# Image normalization by mutual information

Evgeniy Bart* and Shimon Ullman
Dept. of Computer Science and Applied Mathematics
Weizmann Institute of Science
Rehovot, Israel 76100
{evgeniy.bart, shimon.ullman}@weizmann.ac.il

**Abstract**

Image normalization refers to eliminating image variations (such as noise, illumination, or occlusion) that are related to conditions of image acquisition and are irrelevant to object identity. Image normalization can be used as a preprocessing stage to assist computer or human object perception. In this paper, a class-based image normalization method is proposed. Objects in this method are represented in the PCA basis, and mutual information is used to identify irrelevant principal components. These components are then discarded to obtain a normalized image which is not affected by the specific conditions of image acquisition. The method is demonstrated to produce visually pleasing results and to improve significantly the accuracy of known recognition algorithms.

The use of mutual information is a significant advantage over the standard method of discarding components according to the eigenvalues, since eigenvalues correspond to variance and have no direct relation to the relevance of components to representation. An additional advantage of the proposed algorithm is that many types of image variations are handled in a unified framework.

## 1 Introduction

Image normalization refers to eliminating image variations (such as noise, illumination, or occlusion) that are related to conditions of image acquisition and are irrelevant to object identity. The goal is to obtain a standard image with no artifacts arising from the specific conditions in which a particular image was taken. For example, illumination should be neutral, and no noise should be present. Two types of variations and the desired normalization results are shown in Figures 1, 3.

The need for image normalization arises for several reasons. First, image variations described above severely interfere with tasks such as object recognition (e.g. [1]); image normalization can therefore be a useful preprocessing stage for these tasks. In addition, image normalization can be used to facilitate human object perception, for example, to assist recognition.

*http://www.wisdom.weizmann.ac.il/~evgeniy

BMVC 2004 doi:10.5244/C.18.35

In this paper, we present a technique for class-based image normalization. In our scheme, objects are represented in the basis of principal components. Mutual information is used to identify irrelevant components. These components are discarded to obtain a normalized image which is not affected by irrelevant variations. Mutual information provides a principled way to evaluate the relevance of a particular component to the representation, in contrast to eigenvalues used by the previous schemes. Eigenvalues determine the amount of variance accounted for by each component, and have no direct relation to the relevance of the component to representation.

The remainder of this paper is organized as follows. In the next section, relevant previous approaches to image normalization are reviewed. In section 3, we describe the proposed normalization scheme in more detail. Experimental evaluation of the scheme is presented in section 4. The method is demonstrated to produce visually pleasing normalized images and to improve significantly the accuracy of recognition. These results and future directions are discussed in section 5.

# 2 Existing normalization methods

Principal component analysis (PCA) is frequently applied to various image processing tasks. However, its use has been based on empirical observations related to the magnitude of the eigenvalues. This magnitude corresponds to the variance accounted for by the corresponding component. For example, a common technique for eliminating small random noise is to discard principal components with smallest eigenvalues. Eliminating three components with *largest* eigenvalues has been used (e.g. [4]) to handle illumination variations. This method has been theoretically justified in [18] for the case of images uniformly sampled from the viewing sphere. However, in general, eigenvalues need not correspond to the relevance of the corresponding components to representation. For example, if the variance of noise is larger than the variance of some significant components, these components will be removed along with the noise. Sampling of illumination directions is unlikely to be uniform, and in this situation top three eigenvalues will no longer correspond to illumination (see section 4.2). In contrast, mutual information provided by each component is a principled measure of its relevance to the representation.

Two important sources of image variability are random noise and illumination. Below, algorithms specialized to handling either of these tasks are reviewed.

## 2.1 Random noise

A standard noise reduction technique is filtering by mean, Gaussian, or median filters [13]. Additional popular methods include the use of partial differential equations [8] or variational techniques [7]. Independent components of the data are found in [12], and components with small variance are discarded. A disadvantage common to the approaches above is that they can only handle relatively small noise levels. When the magnitude of noise is comparable to the magnitude of salient features in the image, these methods will remove meaningful image parts along with the noise.

### 2.2 Illumination

Humans perceive patches of similar brightness as similar despite significant changes in illumination level. Several algorithms [14, 9] mimic this human ability (called brightness constrancy). However, these algorithms cannot fully overcome changes in illumination direction.

