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

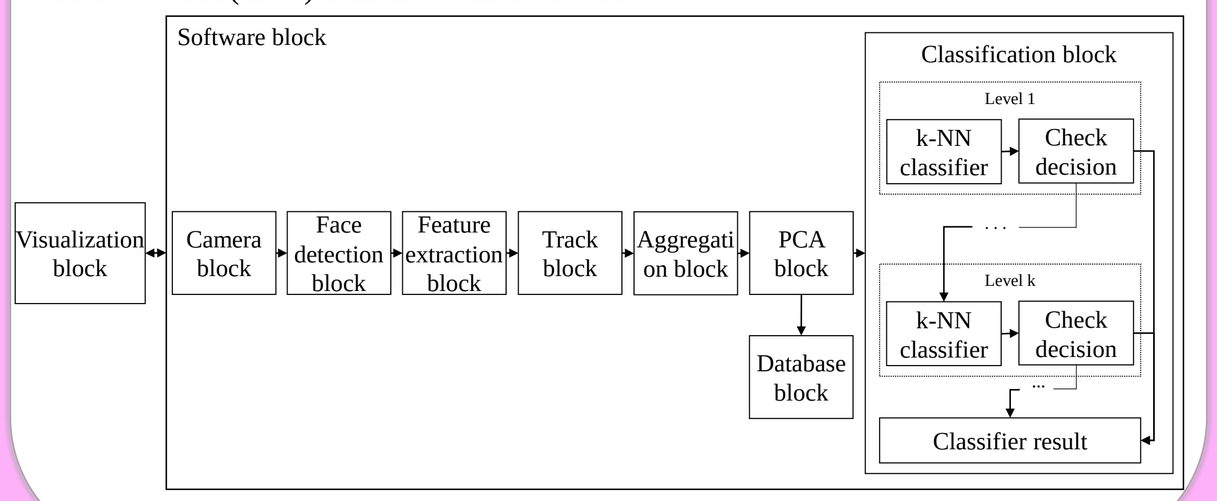
A. Sokolova, A. Savchenko
National Research University Higher School of Economics
Nizhny Novgorod

Annotation

In this work we improve the speed of the nearest neighbor classifiers of a set of points based on sequential analysis of high-dimensional feature vectors. 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 of 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.

Proposed 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 algorithm

The distance between feature level to speed-up the matching

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

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 techniques

1. Average features vectors of each track

$$\rho\big(X(m_1), X(m_2)\big) = \rho\big(\bar{\mathbf{x}}(m_1), \bar{\mathbf{x}}(m_2)\big), \quad \bar{\mathbf{x}}(m_i) = \frac{1}{\Delta t(m_i)}\sum_{t'=t_1(m_i)}^{t_2(m_i)}\mathbf{x}(t)$$

2. The distance between medoids

$$\rho\big(X(m_1), X(m_2)\big) = \rho\big(\mathbf{x}^*(m_1), \mathbf{x}^*(m_2)\big), \mathbf{x}^*(m_i) = \underset{\mathbf{x}(t), t\in[t_1(m_i), t_2(m_i)]}{\operatorname{argmin}}\sum_{t'=t_1(m_i)}^{t_2(m_i)}\rho\big(\mathbf{x}(t), \mathbf{x}(t')\big)$$

3. Comparison of the median features of each track

$$\rho\big(X(m_1), X(m_2)\big) = \rho\big(\mathbf{x}'(m_1), \mathbf{x}'(m_2)\big)$$

Experimental data

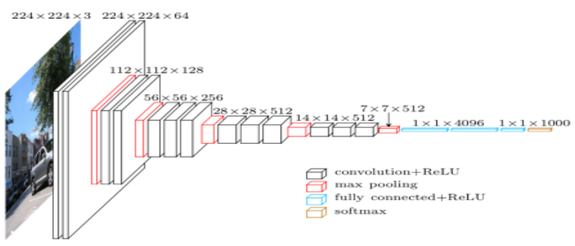
Convolutional neural networks

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

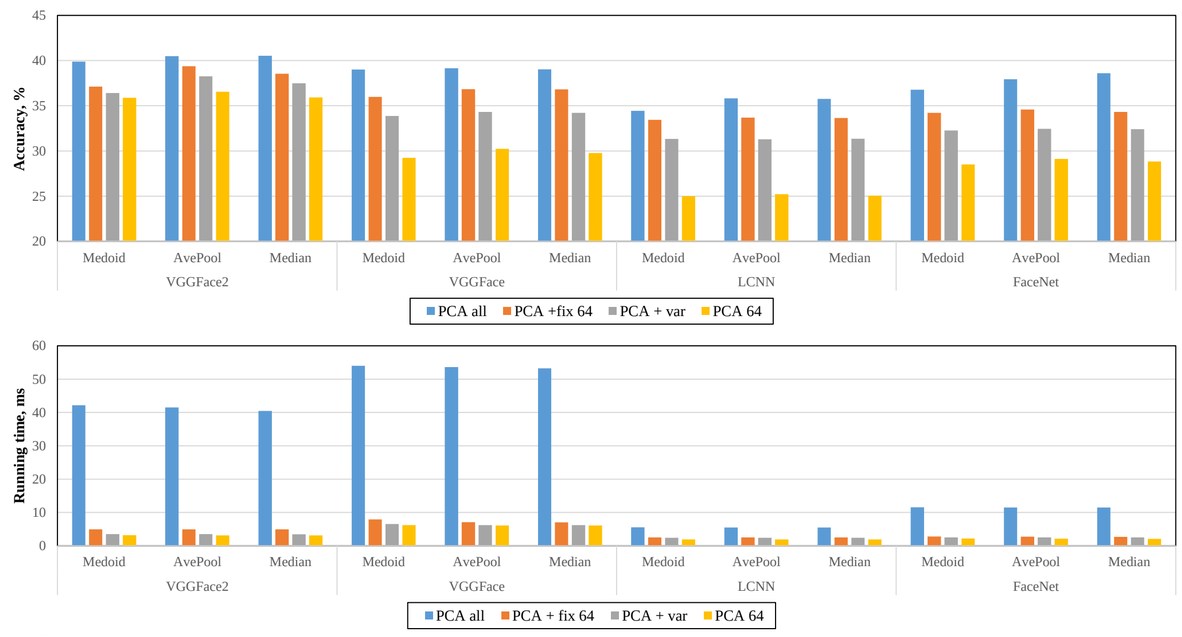
Experimental results

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

Metric	Classifier	VGGFace	LCNN	VGGFace2	FaceNet
Accuracy (%)	k-NN, all components	96.31	97.48	98.66	98.15
	k-NN, 64 components	94.10	96.21	96.95	97.36
	sequential k-NN, fixed no. of components	95.97	96.97	98.32	98.15
	sequential k-NN, variable no. of components	95.80	96.81	97.98	97.94
Time (ms)	k-NN, all components	52.49	5.41	37.14	11.49
	k-NN, 64 components	6.87	2.31	4.12	2.62
	sequential k-NN, fixed no. of components	7.23	2.54	4.85	2.69
	sequential k-NN, variable no. of components	6.10	2.23	3.58	2.59

Here the proposed approach is only 0.3-1% less accurate when compared to the traditional k-NN for all features, though 2-7-times speed-up is obtained. Conventional matching of only fixed number of principal components is 0.6-1.7% less accurate than the proposed implementation of sequential analysis.

The dependencies of recognition accuracy and running time on various aggregation techniques for YTF:



Here the accuracy is much lower when compared to the previous experiment because the LFW dataset contains more images per one subject, while YTF dataset includes a lot of similar video frames that should be aggregated. Our sequential procedure again enables to significantly reduce the running time without noticeable accuracy degradation.

The dependencies of recognition accuracy and running time on various aggregation techniques for LFW(train) – YTF(test):

The most effective results were demonstrated by PCA with variable number of components in each level obtained by increasing explained variance for features extracted by VGGFace2 CNN. The decision-making time in the proposed approach is 10-times lower than the conventional implementation of the k-NN.

Conclusion

In this work we proposed the modification of the k-NN classification method based on sequential analysis of high-dimensional features. Experiments with face recognition and contemporary deep features demonstrated that our approach is up to 8-10-times faster when compared to the original k-NN method, while the accuracy decreases only by 0.1-0.9\%. The main direction for further research is to examine sophisticated classifiers in order to increase the recognition accuracy. Moreover, it is important to use more difficult dissimilarity measures, e.g., implement metric learning to point-to-set and set-to-set distance learning.

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY