天真 - 深度面部识别：触及LFW基准的限制与否？

|  |  |  |
| --- | --- | --- |
| Erjin Zhou  Face++, Megvii Inc. zej@megvii.com | Zhimin Cao  Face++, Megvii Inc. czm@megvii.com | Qi Yin  Face++, Megvii Inc. yq@megvii.com |

Abstract

*随着最近深度学习技术的发展和潜在的大型训练数据集的积累，人脸识别性能迅速提高。 在本文中，我们报告了我们对大数据如何影响识别性能的观察。 根据这些观察结果，我们构建了Megvii人脸识别系统，该系统在LFW基准测试中达到了99.50％的准确度，优于之前的最新技术水平。 此外，我们还报告了实际安全认证方案中的性能。 机器识别与人类表现之间仍存在明显差距。 我们总结了我们的实验，并提出了近期面部识别中面临的三个挑战。 我们为这些挑战指出了几种可能的解决方案。 我们希望我们的工作能够激发社区对研究基准和实际应用之间差异的讨论。*

# INTRODUCTION

LFW基准[8]旨在测试识别系统在无约束环境中的性能，这比许多其他约束数据集（例如，YaleB [6]和MultiPIE [7]）要难得多。近年来，它已成为关于面部识别在性能评估中的事实标准。已经进行了大量的工作来推动其精度限制[3,16,4,1,2,5,11,10,12,14,13,17,9]。

在整个LFW基准测试的历史中，最近的深度学习技术[17,14,13,10,12]获得了惊人的改进。这些系统的主要框架基于多级分类[10,12,14,13]。同时，开发了许多复杂的方法并应用于识别系统（例如，[4,2,10,12,13]中的联合贝叶斯，[10,14]中的模型集合，[10,12]中的多阶段特征， [10,13]中的联合识别和验证学习。实际上，收集大量外部标记数据用于学习深度网络。不幸的是，在研究大数据和识别性能之间的关系方面几乎没有什么工作。

100

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | Gau | DeepID2 ssianFace | DeepID2+ |
|  |  |  |  | DeepID | DeepFace |
|  |  | TL Joint | Bayesian | FR+FCN |  |
|  |  | High | -dim LBP |  |  |
|  | To Baye | m-vs-Pete  sian Face | Hybrid D Learni | eep ng |  |
| Associa | Re  te-Predict | visited |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| Multiple | LE + com | p |  |  |  |

98

**Accuracy**

96

94

92 #Dataset ~ 10K

#Dataset < 100K

90 #Dataset > 100K

88

86

84

2009 2010 2011 2012 2013 2014 2015

## Year

Figure 1. A data perspective to the LFW history. Large amounts of web-collected data is coming up with the recent deep learning waves. Extreme performance improvement is gained then. How does big data impact face recognition?

这促使我们探索大数据如何影响识别性能。

因此，我们收集大量标记的Web数据，并构建卷积网络框架。获得了两个关键观察结果。首先，数据分布和数据大小确实影响识别性能。其次，我们观察到许多现有复杂方法的性能增益随着总数据大小的增加而降低。

根据我们的观察，我们通过简单直接的卷积网络构建我们的Megvii人脸识别系统，无需任何复杂的调整技巧或智能架构设计。令人惊讶的是，通过利用大型网络收集的标记数据集，这种天真的深度学习系统在LFW上实现了最先进的性能。我们达到99.50％的识别准确度，超过人类水平。此外，我们引入了一个新的基准，称为中国ID（CHID）基准，以探索识别系统的概括。 CHID基准测试旨在测试真实安全证书环境中的识别系统，该环境限制中国人并且需要非常低的误报率。不幸的是，实证结果表明，使用网络收集数据和高LFW性能训练的通用方法并不意味着在这种应用驱动的基准测试中可接受的结果。当我们将假阳性率保持在10-5时，真阳性率为66％，这不符合我们的应用要求。

通过总结这些实验，我们报告了人脸识别中的三个主要挑战：数据偏差，非常低的误报标准和交叉因素。 尽管我们在LFW基准测试中获得了非常高的准确度，但这些问题仍然存在并且将在许多特定的实际应用中被放大。 因此，从工业角度来看，我们讨论了指导未来研究的几种方法。 我们关注的主要是数据：如何收集数据以及如何使用数据。 我们希望这些讨论将有助于进一步研究人脸识别。

# A DATA PERSPECTIVE TO FACE RECOGNITION

LFW基准历史的一个有趣观点（见图1）表明隐含的数据积累是性能改进的基础。 从2010年到2014年，数据量增加了100倍（例如，从多个LE中的大约1万个训练样本[3]到DeepFace中的4百万个图像[14]）。 特别是，随着最近的深度学习浪潮，大量的网络收集数据随之而来，并且随后获得了巨大的性能提升。

