HONG KONG JOURNAL OF SOCIAL SCIENCES

香港社會科學學報

第一的第 59 期 (2022 春/夏) Vol. 59 Spring/Summer 2022

Open Access Article

Detection and Recognition of Real-Time Hand Gestures Using YOLO and AlexNet Techniques

Raad Ahmed Mohamed[1,*], Karim Q. Hussein[2,*]

[1] *Computer Science, Iraqi Commission for Computer and Informatics, Informatics Institute for Postgraduate Studies, Baghdad, Iraq*

[2] *Assist. Prof., Computer Science Department, Faculty of Science, Mustansiryha University, Baghdad, Iraq*

Received: March 5, 2022 ▪ Reviewed: May 2, 2022

▪ Accepted: June 8, 2022 ▪ Published: July 29, 2022

Abstract:
At least three thousand five hundred million people in the world cannot hear or speak, they are called the deaf and dumb. Often this segment of society is partially isolated from the rest of community due to the difficulty of dealing, communicating and understanding. To address this problem, many solutions have been proposed that try bridging this gap between this segment and the rest of society and using the technical development of devices, especially computers and mobile phones. The science of image processing with artificial intelligence has been used to generate programs for converting natural speech into a sign language that enables the people with disabilities (the deaf and dumb) understand it. Initially, a set of sign dictionaries were made in the deaf and dumb language, including the Indian Sign Language (ISL), the American Sign Language (ASL), the European Dictionary and so on, and the main reason for this is to simplify the understanding of the sign language. These dictionaries depend mainly on the movement of hands, and one or both hands can be used to form special characters for this conversation. These dictionaries can be used after training this segment of people. Because of technological progress, this process can be automated using a computer and various programming languages in addition to secondary connections (cameras and microphones) and advanced algorithms for artificial intelligence. This goal can be reached by building a special program for communication between people with disabilities (deaf and dumb) and healthy people, or both directions. The research results show that the use of neural networks, especially convolutional neural networks, is very suitable in terms of accuracy, speed of performance and generality in processing the previously unused input data.

Keywords: the deaf and dumb, Indian sign language, American sign language, hand gesture, artificial intelligence, hand detection, convolutional neural networks.

Corresponding Authors: Raad Ahmed Mohamed, Computer Science, Iraqi Commission for Computer and Informatics, Informatics Institute for Postgraduate Studies, Baghdad, Iraq; email: Raadahmed130@yahoo.com; Karim Q. Hussein, Assist. Prof., Computer Science Department, Faculty of Science, Mustansiryha University, Baghdad, Iraq; email: Karimzzm@yahoo.com, karim.q.h@uomustansiriyha.edu.iq

使用优洛和亚历克斯网技术检测和识别实时手势

摘要:
世界上至少有三千五亿人不能听不能说，他们被称为聋哑人。由于交易、沟通和理解的困难，这部分社会通常与社区的其他部分部分隔离。为了解决这个问题，已经提出了许多解决方案，试图弥合这一群体与社会其他人之间的差距，并利用设备的技术发展，尤其是计算机和手机。人工智能图像处理科学已被用于生成将自然语音转换为手语的程序，使残疾人（聋哑人）能够理解它。最初，用聋哑语制作了一套手语词典，包括印度手语（ISL）、美国手语（手语）、欧洲词典等，主要是为了简化手语对手语的理解。这些字典主要依赖于手的动作，可以用一只或两只手来组成这个对话的特殊字符。这些字典可以在训练这部分人之后使用。由于技术进步，除了辅助连接（相机和麦克风）和人工智能的高级算法之外，这个过程还可以使用计算机和各种编程语言实现自动化。这个目标可以通过建立一个特殊的项目来实现残疾人（聋哑）和健康人之间的交流，或双向交流。研究结果表明，使用神经网络，尤其是卷积神经网络，在处理以前未使用的输入数据时，在准确性、性能速度和通用性方面非常适合。

关键词：聋哑人，印度手语，美国手语，手势，人工智能，手部检测，卷积神经网络.

1. Introduction

According to the World Health Organization (WHO), the world's deaf population is estimated to be 250 million individuals (WHO, 2001). This group of people communicate with one another through symbolic language. This pictorial language is referred to as a sign language. Sign Language was developed to assist deaf or hard-of-hearing individuals in communicating with others. Sign language is not an isolated language, and it has significantly advanced over the years. Since the introduction of sign languages, there is evidence of communication among hearing-impaired people (Emond et al., 2015)

Numerous countries have developed their own sign languages, among them the American Sign Language, the French Sign Language, the Indian Sign Language, and the Puerto Rican Sign Language. Gesture-based communication varies by place and contains substantial distinctions from other languages. Understanding the sign language is critical for communicating with hearing challenged individuals and their families. Inadequate comprehension creates substantial barriers to understanding this culture and may result in miscommunication.