Fischer's linear discriminant (FLD) is a general method to identify a set of directions that best separate the given classes. It has been used for face recognition across changes in illumination [4, 27]. A drawback of this technique is that for $c$ classes, it identifies at most $c-1$ directions. When this number is insufficient, alternative methods must be employed to find the remaining directions and determine their relevance. In addition, FLD performs poorly when image should be reconstructed after removing illumination effects. The reason is that image pixels that are highly influenced by illumination are deemphasized by FLD so that they do not affect recognition. As a result, illumination effects are not removed from these pixels upon reconstruction, which produces visually disturbing artifacts. Our experiments confirmed that FLD is not able to achieve visually pleasing reconstruction while reducing illumination effects.

Under the commonly used Lambertian reflectance model, the effects of illumination are linear, with clamping at shadowed points. This allows to synthesize images of an object under arbitrary illumination by linear or convex combinations of several basis images of the same object. A drawback of this approach is the necessity to have several images of each object, taken in controlled conditions, to bootstrap the synthesis. For example, algorithms that use illumination cone [3, 6, 11, 16] require three images without shadows per object. In addition, the computations in illumination cone are time-consuming. The quotient image method [19, 24] uses class-based knowledge to perform synthesis from a single novel image; however, three images per object are still required for training. In addition, the illumination in training images should be the same for all objects. Similar constraints are present in [21, 23, 20, 17, 2, 10, 25]. In contrast, the scheme proposed below can handle uncontrolled training databases with a single image in random illumination per object.

For bilaterally symmetric objects such as faces, the constraint of having multiple training images can be relaxed [26] by reflecting each image horizontally and using this reflection as an additional training image. However, the method is not robust, and the use of a 3D face model is necessary to achieve reasonable reconstruction [26]. In addition, this method does not apply to more general, non-symmetric objects.

## 3 Image normalization by mutual information

In this section, the proposed scheme of image normalization by mutual information is described. The scheme consists of calculating the basis of principal components for the given set of images and discarding the uninformative components. These steps are detailed next.

We assume that a set $\{I_k\}_{k=1}^{N}$ of $N$ training images is given. In addition, class label for each image should be provided by a class variable $C_k$ such that $C_k = c$ if image $k$ belongs to class $c$. For example, the images may depict human faces, and the labels may represent the identity of the person in the image.

Principal components are calculated by performing singular value decomposition of the matrix $M = [\widetilde{I}_1 \ldots \widetilde{I}_N]$ that consists of centered training images: $\widetilde{I}_k = I_k - \mu$ (here $\mu$ is the mean image: $\mu = \frac{1}{N}\sum_k I_k$). The images here are regarded as column vectors of dimension $d$, where $d$ is the number of pixels. $N-1$ principal components will usually be obtained; these $N-1$ components together with the mean image $\mu$ span the original $N$ images.

Each principal component represents some image feature. Some of these features pertain to the objects depicted in the images, and others represent irrelevant sources of image variability. It is natural to use mutual information [5] to estimate the relevance of a particular feature to the object. The intuition is that a relevant feature will provide information about the object's identity and will have high mutual information. In contrast, an irrelevant feature will be independent of the object's identity and will have low mutual information.

To calculate mutual information of a given feature $F$, the joint probability distribution of class label and feature strength, $p(C=c, F=f)$, has to be estimated. The strength of a feature in an image is measured by the projection of the image onto the corresponding principal component. Denoting the image by $X$ and the principal component that corresponds to feature $F$ by $P_F$, the projection is $\langle X, P_F \rangle = P_F^T X$. This is a continuous value, and therefore $p(C=c, F=f)$ is a continuous probability density function. In principle, $p(C=c, F=f)$ can be estimated from the training images and mutual information can be calculated from it. However, estimating a continuous function is likely to require a large number of training examples. To reduce the required amount of training data, a discretization of the values $F$ assumes can be introduced. In the experiments described below, a threshold was used to make features binary. The value of the feature $F$ was set to 1 if $\langle X, P_F \rangle > \theta_F$ and to 0 otherwise. The resulting discrete distribution $p(C=c, F=f)$ can be estimated reliably from a reasonable number of training examples. Mutual information can be expressed in terms of this distribution as follows:

$$I_{\theta_F}(C;F) = \sum_{c=1}^{K} \sum_{f=0}^{1} p(C=c, F=f) \log \frac{p(C=c, F=f)}{p(C=c)\,p(F=f)}, \tag{1}$$

where $K$ is the number of classes. Note that this expression implicitly depends on the threshold $\theta_F$ because the 0/1 value the feature assumes is determined by $\theta_F$. Therefore, it is natural to select the threshold $\theta_F$ for the feature $F$ that maximizes the feature's mutual information: $\theta_F = \arg\max_\theta I_\theta(C;F)$. In the experiments described below, optimal thresholds were selected automatically in this manner.

After the mutual information for each principal component is computed, non-informative components can be discarded to achieve normalization. The results of this normalization by mutual information are show in the next section.

# 4 Experiments

The AR database [15] of face images was used in our experiments. A subset of 83 subjects photographed without eyeglasses was selected. The images were manually aligned and the central face region was cropped. The resulting cropped images were converted to grayscale and the size was reduced to $90 \times 90$ pixels. Examples are shown in Figures 1, 3. Several experiments with different kinds of normalization are summarized below.


(a) MI (b) Original (c) Magnified

Figure 1: Gaussian noise experiment. In (a): mutual information of principal components. Horizontal axis: component index; vertical axis: mutual information. The components were sorted by decreasing mutual information. In (b), left to right: original image, image corrupted by Gaussian noise, normalized image. In (c): a magnified detail of (b).

## 4.1 Gaussian noise

For this experiment, original images were corrupted by additive Gaussian noise of mean zero and standard deviation 8 (intensity values of the images ranged from 0 to 255). An example is shown in Figure 1. Five different noisy samples were created from each original image, resulting in 415 images. 414 principal components were calculated from these 415 noisy samples. Mutual information of each component is shown in Figure 1(a). As can be seen, a large drop in mutual information (defined as $(I_k - I_{k+1})/I_k$, where $I_k$ is the mutual information of the $k$'th component) occurs at 82'nd component[1]. This suggests a natural number of principal components to retain. Therefore, components 83 and beyond were discarded, and noisy samples were reconstructed from 82 remaining components. The results of the reconstruction are shown in Figure 1.

In addition, performance on a recognition task was evaluated before and after normalization. The standard algorithm of recognition by PCA [22] was applied to normalized images. For comparison, performance was also evaluated using non-normalized noisy images. In addition, the standard method of discarding principal components with smallest eigenvalues was tested; we refer to this method as 'normalization by eigenvalues'. The number of discarded components was the same as in normalization by mutual information. The performance of the three methods is summarized in the first row of Table 1. As can be seen, Gaussian noise is a relatively simple case of image corruption, and can be handled by standard techniques. However, this example illustrates that the proposed method can handle many types of image variability in a single framework. More challenging examples are described next.

## 4.2 Illumination: four samples

Face images taken with either left or right flash on (Figure 3(a)) were used in this experiment. Four samples per subject (two with right flash and two with left flash) were used, resulting in 332 images. 331 principal components were calculated from these images, and uninformative components were discarded. The components to discard were again

[1] The reason that 82 informative components are present is that these 82 components and the mean image roughly span the original 83 faces used for this experiment. The remaining components represent random noise and are uninformative.

| | Original | MI | Eigenvalues |
|---|---|---|---|
| Gaussian noise | 100 | 100 | 100 |
| Illumination | 12 | 99 | 12 |
| Occlusion | 14 | 71 | 14 |

Table 1: Recognition accuracy (in %) on original images (first column), on images normalized by mutual information (second column), and on images normalized by eigenvalues (third column). First row: Gaussian noise; second row: illumination; third row: occlusion.

Figure 2: Principal component corresponding to illumination. From left to right: images obtained by adding progressively larger multiples of the component to the mean image.

determined by the location of the largest drop in mutual information. Only one uninformative component was found in this experiment. This component is shown in Figure 2. (Three components corresponding to illumination are predicted in [18]; the reason this prediction is violated is that illumination directions are not sampled uniformly in the data set used for this experiment.) Original images were reconstructed from the remaining components. The results of the reconstruction are shown in Figure 3(b). As can be seen, the proposed method can effectively remove illumination effects to obtain an image with neutral illumination.

Conceivably, a simpler way to remove illumination effects could be to compute a PCA basis of neutrally illuminated images (like the leftmost image in Figure 1(b)) and project the images with flash (Figure 3(a)) onto this basis. However, since neutral images do not incorporate information about the structure of illumination, projections of differently illuminated images will be significantly affected, and reasonably accurate reconstruction will be impossible. Our experiments confirmed that reconstruction quality achieved by this method is not acceptable. (Results are not shown due to disturbing visual artifacts.)

As before, performance on a recognition task was evaluated before and after normal-

(a) Original

(b) Normalized

Figure 3: Illumination experiment with four samples per face. (a): original, (b): normalized images.

(a) (b)

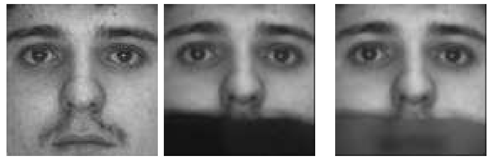
Figure 4: Illumination experiment with a single sample per face. (a): original (left) and normalized (right) images. (b): same as (a) for a different face.

ization and is summarized in Table 1 (second row). As can be seen, selecting principal components by mutual information rather than by eigenvalue significantly improves the results. The technique of eliminating three principal components with largest eigenvalues [4] was also evaluated. The accuracy achieved by this technique was 90%. The performance is poorer because only one component responsible for illumination was present in the data set; as a result, two useful components have been removed along with the irrelevant one. The reason is that eigenvalues cannot discriminate between relevant and irrelevant components. In contrast, the exact number of irrelevant components has been identified by mutual information, and better performance (99%, Table 1) has been obtained.

A number of alternative algorithms described in section 2 (e.g. [19]) can obtain illumination-normalized images (as in Figure 3(b)) when multiple samples of each face are available. Next, we describe how similar results can be obtained by the proposed algorithm from only a single sample per face.

## 4.3 Illumination: single sample

In this experiment, images were similar to those described in section 4.2 (Figure 3(a)), except that only a single sample per face was used (either left or right flash chosen randomly). Since each face appeared in the data set with only one illumination (either left or right), a dependence between illumination and facial identity was present. For example, left illumination direction was compatible with only half of the subjects in the data set. As a result, mutual information between illumination and facial identity was not zero, and the method of rejecting uninformative components used in the previous section would fail. Therefore, images were grouped by illumination direction rather than by facial identity. Note that this grouping can be performed reliably in an unsupervised manner even when the illumination direction is not labeled [1, 11]. The mutual information of each principal component with illumination direction was calculated, and the components with *largest* mutual information were discarded. The intuition is that relevant components describe facial features and will be independent of illumination direction, with mutual information near zero. As in section 4.2, only one principal component responsible for illumination was identified. This component was discarded, and original images were reconstructed from the remaining components. The results of the reconstruction are shown in Figure 4.



(a) Original

(b) Normalized

Figure 5: Occlusion experiment. (a): original images. (b): normalized scarf image.

### 4.4 Occlusion

Neutral face images and images occluded by a scarf (Figure 5(a)) were used in this experiment. Two neutral and two occluded samples were used for each face. Uninformative components were determined by the location of the largest drop in mutual information and discarded. Only one uninformative component was found in this experiment. Original images were reconstructed from the remaining components. Reconstruction results are shown in Figure 5(b). As can be seen, the occlusion was not removed completely. This is due to the fact that occlusion influences the image in a highly non-linear manner and therefore is not handled well by a linear method like PCA. Nevertheless, performance on a recognition task is still improved considerably after normalization by mutual information (Table 1, third row).

## 5 Discussion

A method of image normalization by mutual information was described. The method uses mutual information to identify principal components that are irrelevant to the representation and retains only relevant components. The use of mutual information to determine relevance of the components possesses significant advantages over the standard method of using eigenvalues for this purpose, since eigenvalues correspond to variance and have no direct relation to the relevance of the components.

The method was demonstrated to produce visually pleasing normalization results and to improve significantly the accuracy of recognition across a range of variability sources, such as noise, illumination, and occlusion. Although some of these could be handled by alternative methods, an additional advantage of the proposed algorithm is that it allows handling many types of image variations in a unified framework.

