Literature Review on "Face Recognition in Navigating Humanoids using CNN and PCA"

Team: Kuppa Venkata Krishna Paanchajanya (19BCS063), Karthik Sajjan (19BCS049), Karusala Deepak Chowdary (19BCS050)

Reviewed articles on: Deep Learning in Computer Vision, Computer Vision in Humanoids, Face detection and recognition in Humanoids

Introduction

Artificial intelligence focuses on the ability of a machine to mimic human behavior with greater and productive efficiency. Machine learning achieves AI through algorithms trained with numerous datasets, and thus acts on a given problem with less human intervention. Deep learning, a class of machine learning inspired by the structure of human brain, provides solutions using Artificial Neural Networks at the cost of much higher volume of data required to train our machine.

The transition from the classic CPUs to the Graphics Processing Units (GPUs) and large labelled datasets being publicly available allowed significant acceleration in the training of the deep models. Advancements in Deep learning has witnessed the solutions for a variety of computer vision problems, such as object detection, motion tracking, action recognition, human pose estimation, and semantic segmentation. In this context, I wish to draw your attention to the most important type of deep learning model with respect to their applicability in visual understanding known as the Convolutional Neural Networks (CNN). We shall also name several other types of deep learning models with respect to their applicability in visual understanding and discuss the strengths and limitations of each of those deep learning models. This will facilitate a compare and contrast study with the CNN.

In a navigating humanoid, computer vision plays a pivotal role in perceiving information about the surrounding environment. For object detection, the robot is expected to scan an image and detect a target object which has certain features, such as a human face. Humanoid employs a facial recognition system that matches a human face from the live video feed through the vision systems against a database of faces, typically used to authenticate users or to recognize them for various functions to be performed. We shall discuss the implementation of a face detection and recognition system using for a humanoid using both CNN and as described in [1] using Principal Component Analysis (PCA) [2] used for data reduction and feature extraction.

CNN and its application in Facial Recognition

Convolutional Neural Networks (CNNs or ConvNet) is an artificial neural network model quite popular in analysing visual imagery. A CNN comprises three main types of neural

layers, namely, (i) convolutional layers, (ii) pooling layers, and (iii) fully connected layers. Each type of layer plays a different role.

- **Convolutional Layers:** The convolutional layers in the CNN are the one which make a CNN stand different from a Multi-layer Perceptron (MCP), a simple feedforward artificial neural network. These convolution layers perform what we call a convolution operation wherein the layer receives an input and then outputs the transformed input to the subsequent layer. These transformations are done with the help of **filters (kernels)** essential in pattern detection in the input image. These patterns can be edges, corners or even more complex objects like human faces, animals, objects, etc. Thus, in the convolutional layers, a CNN utilizes various kernels to convolve the whole image as well as the intermediate feature maps, generating various feature maps.
- **Pooling Layers:** Pooling layers are in charge of reducing the spatial dimensions (width x height) of the input volume for the next convolutional layer. The pooling layer does not affect the depth dimension of the volume. The operation performed by this layer is also called subsampling, as the reduction of size leads to a simultaneous loss of information. However, such a loss is beneficial for the network because the decrease in size leads to less computational overhead for the upcoming layers of the network, and also it works against overfitting. Average pooling and max pooling are the most commonly used strategies. Max pooling can lead to faster convergence, select superior invariant features, and improve generalization. Also there are a number of other variations of the pooling layer in the literature, each inspired by different motivations and serving distinct needs, for example, stochastic pooling [3], spatial pyramid pooling [4], and def-pooling [5].
- **Fully Connected Layers**: Following several convolutional and pooling layers, the high-level reasoning in the neural network is performed via fully connected layers. Neurons in a fully connected layer have full connections to all activation in the previous layer, as their name implies. Their activation can hence be computed with a matrix multiplication followed by a bias offset. Fully connected layers eventually convert the 2D feature maps into a 1D feature vector. The derived vector either could be fed forward into a certain number of categories for classification [6] or could be considered as a feature vector for further processing.

Every layer of a CNN thus transforms the input volume to an output volume of neuron activation, eventually leading to the final fully connected layers, resulting in a mapping of the input data to a 1D feature vector. CNNs have been extremely successful in computer vision applications, such as face recognition, object detection, powering vision in robotics, and self-driving cars.

Several other types of deep learning models capable of visual understanding include the Deep Belief Networks (DBNs) and Deep Boltzmann Machines (DBMs) and Stacked (Denoising) Autoencoders.

Comparison of CNNs vs DBNs vs DBMs vs SdAs in Visual Imagery [7]

The CNNs achieve significant performance rates in a variety of visual understanding tasks. However, they have their own merits and demerits. CNNs automatically learn features based on the given dataset. CNNs are invariant to transformations, which is a great asset for certain computer vision applications. On the other hand, they heavily rely on the existence of labelled data, in contrast to the other types of deep learning models such as DBNs/DBMs and SdAs, which can work in an unsupervised fashion.