我们对这种现象感兴趣。 大数据（尤其是大量网络收集的数据）如何影响识别性能？

# MEGVII FACE RECOGNITION SYSTEM

3.1. Megvii Face Classification Database.

我们收集并标记了大量的名人

互联网，简称Megvii人脸分类

（MFC）数据库。 它有500万个标记的面孔，约有20,000个人。 我们手动删除所有出现在LFW中的人。 图2（a）显示了MFC数据库的分布，这是我们稍后将描述的网络收集数据的一个非常重要的特征。

3.2. Naive deep convolutional neural network.

我们开发了一个简单直观的深层网络架构，在MFC数据库上进行了多类分类。 网络包含十层，最后一层是softmax层，在训练阶段设置用于监督学习。 将softmax层之前的隐藏层输出作为输入图像的特征。 面部的最终表示之后是PCA模型，用于减少特征。

**(a) The Distribution of MFC Database**

0

200

400

600

800

1000

301

601

901

1201

1501

1801

**Number of instances**

**Individual ID**

1

3000 6000 9000 12000 15000

18000

**b) Continued Performance Improvement**

**(**

**c) Long-tail Effect**

**(**



96

96.2

96.4

96.6

96.8

97

97.2

97.4

97.6

3000

6000

9000

12000

15000

18000

**Accuracy**

**Number of Individuals**



97

97.1

97.2

97.3

97.4

97.5

97.6

97.7

3000

6000

9000

12000

15000

18000

**Accuracy**

**Number of Individuals**

Figure 2. Data talks. (a) The distribution of the MFC database. All individuals are sorted by the number of instances. (b) Performance under different amounts of training data. The LFW accuracy rises linearly as data size increases. Each sub-training set chooses individuals randomly from the MFC database. (c) Performance under different amounts of training data, meanwhile each sub-database chooses individuals with the largest number of instances. Long-tail effect emerges when number of individuals are greater than 10,000: keep increasing individuals with a few instances per person does not help to improve performance.

# 我们通过简单的L2范数测量两个图像之间的相似性。

# CRITICAL OBSERVATIONS

我们进行了一系列实验，以探索数据对识别性能的影响。 我们首先研究数据大小和数据分布如何影响系统性能。 然后，我们用以前的文献中出现的许多复杂技术报告我们的观察结果，当时他们提出了大量的训练数据集。 所有这些实验都是用我们的十层CNN建立的，适用于整个脸部区域。

4.1. Pros and Cons of web-collected data

网络收集的数据具有典型的长尾特征：少数“富有”的个体有很多个体，而且很多个体“贫穷”，每个人有少数个体（见图2（a））。在本节中，我们首先探讨总数据大小如何影响最终识别性能。然后我们讨论识别系统中的长尾效应。

持续的性能提升。大量的培训数据大大提高了系统的性能。我们通过培训相同的网络来调查这一点，该网络具有从4,000到16,000的不同数量的个人。从MFC数据库中随机抽取个体。因此，每个子数据库保持原始数据分布。图2（b）显示了每个系统在LFW基准测试中的表现。随着数据量的累积，性能会线性提高。

长尾效应。长尾是网络收集数据的典型特征，我们想知道对系统性能的影响。我们首先按实例数量对所有个体进行排序。然后我们训练同一个网络，不同数量的人从4,000到16,000。图2（c）显示了LFW基准测试中每个系统的性能。长尾确实会影响性能。当我们将具有最多实例的前10,000个人作为训练数据集时，会出现最佳性能。换句话说，添加仅有少数实例的个人无助于提高识别性能。实际上，这些人将进一步损害系统的性能。

4.2. Traditional tricks fade as data increasing.

我们已经探索了以前文献中出现的许多复杂方法，并观察到随着训练数据的增加，这些方法在我们的实验中获得的收益很少。 我们尝试过：

•关节贝叶斯：使用独立的高斯变量对人脸表示进行建模[4,2,10,12,13];

•多阶段特征：将最后几个层的输出组合为面部表示[10,12];

•聚类：用层次结构标记每个人，并用粗标签和精细标签学习[15];

•联合识别和验证：在多类分类框架的隐藏层上添加成对约束[10,13]。

所有这些复杂的方法都会给系统带来额外的超参数，这使得训练更加困难。 但是，当我们通过反复试验将这些方法应用于MFC数据库时，根据我们的实验，与简单的CNN架构和PCA减少相比，获得的收益很少。

