УДК 004.89
ББК 32.813
Ч 25

Частикова Вера Аркадьевна
Доцент, кандидат технических наук, доцент кафедры компьютерных технологий и информационной безопасности института компьютерных систем и информационной безопасности Кубанского государственного технологического университета, Краснодар, e-mail: chastikova_va@mail.ru
Васильев Егор Денисович
Студент института компьютерных систем и информационной безопасности Кубанского государственного технологического университета, Краснодар, e-mail: vasilevegor38@gmail.com
Бабич Дмитрий Валерьевич
Студент института компьютерных систем и информационной безопасности Кубанского государственного технологического университета, Краснодар, e-mail: dmitrii_babich@mail.ru

Нейросетевая методика идентификации лиц в видеопотоке в условиях ограниченности данных

(Рецензирована)

Аннотация. *Рассмотрено применение глубоких сиамских сверточных нейронных сетей для решения задачи детектирования и идентификации лиц в видеопотоке в условиях ограниченности данных с целью снижения сложности модели и количества ошибок на этапе обучения. Биометрические данные лица преобразуются в компактные векторы – эмбеддинги. Для решения проблем детектирования применяется метод, основанный на калибровке оценочных векторов (Non-Maximum Suppression);в области идентификации показано использование методов нормализации и L2-регуляризации с последующим применением функции потерь tripletloss. Для входных данных выполняется батч-нормализация; L2-регуляризация производит минимизацию больших весов с сохранением не подверженных переобучению параметров; в свою очередь, tripletloss используется для минимизации евклидова расстояния между эмбеддингами. Применение данных подходов позволяет добиться высоких показателей точности при достаточно нестандартных ракурсах и неравномерной освещенности изображений лиц.*

Ключевые слова: *идентификация в видеопотоке, сверточная нейронная сеть, сиамская нейронная сеть, эмбеддинг, L2-регуляризация, батч-нормализация, tripletloss.*

Chastikova Vera Arkadyevna
Associate Professor, Candidate of Technical Sciences, Associate Professor of Computer Technologies and Information Security Department, Institute of Computer Systems and Information Security, Kuban State University of Technology, Krasnodar, e-mail: chastikova_va@mail.ru
Vasilyev Egor Denisovich
Student of Institute of Computer Systems and Information Security, Kuban State University of Technology, Krasnodar, e-mail: vasilevegor38@gmail.com
Babich Dmitriy Valeryevich
Student of Institute of Computer Systems and Information Security, Kuban State University of Technology, Krasnodar, e-mail: dmitrii_babich@mail.ru

Methods of identification of persons in the video stream in conditions of limited data

Abstract. *The article describes the use of deep Siamese convolutional neural networks for solving problems of detecting and identifying people with a video stream in conditions of limited data in order to reduce the complexity of models and the number of errors at the learning stage. Biometric data of the person is converted into a compact size – embedding. To solve detection problems, a method is applied based on the calibration of evaluation vectors; in this area, the use of methods for normalization and L2 regulation with loss of triplets is shown. For input data, a batch normalization is performed; L2-regularization minimizes large weights while maintaining parameters that are not subject to retraining; in turn, the loss of triplets is used to minimize the Euclidean distances between embeddings. The use of these approaches allows achieving high accuracy rates with a rather non-standard perspectives and uneven illumination of facial images.*

Keywords: *video stream identification, convolutional neural network, Siamese neural network, embedding, L2 regularization, batch normalization, triplet loss.*

Введение

Старые системы контроля и управления доступом (СКУД) становятся менее эффективными ввиду повышения вычислительных мощностей компьютеров и появления новых

методов их обхода. Наиболее эффективным подходом является реализация интеллектуальных систем видеонаблюдения, это позволяет объединить СКУД и видеонаблюдение в одну комплексную систему. Чаще всего такие системы реализуются при помощи технологий компьютерного зрения и машинного обучения, вследствие чего объемы генерируемой информации возрастают. Интеллектуальные системы видеонаблюдения используются повсеместно, но при этом есть определенные сложности в создании обучающего набора данных, так как получаемая информация никак не размечена. Поэтому *целью данного исследования* является повышение эффективности детектирования и распознавания изображений лиц в условиях ограниченности данных.

