Упорядочивание данных в системах видеонаблюдения на основе технологий глубокого обучения

А.Д. Соколова, А.В. Савченко

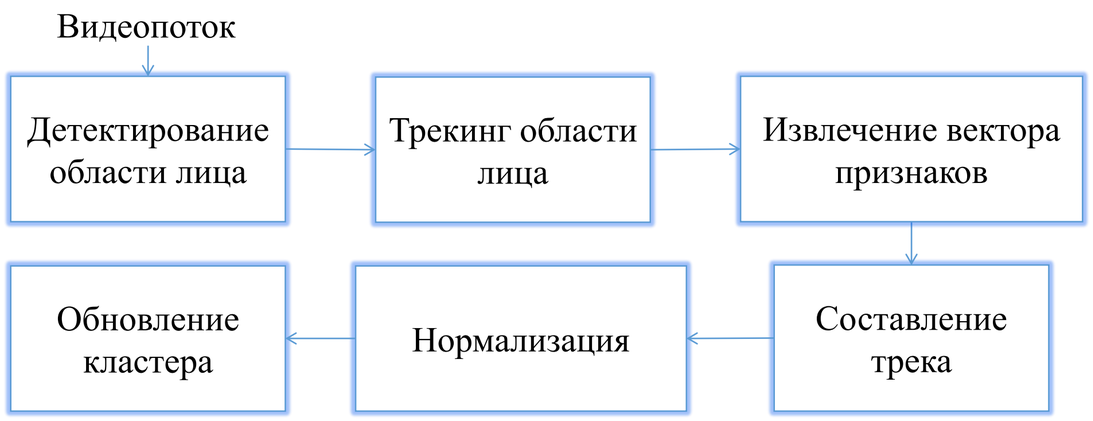
Национальный Исследовательский Университет Высшая Школа Экономики
Нижний Новгород

Аннотация

Рассматривается задача организации информации в системах видеонаблюдения с помощью автоматического выделения групп треков, так, чтобы каждая группа содержала изображения лиц только одного человека. Исследованы методы агрегации векторов признаков каждого кадра, извлекаемых с помощью глубокой сверточной нейронной сети. Треки, содержащие одинаковые лица, группируются с использованием методов верификации лиц и алгоритмов последовательной кластеризации. В экспериментальном исследовании с набором данных YouTubeFaces рассматриваются несколько способов объединения отдельных кадров для получения дескриптора видеодорожки. Показано, что наиболее высокую точность демонстрирует алгоритм сравнения нормализованных признаков, полученных с помощью усреднения векторов признаков всех кадров каждого трека.

Предлагаемый подоход

Задача состоит в том, чтобы разбить набор кадров на M < T последовательных треков $\{X(m)\}, m = 1,2,\dots,M,$ содержащих изображения лица одного человека, а затем объединить похожие треки в кластеры. Каждый m-й трек характеризуется индексами начала $t_1(m)$ и конца $t_2(m)$.



Вычисление расстояния между кадрами

1. Евклидова метрика (L_2)

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

2. Критерий Стьюдента (*t-test*)