# PERFORMANCE EVALUATION

在本节中，我们将系统评估为LFW基准测试和实际安全认证应用程序。 根据我们之前的观察，我们用10,000个最“富有”的人训练整个系统。 我们在四个面部区域上训练网络（即，通过面部界标检测器集中在眉毛，眼睛中心，鼻尖和嘴角）。 图3显示了整个系统的概况。 面部的最终表示是四个特征的连接，然后是PCA以减少特征。

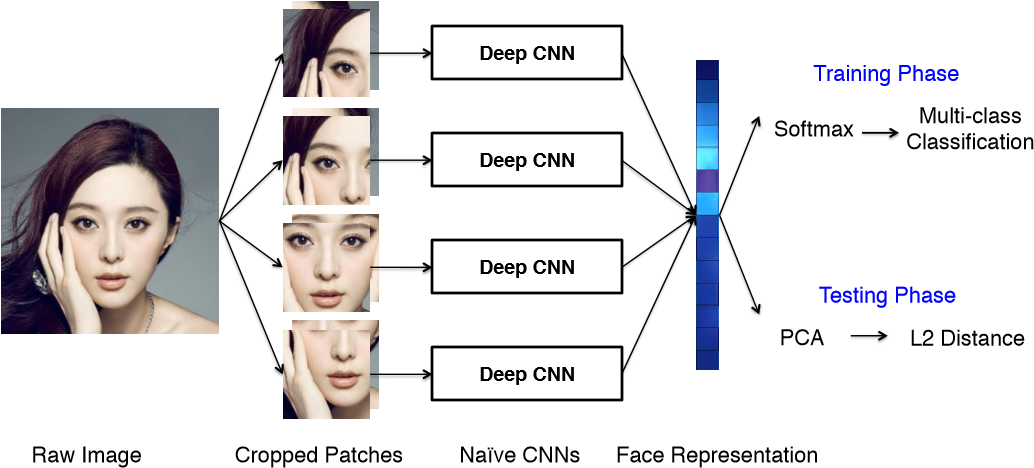


Figure 3. Overview of Megvii Face Recognition System. We design a simple 10 layers deep convolutional neural network for recognition. Four face regions are cropped for representation extraction. We train our networks on the MFC database under the traditional multi-class classification framework. In testing phase, a PCA model is applied for feature reduction, and a simple L2 norm is used for measuring the pair of testing faces.

5.1. Results on the LFW benchmark

我们在LFW基准测试中达到99.50％的准确度，这是现在和人类表现之外的最佳结果。 图4显示了我们系统中的所有失败案例。 除了几对（称为“简单案例”）之外，大多数情况很难区分，甚至是人类。 这些“硬壳”受到几种不同的交叉因素的影响，例如大的姿势变化，化妆很重，玻璃磨损或其他闭塞。 我们指出，没有其他先验（例如，我们已经看过The Hours，所以我们知道棕色头发“Virginia Woolf”是Nicole Kidman），很难纠正最多的保留对。 基于此，如果所有“简单案例”得到解决，我们认为LFW的合理上限约为99.7％。5.2. Results on the real-world application

为了研究识别系统在现实环境中的表现，我们引入了一个新的基准，称为中文ID（CHID）基准。我们离线收集数据集并专注于中国人。与LFW基准测试不同，CHID基准测试是针对中国人的特定领域任务。当我们以非常低的速率保持假阳性时（例如，FP = 10-5），我们对真正的阳性率感兴趣。该基准旨在模拟真实的安全认证环境和测试识别系统的性能。当我们将“99.50％”识别系统应用于CHID基准时，性能不符合实际应用的要求。 “超越人类”系统似乎并没有真正发挥作用。当我们将假阳性率保持在10-5时，真阳性率为66％。图5显示了FP = 10-5标准中的一些失败案例。年龄变化，包括内部变异（即，在不同年龄捕获的同一个人的面部）和相互变化（即，具有不同年龄的人），

|  |
| --- |
| Figure 4. 30 Failed Cases in the LFW benchmark. We present all the failed cases, and group them into two parts. (a) shows the failed cases regarded as “easy cases”, which we believe can be solved with a better training system under the existing framework. (b) shows the “hard cases”. These cases all present some special cross factors, such as occlusion, pose variation, or heavy make-up. Most of them are even hard for human. Hence, we believe that without any other priors, it is hard for computer to correct these cases. |

is a typical characteristic in the CHID benchmark. Unsurprisingly, the system suffers from this variation, because they are not captured in the web-collected MFC database. We do human test on all of our failed cases. After averaging 10 independent results, it shows 90% cases can be solved by human, which means the machine recognition performance is still far from human level in this scenario.

