HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

Organizing Multimedia Data in Video Surveillance Systems Based on Face Verification with Convolutional Neural Networks

Anastasiia D. Sokolova
Email: adsokolova96@mail.ru

Analysis of images, social networks, and texts 2017

Outline

- Image recognition and faces ordering problem
- Proposed two-stage approach of organizing information in video surveillance systems
- Experimental results in face recognition
- Concluding comments and future plans

Image recognition problem

The problem is to split the given video sequence of T frames into subsequences with observations of one person, and then unite different subsequences containing the same person.

Conventional approach

An appropriate tracker algorithm divides the input sequence into $M<T$ disjoint subsequences (tracks , which consist of face image) $\{X(m)\}$, $m = 1,2, ..., M$. The similar objects are sequentially grouped together using clustering methods.

Key idea

Improvement of verification efficiency by the combination for features extracted from individual frames

Proposed approach

Feature extraction is implemented using the deep CNNs trained with an external large dataset.

The highest accuracy was achieved with the average distance:

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

Straightforward aggregation techniques:

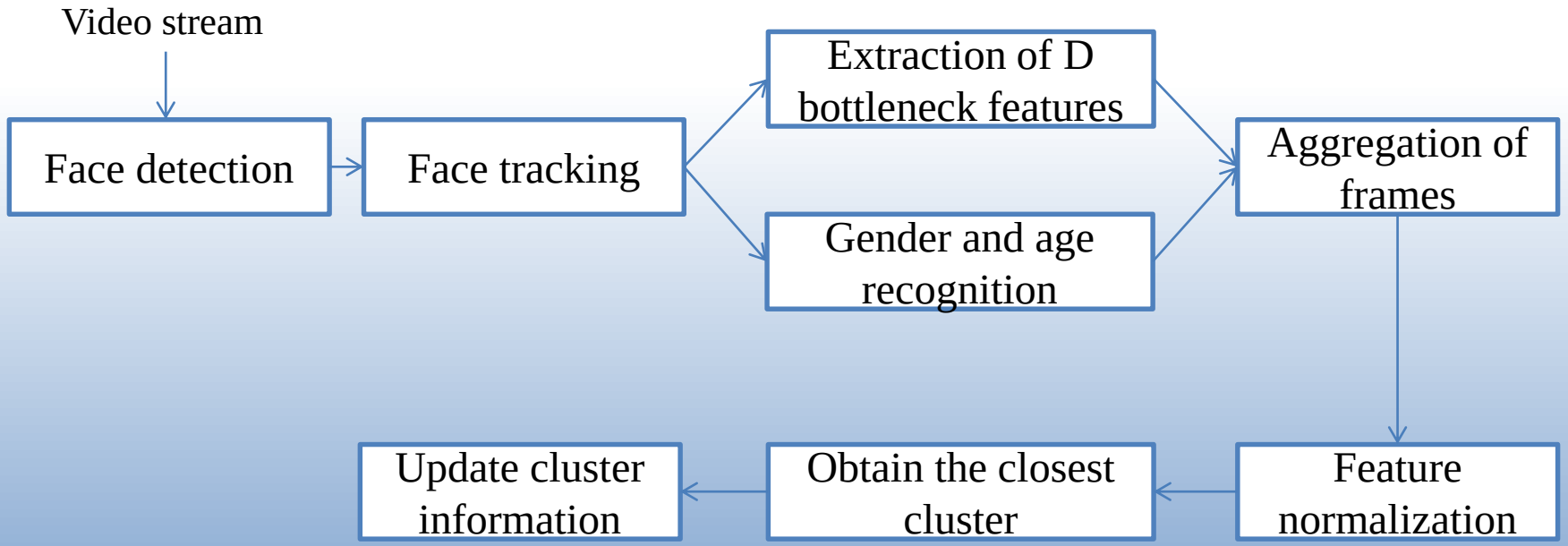
- The distance between tracks is defined as the distance between their medoids:

$$\rho\left(X(m_1), X(m_2)\right) = \rho\left(\mathbf{x}^*(m_1), \mathbf{x}^*(m_2)\right), \mathbf{x}^*(m_i) = \underset{\mathbf{x}(t), t\in[t_1(m_i), t_2(m_i)]}{\operatorname{argmin}} \sum_{t'=t_1(m_i)}^{t_2(m_i)} \rho\left(\mathbf{x}(t), \mathbf{x}(t')\right) \quad (2)$$