Актуальность

Количество получаемой информации в интеллектуальных системах видеонаблюдения растет пропорционально количеству камер, поэтому старые методы контроля и обработки информации становятся неэффективными. Статистика, собранная компанией IHS Markit Ltd по количеству информации, генерируемой камерами, представлена на рисунке 1.

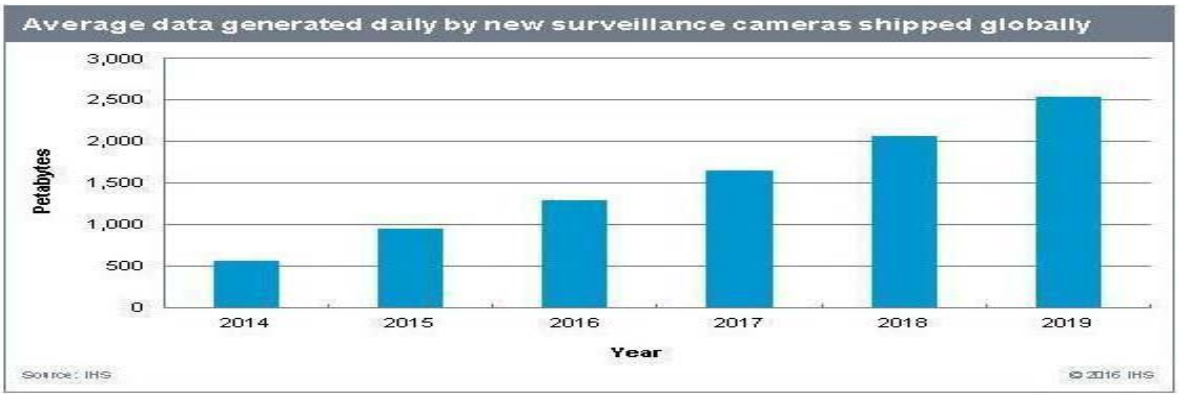

Рис. 1. Статистика по количеству информации, генерируемой камерами

Крупные компании активно развивают рынок анализа данных, чаще всего опираясь на такие модели, как One vs One (Face ID) и One vs All.

Представленная в работе модель занимает перспективную нишу работы с небольшими группами детектируемых и идентифицируемых лиц, то есть она нацелена на решение задачи one vs all в условиях недостаточности данных, характерной для модели one vs one.

Данные (препроцессинг)

Предобработка данных подготавливает изображения для дальнейшего обучения верхних слоев сверточной нейронной сети [1]. Происходит это путем детектирования и выделения лица отдельной сетью (Multi-task Cascaded Convolutional Network (MTCNN)), на выходе получая изображения размером 182x182 пикселей. MTCNN – это многозадачная сверточная нейронная сеть, при последовательном решении нескольких задач и использовании более мощной сверточной нейронной сети повышается точность детектирования лица на основе пяти лицевых точек. Стоит отметить, что важной частью работы является настройка размера матрицы признаков. Для нейросетевой архитектуры, рассмотренной в дальнейшем, была выбрана матрица размером 3x3. Работа многозадачной сверточной нейронной сети разбита на несколько этапов.

На первом этапе используется полная сверточная сеть (Proposal Network). В результате ее работы рассчитываются все возможные варианты изображений лиц и их ограничивающий прямоугольник регрессии векторов. Затем личности калибруются на основе оценочных векторов регрессии ограничивающего окна. Далее используется метод Non-Maximum Suppression (NMS), который объединяет сильно перекрывающиеся лица.

На втором этапе все личности отправляются в сеть Re-fine Network (R-Net), которая дополнительно отклоняет большинство ложных личностей, выполняет калибровку с повтор-

ным сжатием ограничивающего окна и проводит NMS.

На третьем этапе действия аналогичны второму этапу, но используется большая область лица. В результате препроцессинга сеть выведет пять лицевых точек [2].

Архитектура \ обучение

В данной работе для идентификации личности по изображению лица предлагается использовать сиамскую нейронную сеть. Сиамская сеть – это искусственная нейронная сеть, которая использует одинаковые веса в совместной работе на двух разных входных векторах для вычисления сопоставимых выходных векторов [3, 4]. Пример работы сиамской сети продемонстрирован на рисунке 2.

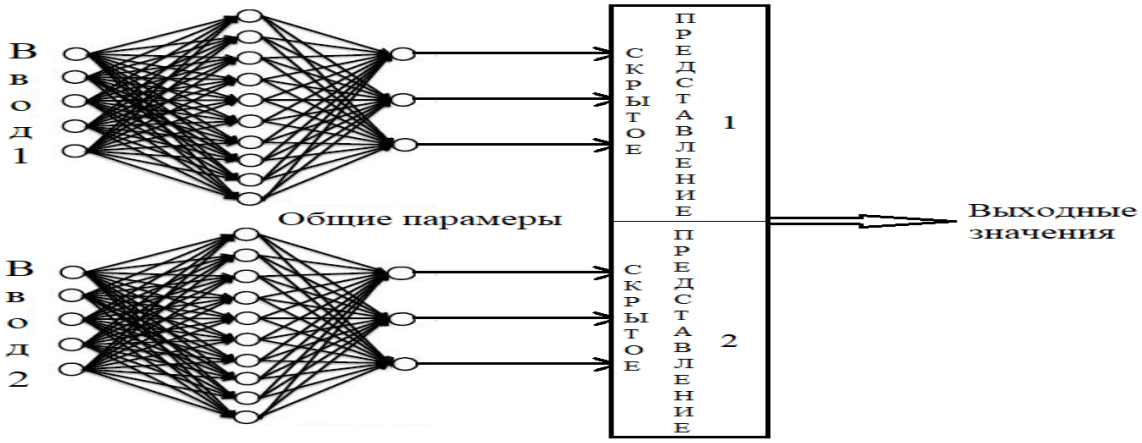

Рис. 2. Пример работы сиамской сети

Достаточно часто предварительно вычисляется один из выходных векторов, таким образом формируя шаблон, с которым сравнивается другой выходной вектор. Этим занимается модель, которая определяет, соответствует ли векторное представление изображений одной и той же личности. Структура модели для решения задачи идентификации личности по изображению лица представлена на рисунке 3. Далее будут рассмотрены составные части предложенной модели.

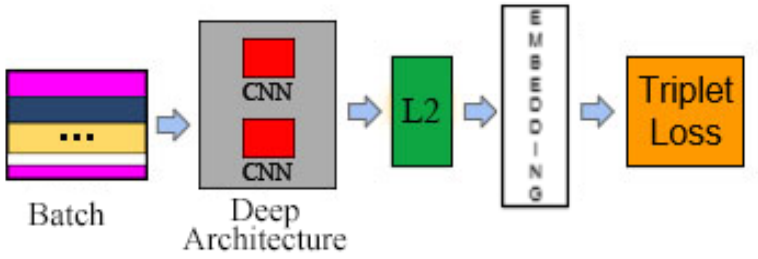

Рис. 3. Структура модели для решения задачи идентификации личности по изображению лица

Представленная в работе модель состоит из пакетного входного слоя и глубокой сверточной нейронной сети (Convolutional Neural Network (CNN)) [5] с последующей нормализацией L2, что приводит к преобразованию изображения лица в эмбеддинг. Далее следует обучение по триплету (набор из 3 изображений идентифицируемого, максимально похожего и максимально непохожего на него). Для задачи идентификации личности характерна проблема настройки гиперпараметров, их настройка регулируется входным набором данных. Предоставленная в работе модель прошла несколько итераций поднастройки.

Батч. По мере распространения сигнала по сети он может исказиться как по математическому ожиданию, так и по дисперсии, что может вызвать несоответствия между градиентами на различных уровнях. Поэтому используются более сильные регуляризаторы, замедляющие темп обучения – в данном случае используется батч-нормализация. Батч – это метод ускорения глубокого обучения, решающий проблему, препятствующую эффективному обу-

чению нейронных сетей. Батч-нормализация предлагает следующее решение – нормализует входные данные, получая нулевое математическое ожидание и единичную дисперсию.

Deep Architecture. Deep Architecture основывается на сверточной нейронной сети. В отличие от обычной нейронной сети [6], слои CNN состоят из нейронов, расположенных в 3-х измерениях – ширине, высоте и глубине. Иначе говоря, в измерениях, формирующих объем. Схематично CNN – это последовательность слоев. Каждый слой преобразует один активационный объем в другой с помощью дифференцируемой функции. Для организации сверточной сети применяются три основных слоя: слой свертки, слой пулинга и полносвязный слой [7].