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

Методы агрегации

1. Сравнение средних векторов признаков каждого трека

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

2. Усреднение попарных расстояний между всеми кадрами

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

3. Вычисление расстояния между медоидами каждого трека

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

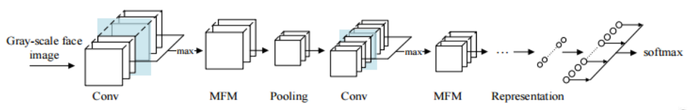
4. Сравнение медиан каждого трека

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

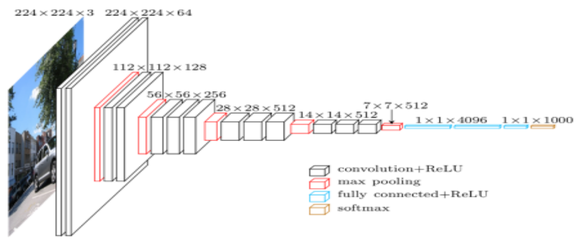
Экспериментальные данные

Сверточные нейронные сети

Lightened CNN (Version C) – 256 элементов



VggNet – 4096 элементов

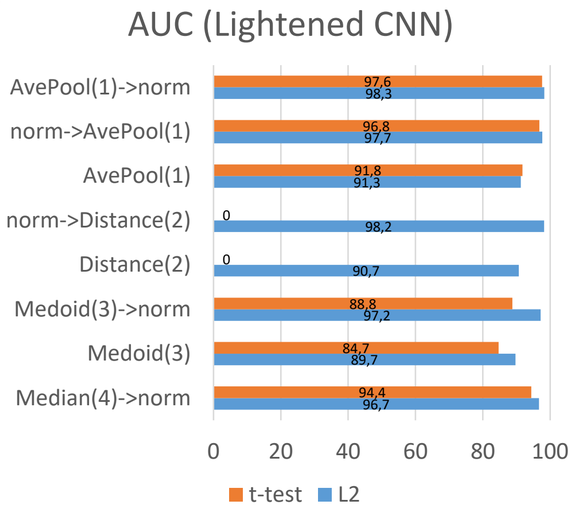


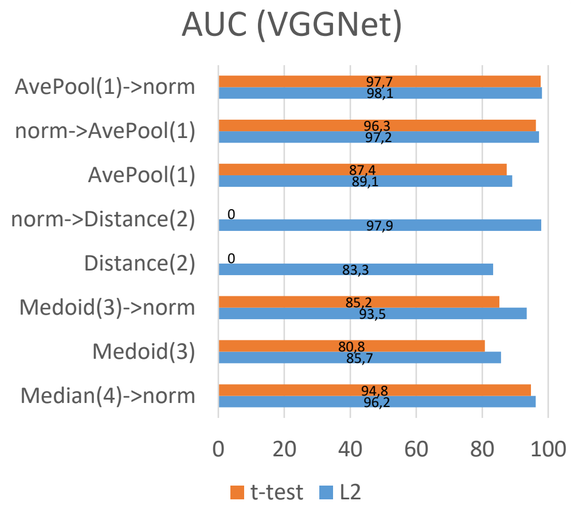
Набор данных

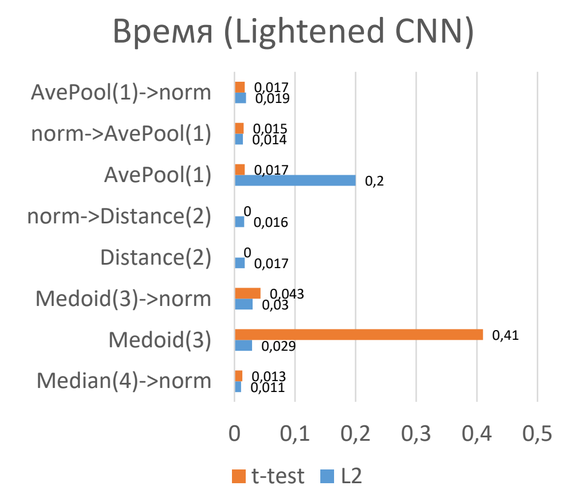
YouTube Faces (YTF):
1595 people
3425 videos
48-6070 frames

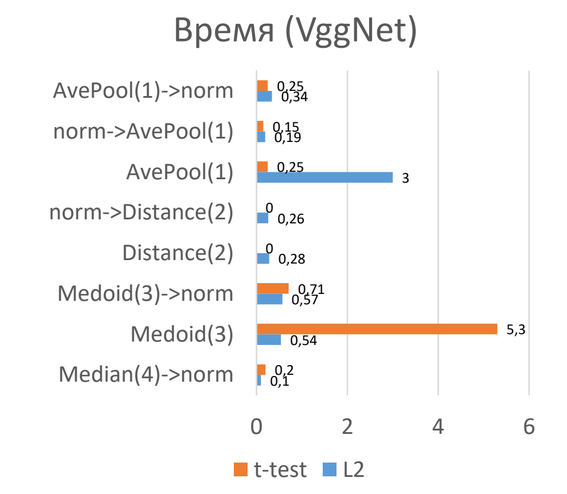
Экспериментальные результаты

Были найдены следующие показатели: AUC (Area under curve), время (сек).

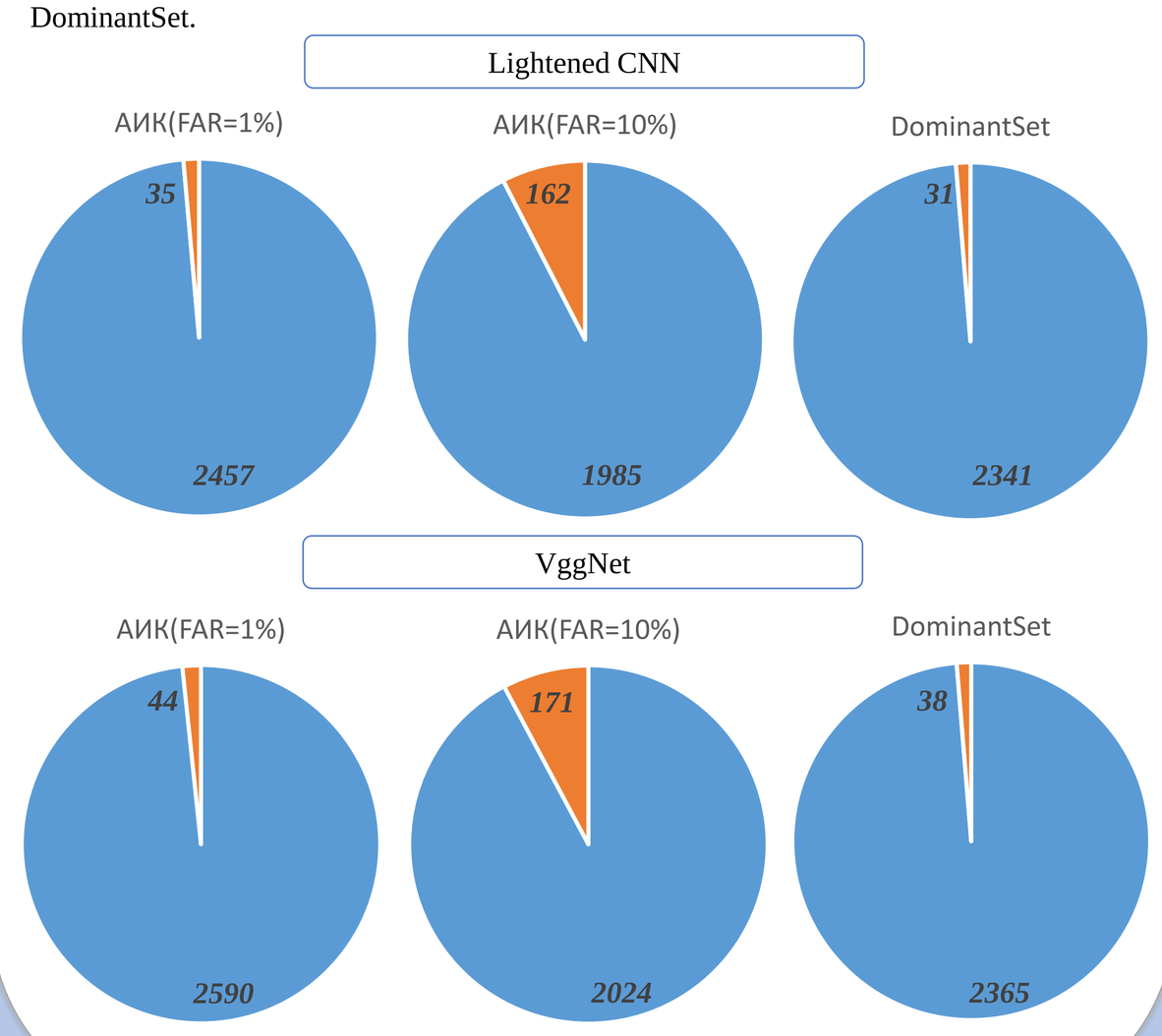








Реализована агломеративная иерархическая кластеризация (АИК), в которой порог для определения результирующих кластеров определялся по фиксированному значению FAR. Кроме того, использовался алгоритм кластеризации из библиотеки DominantSet.



Заключение

Исследована задача кластеризации видеопоследовательностей в системах видеонаблюдения. Основной акцент был сделан на вычислении степени близости видеотреков с использованием агрегации векторов признаков, извлеченных с помощью глубоких сверточных нейронных сетей. Эксперименты продемонстрировали, что наибольшей точностью и вычислительной эффективностью для задачи верификации пользователя по видеоизображению лица характеризуется усреднение векторов признаков всех кадров трека с последующей нормировкой.

ITNT-2018

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
НИЖНИЙ НОВГОРОД