- Average features of each track are matched:

$$\rho\left(X(m_1), X(m_2)\right) = \rho\left(\bar{\mathbf{x}}(m_1), \bar{\mathbf{x}}(m_2)\right), \quad \bar{\mathbf{x}}(m_i) = \frac{1}{\Delta t(m_i)} \sum_{t'=t_1(m_i)}^{t_2(m_i)} \mathbf{x}(t) \quad (3)$$

Proposed algorithm



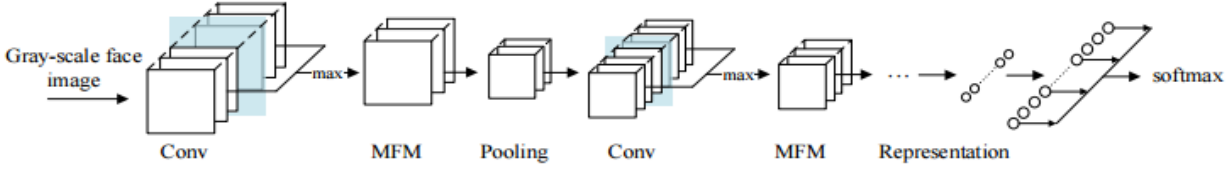
Datasets

- LFW (Labeled Faces in the Wild)
  - 1680 people
  - 13000 images
  - 1-10 frames
- YTF (YouTube Faces)
  - 1595 people
  - 3425 videos
  - 48-6070 frames

Face detection and feature extraction

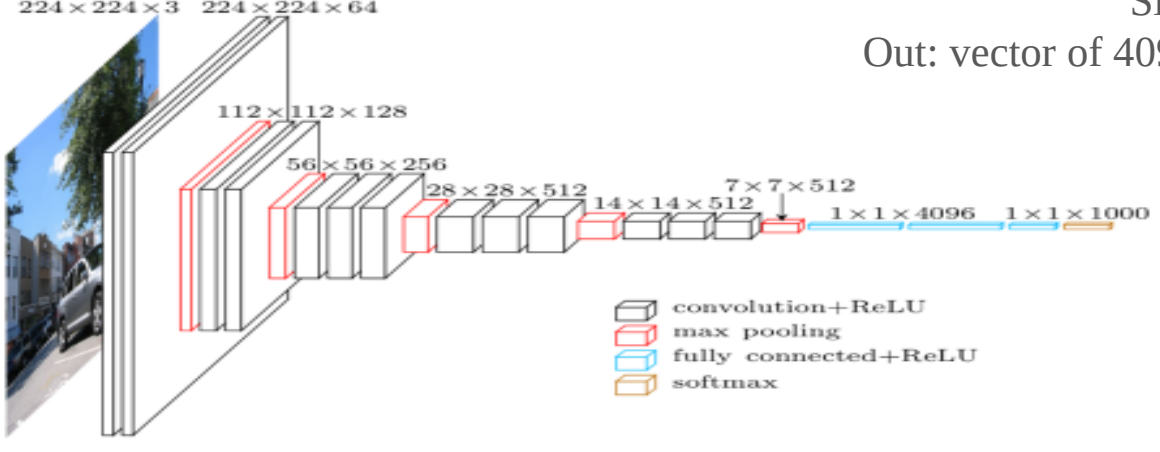
- Lightened CNN (version C)

