Video clustering in surveillance systems using deep embeddings

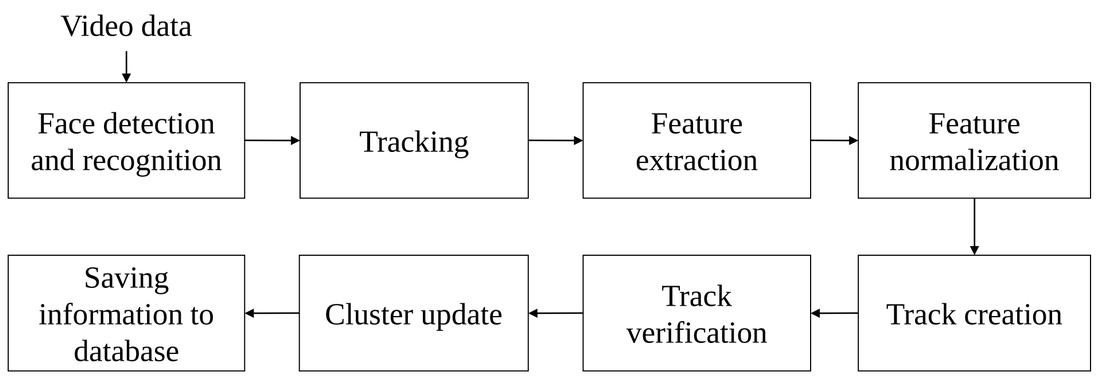
A.D. Sokolova, A.V. Savchenko
National Research University Higher School of Economics
Nizhny Novgorod

Introduction

We consider the organizing of data in video surveillance systems by grouping the video tracks, which contain identical faces. The facial regions are detected in each frame and a video stream is split into sequences with one person. Aggregation of the features of individual frames extracted using deep convolutional neural networks are used in order to obtain a descriptor of video track. The tracks with identical faces are grouped using the known face verification algorithms and clustering methods.

Proposed approach

The task is to divide the input video sequence of $T>1$ frames into $M < T$ subsequent tracks $\{X(m)\}$, $m = 1,2,...,M$ contained face images of one person and cluster similar tracks. Each m-th track is characterized by the indices of its start t_1(m) and end frame t_2(m).



Distance measurement

1. Euclidean metric (*L2*)

$$\rho\left(x_1(t), x_2(t)\right) = \sqrt{\sum_{k=1}^{N} (x_{1k}(t) - x_{2k}(t))^2}$$

2. Student t-criteria $(t - test)$

$$t = \frac{\rho(X(m_1), X(m_2))}{\sqrt{\frac{D(m_1)}{\Delta t(m_1)} + \frac{D(m_2)}{\Delta t(m_2)}}}$$

Aggregation techniques

1. The mean pairwise distances between all frames

$$\rho\left(X(m_1), X(m_2)\right) = \frac{1}{\Delta t(m_1)\Delta t(m_2)} \sum_{t=t_1(m_1)}^{t_2(m_1)} \sum_{t`=t_1(m_2)}^{t_2(m_2)} \rho(\mathrm{x}(t), \mathrm{x}(t`))$$

2. The distance between their medoids

$$\rho\left(X(m_1), X(m_2)\right) = \rho\left(\mathrm{x}^*(m_1), \mathrm{x}^*(m_2)\right),$$

$$\mathrm{x}^*(m_i) = \underset{\mathrm{x}(t), t\in[t_1(m_i), t_2(m_i)]}{\operatorname{argmin}} \sum_{t`=t_1(m_i)}^{t_2(m_i)} \rho(\mathrm{x}(t), \mathrm{x}(t`))$$

3. Average features vectors of each track are matched

$$\rho\left(X(m_1), X(m_2)\right) = \rho\left(\bar{\mathrm{x}}(m_1), \bar{\mathrm{x}}(m_2)\right),$$

$$\bar{\mathrm{x}}(m_i) = \frac{1}{\Delta t(m_i)} \sum_{t`=t_1(m_i)}^{t_2(m_i)} \mathrm{x}(t`)$$

4. Comparison of the median features $\mathrm{x}`(m_i)$ of each track

$$\rho\left(X(m_1), X(m_2)\right) = \rho\left(\mathrm{x}`(m_1), \mathrm{x}`(m_2)\right)$$

5. Learnable pooling of video features of the *m*–th track

$$r(m) = \sum_{t=t_1(m)}^{t_2(m)} a(t)\mathrm{x}(t)$$

$$a(t) = \frac{\exp(q^t \mathrm{x}(t))}{\sum_{t=t_1(m)}^{t_2(m)} \exp(q^t \mathrm{x}`(t))}$$

In order to improve the quality of learnable pooling it is significant to use two sequential attention blocks

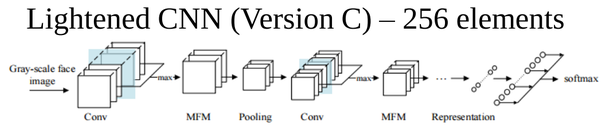
$$q^{(1)} = \tanh(Wr^{(0)} + b)$$

where W and b are the learnable weight matrix and bias vector of the neurons respectively.

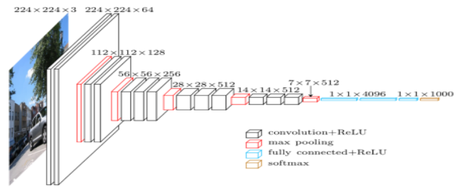
Experimental data

Convolutional neural networks

Lightened CNN (Version C) – 256 elements



VggFace – 4096 elements
VggFace2 – 2048 elements

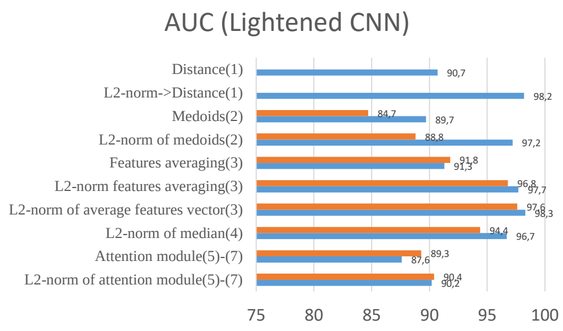


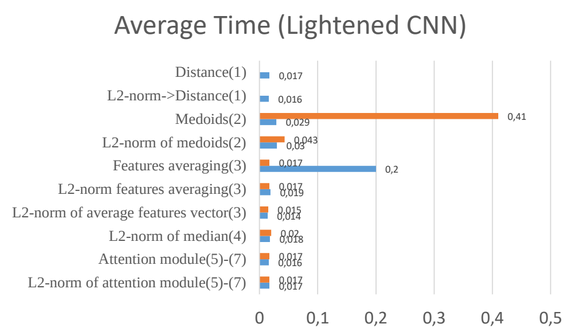
Dataset

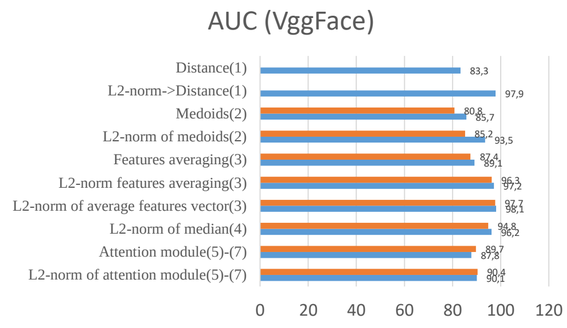
YouTube Faces (YTF):
1595 people
3425 videos
48-6070 frames

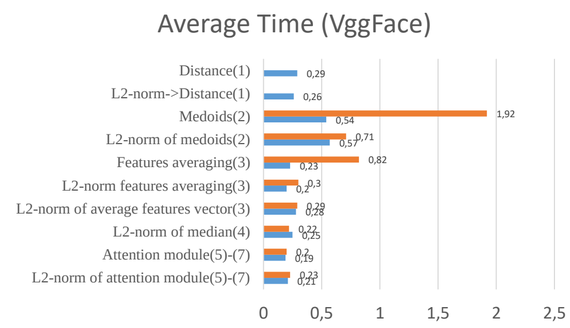
Experimental results

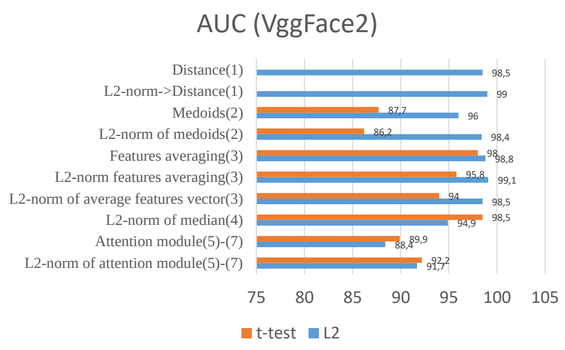
The AUC (Area Under Curve) and average time were computed:

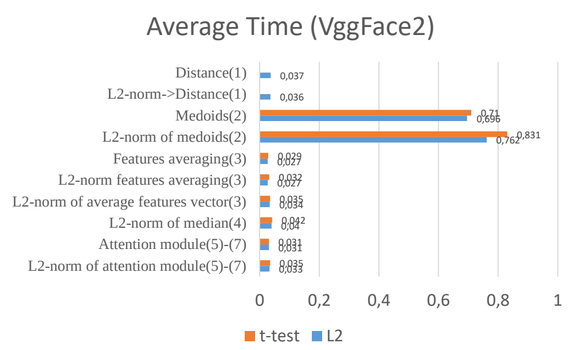




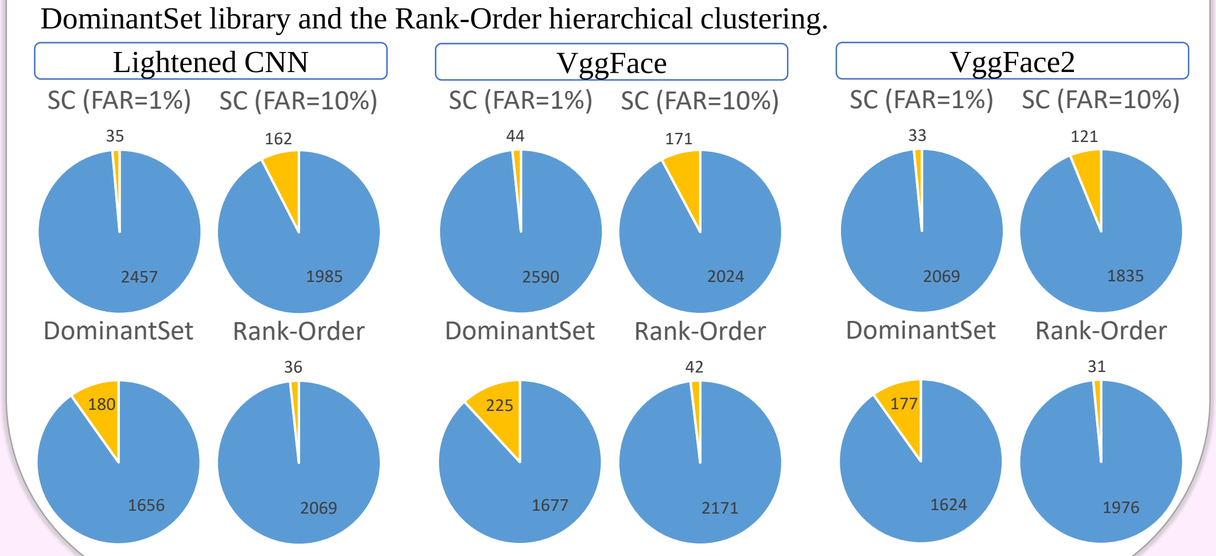








We implemented sequential clustering (SC) where threshold for resulting clusters is set by fixing the FAR value. In addition, we examined the clustering algorithm from the DominantSet library and the Rank-Order hierarchical clustering.



Conclusion

The problem of video subsequences clustering for video surveillance systems was solved. In particular, we focused on calculating the degree of proximity of video tracks using the aggregation of features vectors extracted by deep CNNs. We experimentally compare frame aggregation methods using the YouTubeFaces dataset and contemporary neural networks (LightenedCNN, VGGFace, VGGFace2). Experimental study demonstrated that the features vectors averaging of all frames and subsequent normalization lead to the highest accuracy of video face verification. In the future work we plan to analyze other clustering algorithms deeper in order to achieve low calculation complexity and high accuracy of data processing.

VisionLabs
MACHINES CAN SEE

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY
NIZHNY NOVGOROD