Application of CNN in Face Recognition

Face recognition is one of the trending computer vision applications with great commercial interest as well. A feature extractor in a face recognition system extracts features from an aligned face to obtain a low-dimensional representation, based on which an appropriate classifier makes predictions. CNNs brought about a change in the face recognition field owing to their feature learning and transformation invariance properties.

Google's FaceNet [8] and Facebook's DeepFace [9] are both based on CNNs. Although DeepFace attains great performance rates, its representation is not easy to interpret because the faces of the same person are not necessarily clustered during the training process. On the other hand, FaceNet defines a triplet loss function on the representation, which makes the training process learn to cluster the face representation of the same person. Furthermore, CNNs constitute the core of OpenFace [10], an open-source face recognition tool, which is of comparable accuracy, and is suitable for mobile computing, because of its smaller size and fast execution time.

Face Detection and Recognition System using PCA

Image Pre-processing using Histogram Equalization

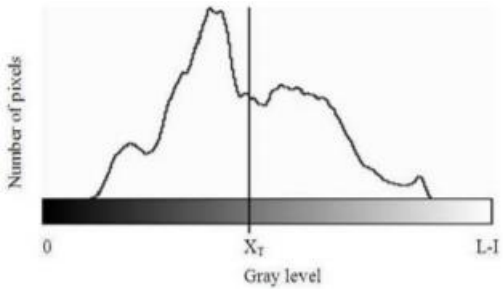
Histogram equalization [11] is a method to improve lighting in an image processing by adjusting its contrast using the image's histogram. The areas with lower local contrast can gain a higher contrast in this way. Histogram equalization is a basic technique from spatial domain of image processing. For implementing histogram equalization, it can be thought as palette change. Given L as maximum Gray scale, histogram from digital image with Gray scale span [0, L-1] is a discrete function:

$$h(r_k) = n_k$$

where r_k is k^{th} Gray scale value, and n_k is the number of pixels in image that have r_k's gray scale value. The figure below shows the histogram of sample image. If the distance from 0 to L-1 divided into 2 parts with X_T as the threshold intensity, then this separation produces 2 histograms. The first histogram has scope from 0 to X_T, and the second histogram has scope from X_T to L-1. Given an image with M x N pixels and L shows the maximum gray scale value, histogram equalization transformation T mapped input value r_k (where k=0, 1, 2,…, L-1) to output value S_k as follows:

$$S_K = T(r_k) = \frac{L-1}{MN}\sum_{j=0}^{k} n_j$$

With the histogram equalization transformation, the lighting in the image can be corrected effectively.



Face detection and recognition system can be divided into four steps, that is:

- Face Detection
- Face Alignment
- Feature Extraction
- Feature Matching

Face Detection

Viola-Jones [12] is chosen as the face detection method which provides a complete framework for extracting and recognizing image features. A combination of several algorithms in Viola-Jones method including HAAR-like features, integral image, boosting algorithm, and cascade classifier provides a robust and fast detector for the human face detection.

Face Recognition

Face recognition is the process where computer analyses the detected face image to identify it by comparing to the known face image in database. Face recognition process can be divided into two steps: feature extraction and feature matching.

In feature extraction, we extract unique features from the detected face image. Here, PCA (Principle Component Analysis) [2] is used for this task. According to Shakhnarovich and Moghaddam [13], there are some difficulties in face recognition that is, to handle high dimension, especially recognition context based on similarity and matching is expensive based on computation.

Therefore, dimension reduction technique is needed to build face recognition system. One of the fast and reliable algorithms for dimension reduction is PCA (Principal Component Analysis). Thus, PCA has two roles in face recognition that is: for extracting image feature and also in reducing image dimension.

PCA

Principal Component Analysis is a standard technique that is used in statistical pattern recognition and signal processing for data reduction and extraction features [2]. In PCA, the enquired and known images must be of the same size. Therefore, a normalization is needed to line-up the eyes and the mouths across all images. Each image is treated as one vector. All images of the training set are stored in a single matrix Γ and each row in the

matrix represents an image. The average image has to be calculated and then subtracted from each original images in Γ. Then calculate the eigenvectors and eigenvalues of the covariance matrix C. These eigenvectors are called eigenfaces. An eigenface is the result of the reduction in dimensions which removes the useless information and decomposes the face structure into the uncorrelated components.