Size: 119 Mb
Out: vector of 256 elements



- VGGNet

Size: 553 Mb
Out: vector of 4096 elements



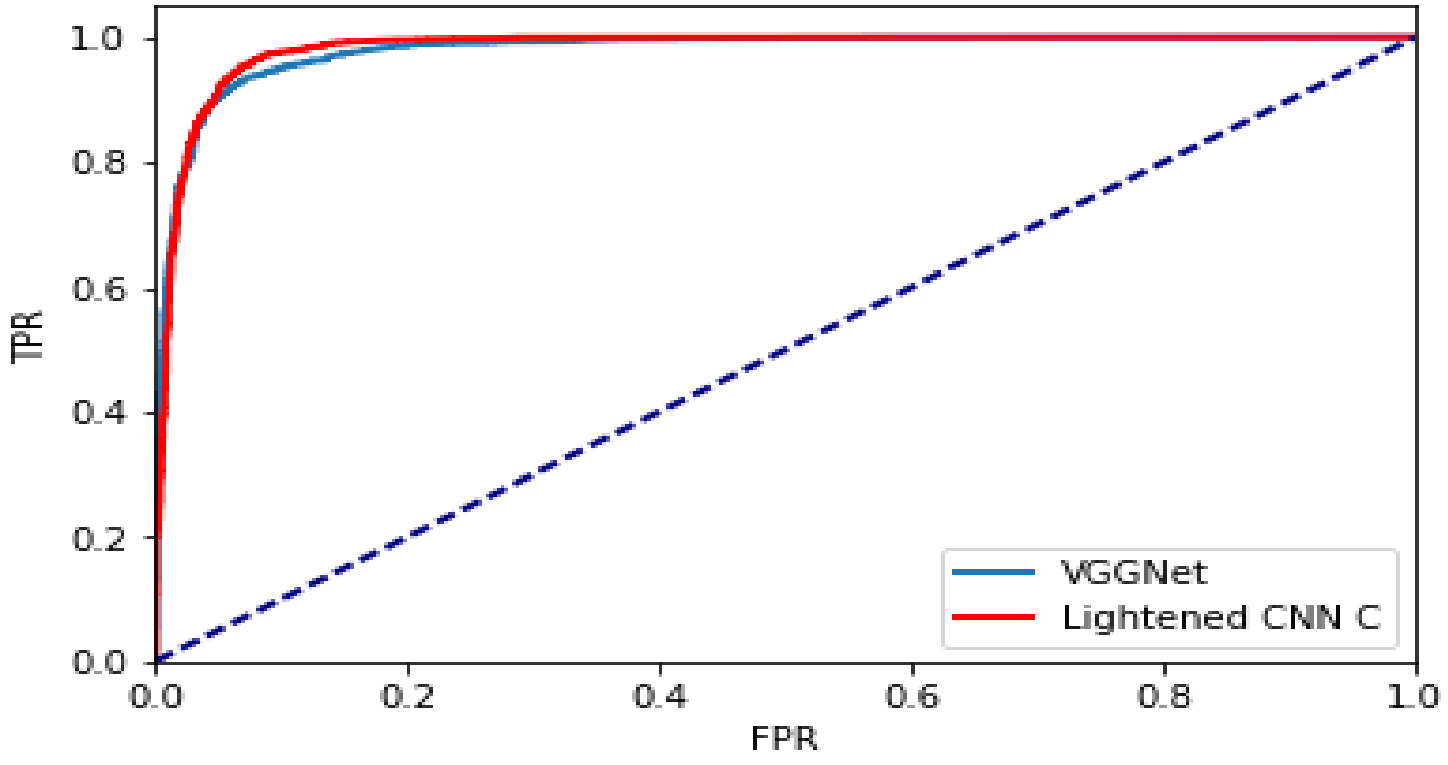
Face identification

0.4285

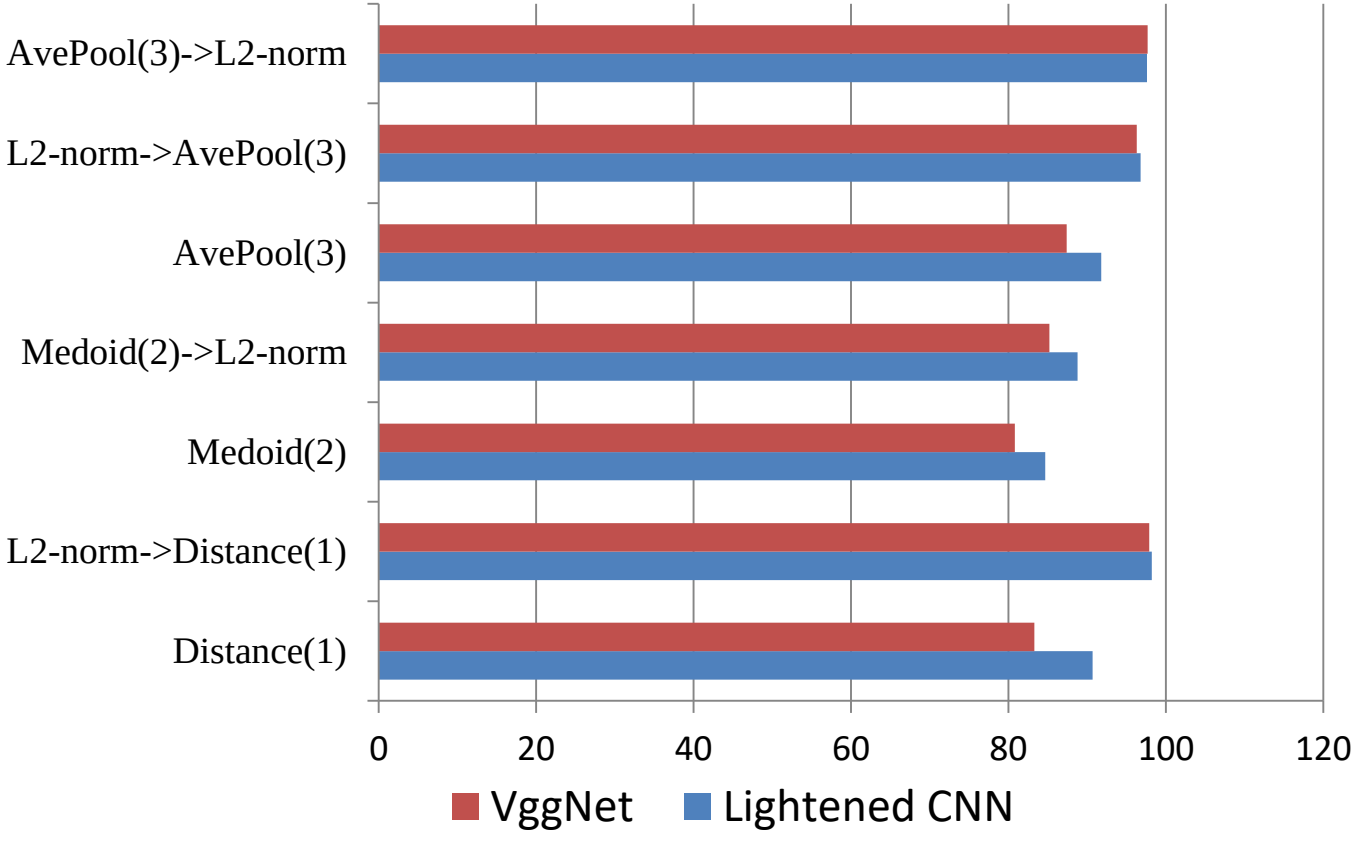
0.521

0.8934

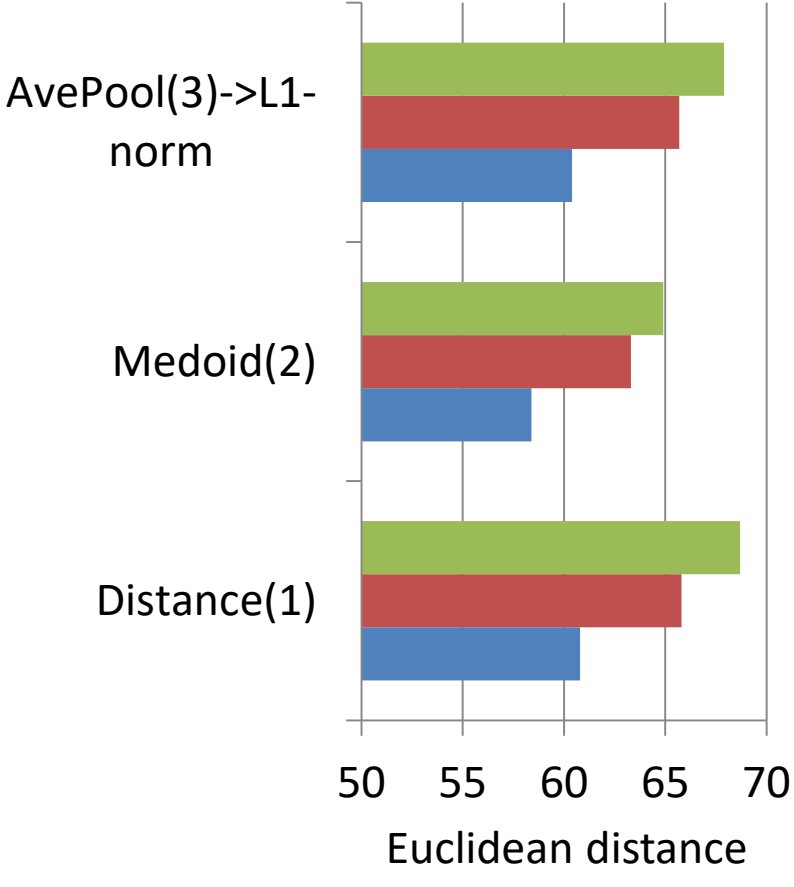
Algorithm accuracy

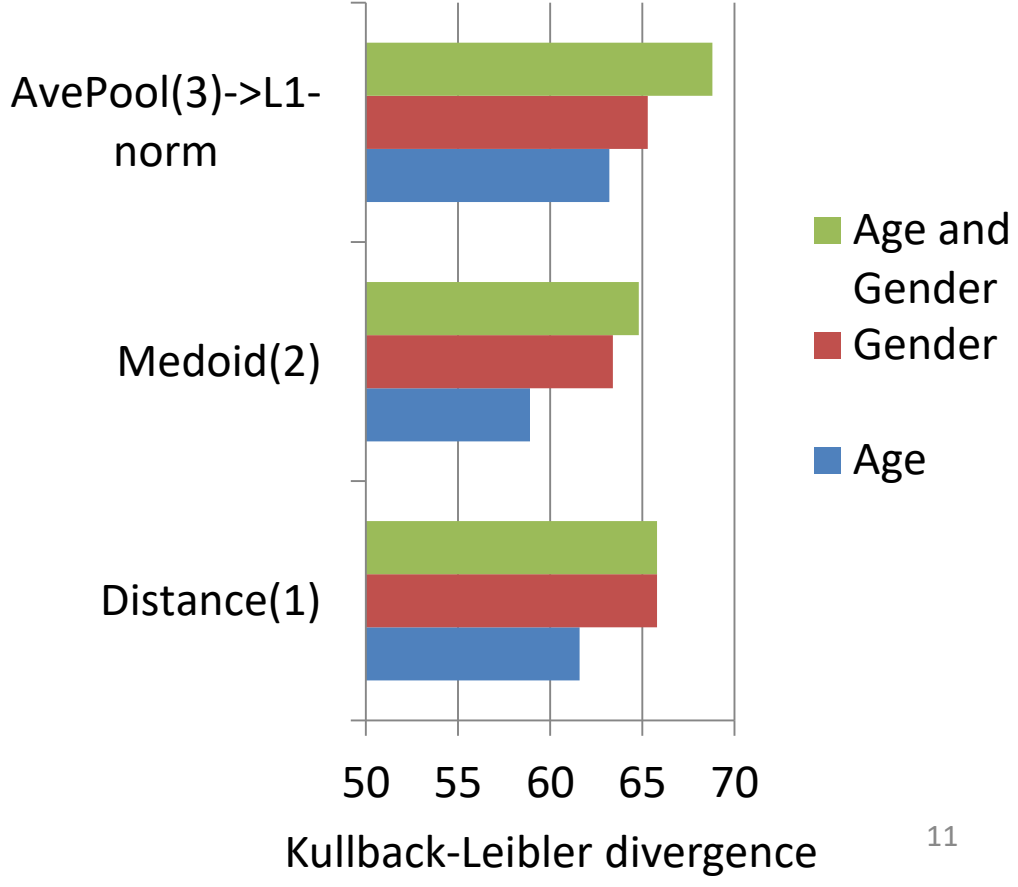


Area under curve (YTF)

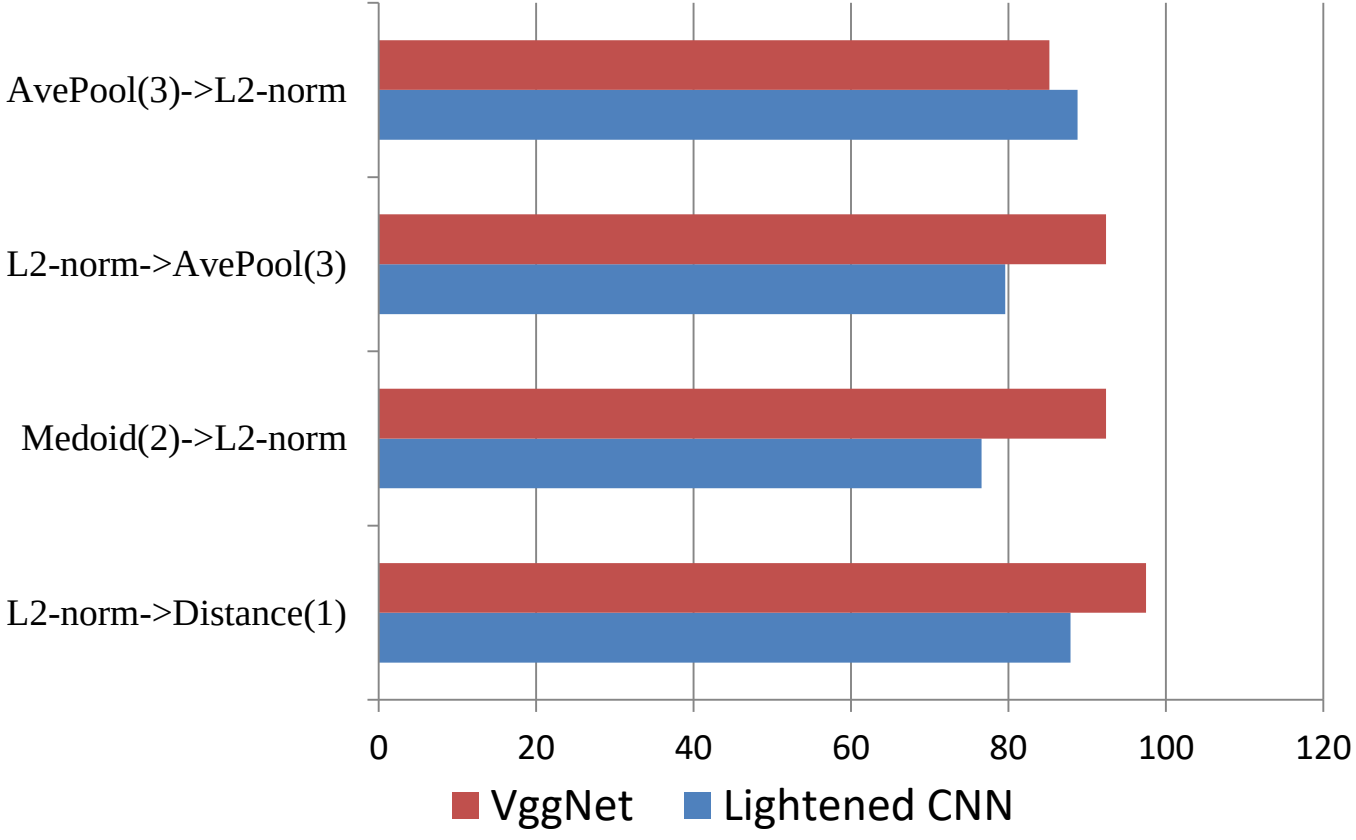


Area under curve (Age and Gender)

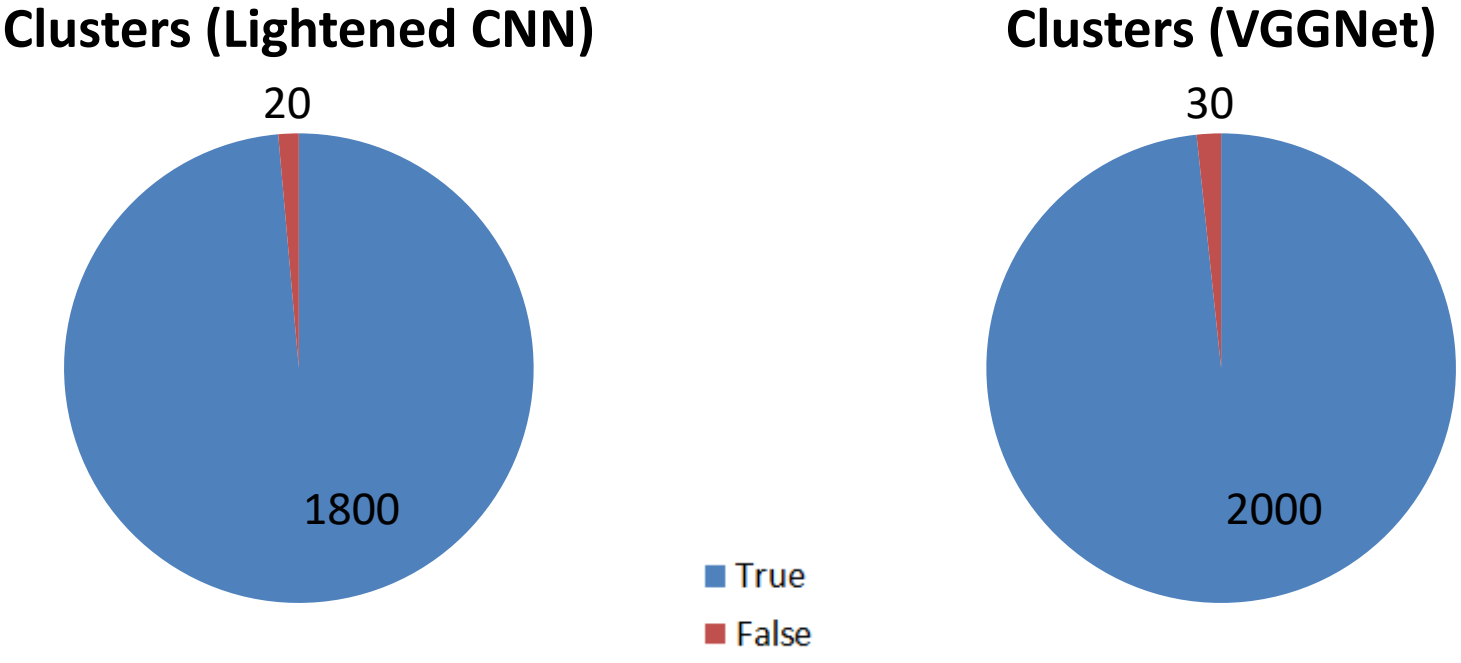




Area under curve (IJB-A)



Clustering results



threshold = 0.0569

Conclusion

- Our algorithm is based on the ways to efficiently compute the dissimilarity of video tracks by using rather simple aggregation techniques
- The most accurate and computationally cheap technique involves the L_2-normed average vector of unnormalized frame features
- Use aggregation of normalized features is usually less accurate

Future work

- Research more sophisticated distances between video tracks, e.g., metric learning or statistical homogeneity testing
- Usage approximate nearest neighbor search
- Introduction the weighing for different features including age and gender probabilities to make our algorithm more accurate

Thank you for attention!