# CHALLENGES LYING AHEAD

根据我们对两个基准的评估，我们总结了人脸识别的三个主要挑战。

数据偏见。网络收集数据的分布极不平衡。我们的实验表明，每个人少数实例的人在一个简单的多类分类框架中不起作用。另一方面，我们意识到大规模的网络收集数据只能提供一个起点;它是人脸识别的基准。大多数网络收集的面孔来自名人：微笑，化妆，年轻，美丽。它远离日常生活中捕获的图像。尽管LFW基准测试具有高精度，但其性能仍然难以满足实际应用中的要求。

误报率极低。真实世界的人脸识别具有比我们在先前的识别基准中处理的更多样化的标准。正如我们之前所说的那样，在大多数安全认证场景中，当假阳性保持在非常低的速率时，客户更关注真正的阳性率。虽然我们在LFW基准测试中实现了非常高的精度，但我们的系统在这些真实环境中仍然远远不能满足人类的性能要求。

交叉因素。 在关于LFW和CHID基准的失败案例研究中，姿势，遮挡和年龄变化是影响系统性能的最常见因素。 然而，我们仍然缺乏对这些交叉因素的充分调查，也缺乏一种有效的方法来清楚和全面地处理它们。

**False**

**Positive**



**False**

**Negative**

Figure 5. Some Failed Cases in the CHID Benchmark. The recognition system suffers from the age variations in the CHID benchmark, including intra-variation (i.e., same person’s faces captured in different age) and inter-variation (i.e., people with different ages). Because little age variation is captured by the web-collected data, not surprisingly, the system cannot well handle this variation. Indeed, we do human test on all these failed cases. Results show that 90% failed cases can be solved by human. There still exists a big gap between machine recognition and human level.

# FUTURE WORKS

大量网络收集的数据帮助我们在LFW基准测试中实现了最先进的结果，超越了人类的表现。但这只是人脸识别的新起点。这一结果的重要性在于表明人脸识别能够走出实验室并进入我们的日常生活。当我们面对实际应用程序而不是简单的基准测试时，我们仍然需要做很多工作。

我们的实验确实强调数据是识别系统中的一个重要因素。我们将以下问题作为工业视角提出，以期对未来的人脸识别研究有所期待。

一方面，开发更智能，更有效的方法挖掘领域特定数据是提高性能的重要方法之一。例如，视频是可以通过自发的弱标记面部提供大量数据的数据源之一，但我们尚未完全探索并将其应用于大规模人脸识别。另一方面，数据合成是生成更多数据的另一个方向。例如，手动收集具有人内年龄变化的数据非常困难。因此，可靠的年龄变化发生器可能会有很大帮助。 3D面部重建也是合成数据的有力工具，尤其是在物理因素建模方面。

我们的观察之一是在简单的多类分类框架中存在长尾效应。如何有效地使用长尾网络收集的数据是一个有趣的问题。此外，如何将通用识别系统转移到特定领域的应用程序仍然是一个悬而未决的问题。

本报告提供了我们关于人脸识别的工业观点，我们希望我们的实验和观察能够激发学术界和工业界的讨论，并进一步改进人脸识别技术。

# References

1. T. Berg and P. N. Belhumeur. Tom-vs-pete classifiers and identity-preserving alignment for face verification. In *BMVC*, volume 2, page 7. Citeseer, 2012.
2. X. Cao, D. Wipf, F. Wen, G. Duan, and J. Sun. A practical transfer learning algorithm for face verification. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3208–3215. IEEE, 2013.
3. Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2707–2714. IEEE, 2010.
4. D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *Computer Vision–ECCV 2012*, pages 566–579. Springer, 2012.
5. D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: Highdimensional feature and its efficient compression for face verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3025–3032. IEEE, 2013.
6. A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.
7. R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.
8. G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.
9. C. Lu and X. Tang. Surpassing human-level face verification performance on lfw with gaussianface. *arXiv preprint arXiv:1404.3840*, 2014.
10. Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2014.
11. Y. Sun, X. Wang, and X. Tang. Hybrid deep learning for face verification. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1489–1496. IEEE, 2013.
12. Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1891–1898. IEEE, 2014.
13. Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. *arXiv preprint arXiv:1412.1265*, 2014.
14. Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701–1708. IEEE, 2014.
15. Z. Yan, V. Jagadeesh, D. DeCoste, W. Di, and R. Piramuthu. Hd-cnn: Hierarchical deep convolutional neural network for image classification. *arXiv preprint arXiv:1410.0736*, 2014.
16. Q. Yin, X. Tang, and J. Sun. An associate-predict model for face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 497–504. IEEE, 2011.
17. Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonical-view faces in the wild with deep neural networks. *arXiv preprint arXiv:1404.3543*, 2014.