Computationally Efficient Algorithms of Image Recognition Based on Sequential Analysis of Deep Neural Network

A. Sokolova, A. Savchenko

National Research University Higher School of Economics – Nizhny Novgorod

ANNOTATION

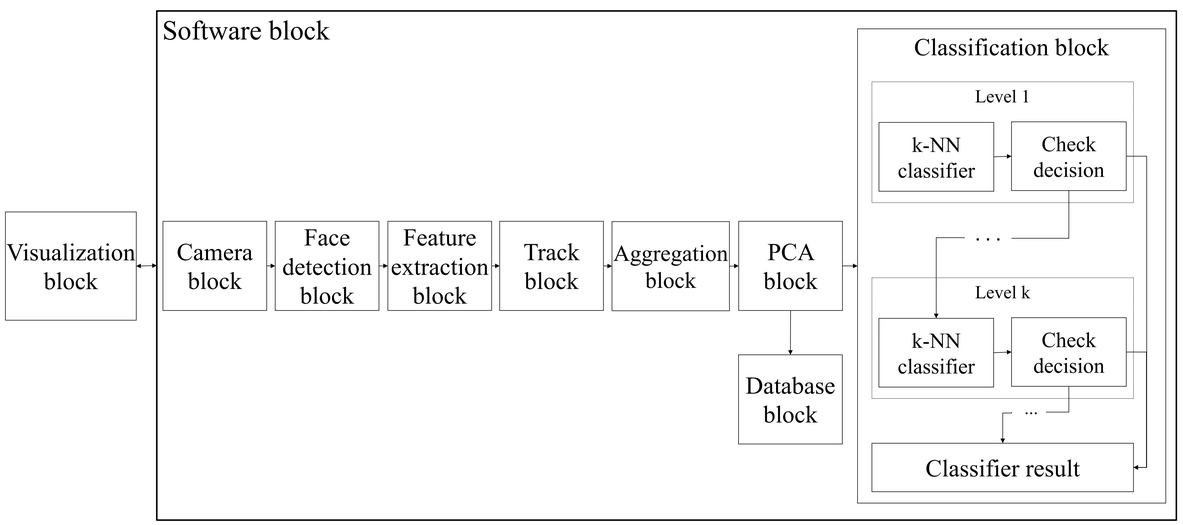
Each input object is associated with a sequence of principal component scores of aggregated features extracted by deep neural network. The number of components in each element of this sequence is dynamically chosen based on explained proportion of total variance for the training set. We propose to process the next element with higher explained variance only if the decision for the current element is unreliable. This reliability is estimated by matching the ratio of the minimum distance and all other distances with a certain threshold. Experimental study for face recognition with the Labeled Faces in the Wild and YouTube Faces datasets demonstrates the decrease of running time up to 10 times when compared to conventional instance-based learning

GOAL

Improve the speed of the nearest neighbor classifiers of a set of points based on sequential analysis of high-dimensional feature vectors

APPROACH

The task of set-of-points classification is formulated as follows. It is required to assign an input set of T feature vectors x(t) = [x1(t),...,xD(t)] into one of C classes. They are specified by a training set of R≥C points xr= [xr1,...,xrD], r∈{1,...,R}, which class label c(r)∈{1,...,C} is known. We assume that dimensionality D is rather high, e.g. when deep convolutional neural network (CNN) is used for feature extraction



CLASSIFICATION

- The distance between feature level to speed-up the matching

$$\rho\left(\tilde{x}^{(l)},\tilde{x}^{(l)}\right)=\rho\left(\tilde{x}^{(l-1)},\tilde{x}^{(l-1)}\right)+\sum_{d=d_{(l-1)}+1}^{d_l}\rho\left(\tilde{x}_d,\tilde{x}_{r;d}\right)$$

- Nearest neighbor class

$$c_l^*=\underset{c\in C_l}{\operatorname{argmin}}\,\rho_c(\tilde{x}^{(l)})$$

- The set of candidates

$$C_{l+1}=\left\{c\in C_l\left|\frac{\rho_c(\tilde{x}^{(l)})}{\rho_{c_l^*}(\tilde{x}^{(l)})}\le\delta\right.\right\}$$

AGGREGATION

- Average features vector $\overline{x}=\frac{1}{T}\sum_{t=1}^{T}x(t)$
- Medoids $x^*=\underset{x(t)}{\operatorname{argmin}}\sum_{t'=1}^{T}\rho(x(t),x(t'))$
- Median features $x'=[x'_1,\dots,x'_D]$

EXPERIMENTAL DATA

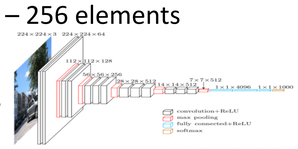
CNNs:

Lightened CNN (Version C) – 256 elements

FaceNet – 512 elements

VggFace2 – 2048 elements

VggFace – 4096 elements



DATASETS:

YouTube Faces (YTF)

1595 people, 3425 videos, 48-6070 frames

Labeled Faces in the Wild (LFW)

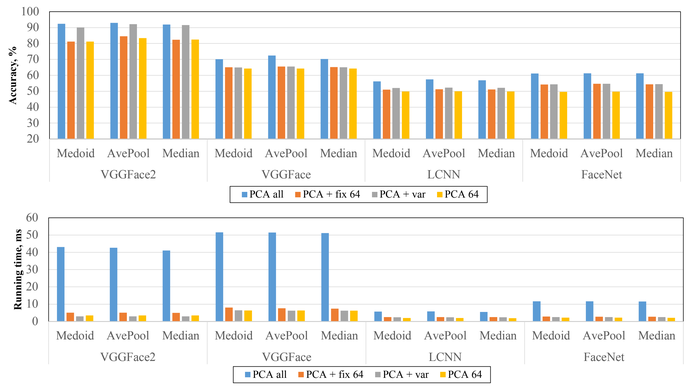
1680 people, 13000 facial photos

IAPRA Janus Benchmark (IJB-C)

31334 facial of 3531 people, 117542 frames of 11779 videos

RESULTS

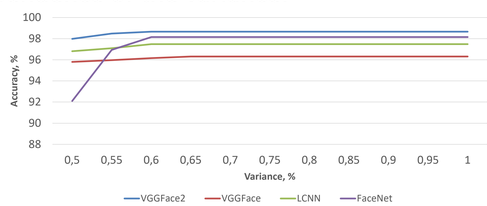
Face recognition results for the k-NN classifier, LFW-YTF dataset:



Face recognition results for the k-NN classifier, IJB-C dataset:

Metrics	Classifier	VGGFace2	FaceNet
Accuracy (%)	k-NN, all components	87.24	53.08
	k-NN, 64 components	82.53	49.73
	sequential k-NN, fixed no. of compone	85.50	51.29
	sequential k-NN, variable no. of components	86.07	52.36
Time (ms)	k-NN, all components	42.17	16.63
	k-NN, 64 components	3.12	3.11
	sequential k-NN, fixed no. of compone	3.64	3.22
	sequential k-NN, variable no. of components	3.42	3.09

The dependence of the variance:



The proposed approach is only 0.3-1% less accurate when compared to the traditional k-NN for all features, though 2-10-times speed-up is obtained

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY