represented in the principal components basis. Principal components space $I_{PC}$ is a subspace of image space $I$. This implies that vector representing an arbitrary recognized image can be situated beyond the principal components space. Thus, vector reconstructed using Karhunen-Loeve expansion can differ from original vector. If this difference is too high one can make a decision that present image does not belong to the variety of recognized classes.

Following recognition scheme for one image space is proposed by Turk and Pentland.\(^3\) There are four possible situations for every vector in an image space. They are presented in the following table and illustrated in Figure 6:

<table>
<thead>
<tr>
<th>PC Space $I_{PC}$</th>
<th>Known images</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>near</td>
<td>near</td>
</tr>
<tr>
<td>2</td>
<td>near</td>
<td>far</td>
</tr>
<tr>
<td>3</td>
<td>far</td>
<td>near</td>
</tr>
<tr>
<td>4</td>
<td>far</td>
<td>far</td>
</tr>
</tbody>
</table>
4. RECOGNITION IN MULTIPLE IMAGE SPACES

Current version of system implements PCA in spaces of photo-images of objects and their relief maps. So, the implementation of proposed recognition scheme is not straightforward as it is in the case of one space.

Obviously, heterogeneous images (i.e. images belonging to different spaces) have different statistical characteristics. PCA distinguishes first eigenvectors as vectors corresponding to maximal dispersion of image set. If one will use vectors consisting of features from heterogeneous images, first principal components will reflect structure of a homogeneous subspace, that has maximal dispersion of features. Structure of other subspaces will be reflected by vectors of higher order and this structure will be hidden by noises of the former subspace. That’s why it is undesirable to use PCA on vectors consisting of heterogeneous features (for instance, vector that includes photo-image and relief simultaneously).

Thus, heterogeneous images are processed separately, i.e. for each type of them separate principal components space is built. Problem arises of synthesizing decision rule, that will consider information obtained in different image spaces. The following measure is proposed:

\[
R_{c,v^k}(c,v^k) = \sum_i W_i \frac{1}{R_i^2(c,v^k)}.
\]

(1)

where \(R_i\) is a measure (Euclidean distance) in \(i\)-th image space, \(c\) is a compound image to be classified (image consisting of heterogeneous images is called here compound image), \(v^k\) is \(k\)-th compound image of training set, \(W_i\) are weights of heterogeneous spaces, calculated as:

\[
W_i = \frac{D_i}{D_i^c}.
\]

(2)

where \(D_i\) is dispersion of all images in \(i\)-th space and \(D_i^c\) is the averaged dispersion by classes in \(i\)-th space. \(R(c,v^k)\) can be considered as a normalized distance for compound images. In terms of normalized distance decision rule may be defined as:

\[
N = \arg \min \{ R(c,v^1), \ldots, R(c,v^K), T \}.
\]

(3)

In case that \(T\) is not the minimum, \(N\) is a number of image, closest to classified image \(c\). If distance exceeds threshold \(T\), then it is concluded that image \(c\) does not belong to any of known classes. Otherwise classified image is considered to belong to the class that closest image belongs to. The level of the threshold depends on implementation. For example, the use of threshold is unnecessary if one knows a priori that image does belong to one of existing classes.
As one can see, a direct sum of measures (or squares of measures) in separate image spaces was not used. It is made by the following reason. Consider heterogeneous image of object, consisting of pair of photo-images. Assume first image of pair is in situation 1 of one-image-space recognition scheme (see Figure 6) and second is in situation 2. Then for direct sum most of it will fall to the share of second “bad” image. As for the reciprocal, used in measure, the main share will be of first “good” image and second image will not worsen this measure too much. This argument is based on the experimental knowledge that usually there are two different situations for images of the same class and for images of different classes. Images of the same class sometimes happen to be very close to each other. As for the images of different classes, they may be situated near each other, like most of the same class images, but not so close as the same class images happen to be. The measure built must not reduce the effect of this behavior and the measure even strengthens this effect, as it increases the probability of considering of such “good” images.

Another improvement of measure, made recently, is that measures $R_i$ are taken separately. Then formula looks like:

$$
\frac{1}{R^2(c,k)} = \sum_{i} W_i \min_{j} \left[ R^2_i(c_j, v^{i,j}) \right].
$$

(4)

Expression in the denominator means that minimum of measure $R_i$ is taken by all known images of given class in $i$-th image space. Thus, measure of image is taken relative to whole class of images, rather than to single compound image. This way the distance from class can be defined, rather than distance from image. As experiments show, implementation of this measure increases reliability of recognition.

The measure was also implemented for continuous frame-grabbing, that gives a sequence of images (maybe, from a single camera) that changes in a given period of time, rather than a set of images from different points of view, but in a single moment in time. The measure is calculated for a set of images, taken at different moments. Weight was revised to take into account different significance of moments of shooting simply by multiplication coefficient. For instance, images, shot just before processing are more significant than those shot some moments ago. The measures of “just-shot” images are added to the sum and decision rule is applied to corrected normalized distance. This allows to proceed with recognition in real-time mode using previously obtained information if first compound images does not yield reliable information about object of recognition.

5. DISCUSSION

Sets of images of about 100 stereo-images (that is 200 photo-images and 100 elevation maps) were used for determining optimal number of dimensions of Principal Components spaces in image spaces of photo-images and relieves. Number of Principal Components required for face recognition was determined to be 20-30 vectors in every image space. Thus, number of features was reduced in hundreds of times relative to raw input data quantity of about $10^4$-$10^5$ values for each image. This allows to classify images on large databases in a real-time mode and store processed images in a very compact form.

Experiments show that image preprocessing is of crucial importance for reliability of recognition. Therefore one should thoroughly choose and implement preprocessing methods before using our algorithm. Different methods supposed to have clusterization and feature selection characteristics were applied. They are: a) finding face border and orientation on photo-image with following procedures of clipping, scaling, shifting and rotating of original image, b) normalizing of brightness and contrast. The possibility of enhancing the face location procedure by adding algorithms of finding important face features as eyes, nose, mouth is studied.

Compound images consisting of pair of photo-images and elevation map, reconstructed from this pair are used for recognition in the system discussed. Also sets of images are used, consisting of frames, shot at different moments of time. It is possible to extend method by adding other types of images to the structure of compound image. It can be images of the most informative parts of human faces, particularly eyes and separately nose and mouth or face sketch which is a picture of edges obtained from original image by applying appropriate filters.

Recognition using compound images was tested on the database of about 600 stereo-images of 200 persons and recognition accuracy achieved was about 95%. This is about two times more reliable than using simple photo-images. Most wrong cases were due to difference in angle of view or facial expression.
6. ACKNOWLEDGEMENTS

The authors wish to thank Serge L. Chernoff for his thought-provoking discussions and Serge A. Zorin for his help in laboratory system development. Also, we would very much like to thank Vladimir I. Tsurkov for his supervision.

7. REFERENCES