The training database consists of M images of same size. The images are normalized by converting each image matrix to equivalent image vector Γ_i. The training set matrix Γ is the set of image vectors, that is:

Training set $\Gamma = [\Gamma_1, \Gamma_2, \ldots \Gamma_M]$

The mean face (Ψ) is the arithmetic average vector as given by: $\Psi = \frac{\sum_{i=1}^{M} \Gamma_i}{M}$

The deviation vector for each image Φ_i is given by $\Phi_i = \Gamma_i - \Psi$ for I = 1, 2, 3, …M

Consider a difference matrix $A = [\Phi_1, \Phi_2, \ldots \Phi_M]$ which keeps only the distinguishing features for face images and removes the common features. Then eigenfaces are calculated by find the Covariance matrix C of the training image vectors by: $C = A.A^T$

Due to large dimension of matrix C, we consider matrix L of size (M_t x M_t) which gives the same effect with reduced dimension. The eigenvectors of C (matrix U) can be obtained by using the eigenvectors of L (matrix V) as given by: $U_i = AV_i$

The eigenfaces are: eigenface = $[U_1, U_2, \ldots U_M]$

Instead of using M eigenfaces, the certain value m' ≤ M is chosen as the eigenspace. Then the weight of each eigenvector ω_i to represent the image in the eigenface space, is given by: $\omega_i = U_i^T(\Gamma - \Psi)$, i = 1, 2, 3, …m'

Weight matrix $\Omega = [\omega_1, \omega_2, \ldots \omega_{m'}]^T$

Average class projection = $\Omega_\Psi = \frac{\sum_{i=1}^{X_i} \Omega_i}{X_i}$

Finally, the Euclidean distance δ_i is used to find out the distance between two face keys vectors and is given by: $\delta_i = ||\Omega - \Omega_{\Psi i}|| = \sum_{k=1}^{M}(\Omega_k - \Omega_{\Psi ik})$

Euclidian distance is one of the methods that can be used to match a new face image to the existing face image in the database. Smaller the Euclidian distance, more is the image similar to the one available in the database.

Conclusion

We have discussed above two methods of facial detection and recognition viz., using Convolutional Neural Networks (CNN) and Principal Component Analysis (PCA). The live video feed perceived by the frontal cameras of the humanoid shall be first converted into grayscale image frames, to reduce computation in the image processing. The CNN deep learning model then convolves the image data across various layers, reduces spatial dimensions using any pooling technique and then recognises the particular person based

on the already trained human faces. In the case of PCA, the gray scale image will be initially corrected using histogram equalisation method. The eigenface of the gray scale image and subsequently its weight matrix shall be computed which are further used to calculate the Euclidean distance with respect to the weight matrix of every eigenface in the database to recognize the face. The accuracy of the PCA face recognition is about 93% as per [1]. Using CNN as classifiers, the face recognition systems reach the accuracy of about 95% [14]. In comparison with this result, the PCA system which uses simple Eulidean distance as classifier has comparable performance.

References

[1] Christian Tarunajaya1 , Oey Kevin Wijaya1 , Reinard Lazuardi Kuwandy1 , Heri Ngarianto1 , Alexander A S Gunawan1 , and Widodo Budiharto1, IPTEK, The Journal for Technology and Science, Vol. 26, No. 2, August 2015 Development of Intelligent Humanoid Robot with Face Recognition Features

[2] M. Turk and A. Pentland, "Eigenfaces for Recognition," Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71-86, 1991

[3] H. Wu and X. Gu, “Max-Pooling Dropout for Regularization of Convolutional Neural Networks,” in Neural Information Processing, vol. 9489 of Lecture Notes in Computer Science, pp. 46–54, Springer International Publishing, Cham, 2015.

[4] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition,” in Computer Vision – ECCV 2014, vol. 8691 of Lecture Notes in Computer Science, pp. 346–361, Springer International Publishing, Cham, 2014.

[5] W. Ouyang, X. Wang, X. Zeng et al., “DeepID-Net: Deformable deep convolutional neural networks for object detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, pp. 2403–2412, USA, June 2015.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS '12), pp. 1097–1105, Lake Tahoe, Nev, USA, December 2012.

[7] Deep Learning for Computer Vision: A Brief Review, HindAwi, Volume 2018 |Article ID 7068349

[8] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: a unified embedding for face recognition and clustering,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '15), pp. 815–823, IEEE, Boston, Mass, USA, June 2015.

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: closing the gap to human-level performance in face verification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14), pp. 1701–1708, Columbus, Ohio, USA, June 2014.

[10] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, “Openface: a general-purpose face recognition library with mobile applications,” CMU-CS-16-118, CMU School of Computer Science, 2016.

[11] R. C. Gonzalez and R. E. Woods, Digital Image Processing (3rd Edition), New Jersey: Prentice Hall, 2007.

[12] "B. Senjaya, A. A. S. Gunawan and J. P. Hakim, "Pendeteksian Bagian Tubuh Manusia untuk Filter Pornografi dengan Metode Viola-Jones," Jurnal ComTech, vol. 3, no. 1, pp. 482-489, 2012".

[13] G. Shakhnarovich and B. Moghaddam, "Face Recognition in Subspaces," in Handbook of Face Recognition, London, SpringerVerlag London, 2004, pp. 141-168.

[14] A. Eleyan and H. Demirel, "PCA and LDA based Neural Networks for Human Face Recognition," in Face Recognition, Vienna, I-Tech Education and Publishing, 2007, pp. 93-106.