Note that although experiments were performed using face images, the method itself is not limited to faces. Successful applications of PCA and similar techniques to additional object classes are described e.g. in [21].

The experiments described above were performed using the direct (pixel-based) image representation. Additional image representations, such as warp maps [23, 21], may be more suitable to handle image variations resulting from 3D rotations or non-rigid motions, such as facial expressions. Exploring these representations is a subject of future research.

## References


[1] Y. Adini, Y. Moses, and S. Ullman. Face recognition: the problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:721–732, 1997.

[2] R. Basri and D. W. Jacobs. Photometric stereo with general, unknown lighting. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 374–381, 2001.

[3] P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible lighting conditions? In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–277, 1996.

[4] Peter N. Belhumeur, Joao Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. In *Proceedings of the European Conference on Computer Vision*, pages 45–58, 1996.

[5] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[6] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Generative models for recognition under variable pose and illumination. In *IEEE Conf. on Automatic Face and Gesture Recognition*, pages 277–284, 2000.

[7] Guy Gilboa, Nir Sochen, and Yehoshua Y. Zeevi. Texture preserving variational denoising using an adaptive fidelity term. In *Proc. VLSM*, 2003.

[8] Guy Gilboa, Nir Sochen, and Yehoshua Y. Zeevi. Image enhancement and denoising by complex diffusion processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.

[9] Ralph Gross and Vladimir Brajovic. An image preprocessing algorithm for illumination invariant face recognition. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*. Springer, June 2003.

[10] P. W. Hallinan. A low-dimensional representation of human faces for arbitrary lighting conditions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 995–999, 1994.

[11] Jeffrey Ho, Ming-Hsuan Yang, JongWoo Lim, Kuang-Chih Lee, and David Kriegman. Clustering appearances of objects under varying illumination conditions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–18, 2003.

[12] A. Hyvarinen, P. O. Hoyer, and E. Oja. Sparse code shrinkage: Denoising by nonlinear maximum likelihood estimation. In *Advances in Neural Information Processing Systems*, pages 473–479, 1998.

[13] Ramesh Jain, Rangachar Kasturi, and Brian G. Schunck. *Machine Vision*. McGraw-Hill, 1995.

[14] D. J. Jobson, Z. Rahman, and G. A. Woodell. A multi-scale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing: Special Issue on Color Processing*, 1997.

[15] A. M. Martinez and R. Benavente. The AR face database. Technical Report 24, CVC, 1998.

[16] M. Osadchy and D. Keren. Detection under variation of pose and illumination. *Computer Vision and Image Understanding*. To appear.

[17] M. Osadchy, M. Lindenbaum, and D. W. Jacobs. Whitening for photometric comparison of smooth surfaces. In *Proceedings of the European Conference on Computer Vision*, 2004. To appear.

[18] Ravi Ramamoorthi. Analytic PCA construction for theoretical analysis of lighting variability, including attached shadows, in a single image of a convex Lambertian object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1322–1333, 2002.

[19] A. Shashua and T. Riklin-Raviv. The quotient image: Class based re-rendering and recognition with varying illuminations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):129–139, 2001.

[20] Terence Sim and Takeo Kanade. Illuminating the face. Technical Report CMU-RI-TR-01-31, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, September 2001.

[21] Joshua B. Tenenbaum and William T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

[22] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[23] Thomas Vetter and Tomaso Poggio. Linear object classes and image synthesis from a single example image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):733–742, 1997.

[24] Haitao Wang, Stan Z. Li, Yangsheng Wang, and Weiwei Zhang. Illumination modeling and normalization for face recognition. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2003.

[25] L. Zhao and Y. H. Yang. Theoretical analysis of illumination in PCA-based vision systems. *Pattern Recognition*, 32(4):547–564, 1999.

[26] W. Zhao and R. Chellappa. Symmetric shape-from-shading using self-ratio image. *International Journal of Computer Vision*, 45:55–75, 2001.

[27] W. Zhao, A. Krishnaswamy, R. Chellappa, D. Swets, and J. Weng. Discriminant analysis of principal components for face recognition. In H. Wechsler, P.J. Phillips, V. Bruce, F. F. Soulie, and T. S. Huang, editors, *Face Recognition: From Theory to Applications*, pages 73–85. Springer-Verlag, 1998.