L2. L2 – регулятор, сокращающий весовые коэффициенты слоев. Обычно первопричиной переобучения является сложность модели, слишком высокая для решаемой задачи и имеющегося обучающего множества. Задача регулятора – понизить сложность модели, сохранив количество ее параметров. Регуляризация выполняется посредством наложения штрафов на веса с наибольшими значениями, минимизируя их.

Эмбеддинг. Векторное представление изображений – это сопоставление произвольной сущности (фрагмента изображения) некоторому вектору. Сопоставив весь набор данных, получаем набор векторов, расположенных в многомерном евклидовом пространстве, имеющих определенное евклидово расстояние между собой.

Пример векторного представления изображений представлен на рисунке 4.

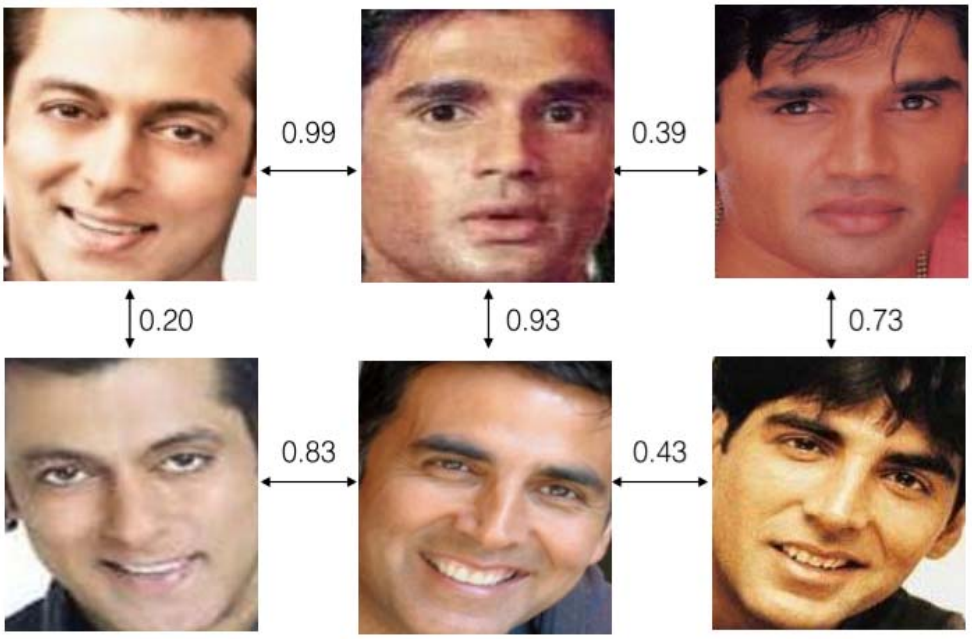

Рис. 4. Пример векторного представления изображений

Важнейшая часть предлагаемого в работе подхода заключается в сквозном обучении всей системы. Для этого используется функция Triplet Loss, которая напрямую отражает повышение эффективности детектирования и распознавания лиц.

Эмбеддинг представляется как $f(x) \in R^D$, где x (изображение эмбеддинга) помещается в d-мерное евклидово пространство (R^D), ограничивающее его нахождение в двумерной гиперсфере, то есть $\|f(x)\|_2 = 1$. Другими словами, необходимо убедиться, что изображение x_i^a (Anchor) конкретного человека ближе ко всем другим изображениям x_i^p (Positive) того же человека, чем к любому изображению x_i^n (Negative) любого другого человека [8]. Наглядное представление результата выполнения функции Triplet Loss показано на рисунке 5.

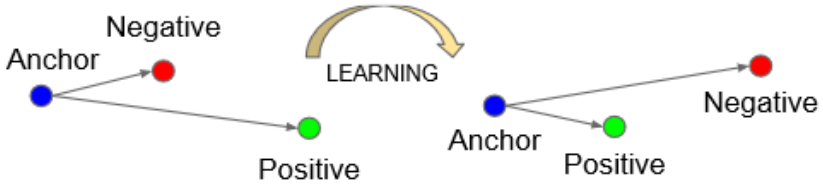

Рис. 5. Представление функции Triplet Loss

Triplet Loss. Triplet Loss минимизирует расстояние между Anchor и Positive, оба из которых имеют одинаковый идентификатор, и максимизирует расстояние между Anchor и Negative другого идентификатора (1).

$$\left\| f\left(x_i^a\right) - f\left(x_i^p\right) \right\|_2^2 + a < \left\| f\left(x_i^a\right) - f\left(x_i^n\right) \right\|_2^2 \qquad \forall \left(f\left(x_i^a\right), f\left(x_i^p\right), f\left(x_i^n\right) \right) \in T, \tag{1}$$

α – поле, которое находится между Positive и Negative и не позволяет функции сходиться к 0 $\left(\left\| f\left(x_i^a\right) - f\left(x_i^p\right) \right\|_2^2 = \left\| f\left(x_i^a\right) - f\left(x_i^n\right) \right\|_2^2 = 0 \right)$.

Отсюда можно сделать вывод, что потери сводятся к минимуму при (2):

$$L = \sum_U^N \left[\left\| f\left(x_i^a\right) - f\left(x_i^p\right) \right\|_2^2 - \left\| f\left(x_i^a\right) - f\left(x_i^n\right) \right\|_2^2 + a \right]. \tag{2}$$

Генерация всех возможных триплетов привела бы к большому множеству триплетов, которые легко выполняются (то есть выполняются ограничения в уравнении (1)). Эти триплеты не будут способствовать обучению и приведут к более медленной конвергенции, поскольку они все равно будут проходить через сеть.

Чтобы обеспечить быструю сходимость, важно выбрать триплеты, которые нарушают ограничение триплета в уравнении (1). Подобные триплеты называются жесткими. Это означает, что, учитывая x_i^a, необходимо выбрать x_i^p (Hard positive) такой, что $\operatorname{argmax}_{x_i^p} \left\| f\left(x_i^a\right) - f\left(x_i^p\right) \right\|_2^2$ и аналогично x_i^n (Hard negative) такой, что $\operatorname{argmin}_{x_i^n} \left\| f\left(x_i^a\right) - f\left(x_i^n\right) \right\|_2^2$. Но невозможно вычислить argmin и argmax во всем тренировочном наборе, так как это может привести к ошибкам в обучении, поскольку неправильно маркированные и плохо изображенные лица будут доминировать над жестко размеченными положительными и отрицательными изображениями лиц.

Есть два варианта, которые позволяют избежать этой проблемы:

1. Создание триплета в автономном режиме каждые n шагов, используя последнюю контрольную точку сети и вычисляя

$$\operatorname{argmax}_{x_i^p} \left\| f\left(x_i^a\right) - f\left(x_i^p\right) \right\|_2^2 \quad \text{и} \quad \operatorname{argmin}_{x_i^n} \left\| f\left(x_i^a\right) - f\left(x_i^n\right) \right\|_2^2$$

на подмножестве данных.

2. Генерацию триплета в режиме реального времени можно сделать, выбрав жесткие положительные / отрицательные образцы из мини-батч.

Основным методом выбора жесткого триплета является онлайн-генерация. Используются большие мини-батчи в порядке нескольких тысяч образцов и рассчитываются только $\operatorname{argmax}_{x_i^p}$ и $\operatorname{argmin}_{x_i^n}$ в мини-батч.

Чтобы иметь осмысленное представление о Anchor-Positive расстояниях, необходимо присутствие в каждом мини-батче минимального количества образцов любого единичного идентификатора. Кроме того, к каждому мини-батчу добавляются случайно выбранные отрицательные изображения лиц (отличные от Anchor изображения).

Выбор Hard-negative на практике может привести к badlocal-минимумам (то есть таких минимумов, у которых не минимальное количество ошибок) на ранних этапах обучения,

в частности, это приводит к схождению модели (то есть $f(x)=0$). Для того чтобы смягчить это, необходимо выбрать x_i^n таким образом, чтобы:

$$\left\| f\left(x_i^a\right) - f\left(x_i^p\right) \right\|_2^2 < \left\| f\left(x_i^a\right) - f\left(x_i^n\right) \right\|_2^2. \quad (3)$$

Такие Negative-точки называются Semi-Hard, так как они находятся дальше от Anchor, чем Positive-образцы, но все же относятся к hard-категории, потому что квадратичное расстояние близко к положительному расстоянию Anchor. Данные негативы лежат внутри поля α.

Заключение

По результатам исследования можно сделать следующие выводы:

1. Правильно подобранные гиперпараметры алгоритма обучения наряду с функцией потерь и функцией активации позволяют значительно повысить точность идентификации и детектирования.
2. Использование таких методов, как батч и L2-регуляризация, значительно ускоряет процесс обучения модели, при этом понизив ее сложность.
3. Выбор жестких триплетов в функции Triplet Loss повышает эффективность обучения модели.

Примечания:

1. Частикова В.А., Жерлицын С.А., Воля Я.И. Нейросетевой подход к решению задачи построения фоторобота по словесному описанию // Известия Волгоградского государственного технического университета. 2018. № 8 (218). С. 63–67.
2. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks / Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Yu Qiao // IEEE Signal Processing Letters (SPL). 2016. Vol. 23, No. 10. P. 1499–1503.
3. Chopra S., Hadsell R., LeCun Y. Learning a similarity metric discriminatively, with application to face verification // IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). 2005. Vol. 1, No. 1. P. 539–546. DOI: 10.11.
4. DeepFace: Closing the Gap to Human-Level Performance in Face Verification / Y. Taigman, M. Yang, M. Ranzato, L. Wolf // IEEE Conference on Computer Vision and Pattern Recognition. 2014. P. 1701–1708. DOI: 10.1109/CVPR.2014.220.
5. Lin M., Chen Q., Yan S. Network in network // International Conference on Learning Representations (ICLR). arXiv preprint arXiv: 1312.4400 2013. No. 2, 4, 6.
6. Сравнительный анализ некоторых алгоритмов роевого интеллекта при обнаружении сетевых атак нейросетевыми методами / В.А. Частикова, М.П. Малыхина, С.А. Жерлицын, Я.И. Воля // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2017. № 129. С. 106–115.
7. Wilson D.R., Martinez T.R. The general inefficiencyof batch training for gradient descent learning // Neural Networks. 2003. No. 16 (10). P. 1429–1451.
8. Weinberger K.Q., Blitzer J., Saul L.K. Distance metriclearning for large margin nearest neighbor classification // NIPS. MIT Press, 2006. No. 2, 3.

References:

1. Chastikova V.A., Zherlitsyn S.A., Volya Ya.I. Neuronetwork approach to the solution of the problem of construction of an identikit using a verbal description // News of Volgograd State Technical University. 2018. No. 8 (218). P. 63–67.
2. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks / Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Yu Qiao // IEEE Signal Processing Letters (SPL). 2016. Vol. 23, No. 10. P. 1499–1503.
3. Chopra S., Hadsell R., LeCun Y. Learning a similarity metric discriminatively, with application to face verification // IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). 2005. Vol. 1, No. 1. P. 539–546. DOI: 10.11.
4. DeepFace: Closing the Gap to Human-Level Performance in Face Verification / Y. Taigman, M. Yang, M. Ranzato, L. Wolf // IEEE Conference on Computer Vision and Pattern Recognition. 2014. P. 1701–1708. DOI: 10.1109/CVPR. 2014. 220.
5. Lin M., Chen Q., Yan S. Network in network // International Conference on Learning Representations (ICLR). arXiv preprint arXiv: 1312. 4400 2013. No 2, 4, 6.
6. Comparative analysis of some algorithms of swarm intelligence when detecting network attacks by neural network methods / V.A. Chastikova, M.P. Malykhina, S.A. Zherlitsyn, Ya.I. Volya // Polythematic Network Electronic Scientific Journal of Kuban State Agrarian University. 2017. No. 129. P. 106–115.
7. Wilson D.R., Martinez T.R. The general inefficiencyof batch training for gradient descent learning // Neural Net-works. 2003. No. 16 (10). P. 1429–1451.
8. Weinberger K.Q., Blitzer J., Saul L.K. Distance metriclearning for large margin nearest neighbor classification // NIPS. MIT Press, 2006. No. 2, 3.