The sign language is a collection of signs that the signing deaf community uses to communicate with one another. Sign language is a combination of gestures, movements, signs, and facial expressions that match to normal language letters and words. Sign languages are not universal. Each country has its own distinctive sign language. As is the case with India, each country has its own sign language (ISL), which is not a visual representation of English (Bhame et al., 2014a).

Sign language recognition systems serve as an intermediary between the hearing and deaf-dumb communities. It catches the signer's gesture and converts it to a recognized language. The term "Hand Gesture Recognition" refers to the process of tracking and comprehending human gestures in the form of meaningful orders. The detection of hand gestures using vision is classified into two categories: gestures, both static and dynamic (Bhame et al., 2014b).

Static hand gestures are the position and orientation of the hand in space for a specified amount of time without movement, whereas dynamic hand gestures contain actions of the hand, such as waving or finger movements. This article proposes a method for hand gesture identification that uses an image database for comparison; it simply records the input image using a simple web cam and then processes and recognizes the gesture using YOLO and ELEXNET for each gesture. This enables it to recognize objects quickly and so be employed in real-time applications. The suggested system constructs a real-time PC-based system for automatic recognition of static hand gestures in Indian (Devanagari) Sign Language by using image processing.

2. Related Works

Recent advances in deep learning have resulted in the presentation of accurate and efficient models for real-time applications. The most accurate systems, on the other hand, create many modalities, such as optical flow, from RGB input frames. Due to the high computational cost of method, real-time performance is hindered. We avoid optical flow calculations entirely by developing a real-time hand motion detection technique based on RGB frames and hand segmentation masks. We employ a lightweight semantic segmentation technique (FASSD-Net) to enhance the accuracy of two very efficient HGR methods: Temporal Segment Networks (TSN) and Temporal Shift Modules (TSM) (TSM). We demonstrate the proposal's effectiveness by examining our IPN Hand dataset, which contains

thirteen diverse gestures for interacting with touchless displays. The experimental results show that our technique significantly surpasses the original TSN and TSM algorithms in terms of accuracy while retaining real-time performance (Benitez-Garcia et al., 2021).

In 2014, Shreya Shi Narayan Sawant used MATLAB to create a real-time Sign Language Recognition system capable of detecting 26 different Indian Sign Language motions. A webcam was used to catch the signs. To extract features from these indicators, they are pre-processed using the HSV color model. The resulting characteristics are compared using the Principal Component Analysis (PCA) approach. After comparing the captured sign's properties to those in the testing database, the least Euclidean distance is determined for sign recognition. Finally, the recognition of gestures is converted into written and audio formats. This technology enables deaf-blind individuals to communicate (Sawant, 2014). Traditional methods, such as skin detection, picture filtering, image segmentation, and template matching have been used to identify hand gestures. The device is capable of deciphering (ASL) sign language used in the United States of America (Farzi and Tarjomannejad, 2015).

3. Research Method

In the first section of the software, the palm of the hand is determined from the captured image stream using artificial intelligence and image processing. This section contains the following:

Using various cameras, photographing the individual is being addressed by hand or by pointing.

- Scanning the image to determine the areas of the right and left palm. This part of the operation is very difficult, as we need a high level of professionalism to isolate this part of the hand only without the rest of the body parts. In the traditional methods used, the hand, forearm and palm are identified and isolated in a way of distinguishing colors, especially the color of the natural skin of the human being. This affects the degree of color, distinguishing the color and thus the difficulty of identifying this part of the hand. Because of these reasons and the failure of traditional methods, the use of artificial intelligence methods represented by neural networks of various types (such as Yolo-NET, R-CNN,) had to be used, to determine the optimal model to achieve standards of accuracy and processing speed. After comparing the results obtained from these models, it was found that the optimal network type for the performance of this section is the YOLO-2 neural network, shown in Figure 1.

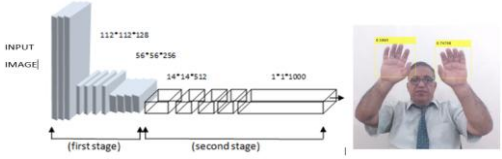

Figure 1. YOLO-2 neural network architecture

Part Two: Creating a graphical dictionary of the letters used in deaf and dumb education, since the Indian dictionary, which contains 26 letters, is insufficient, as shown in Figure 2, it was adopted by adding a symbol to the space, increasing the total number of symbols to 27 and the number of images representing these symbols to thirty thousand for usage in this stage of the program. After generating the image database for the dictionary, neural networks such as ALEX NET, Vgg, Dark NET were used to identify (classification) the images of characters extracted from the first part of the program after experimenting with the traditional methods used for classification after isolating the

Parameters of the character images, such as (SURF, ORB, SIFT & Hu Moments) failed to recognize characters when changing the shape, orientation, the rotation of the images used (global). The most used networks could recognize images and identify the corresponding symbols from the dictionary well, and the best among them was the ALEX type neural network, which was adopted in this part of the program.

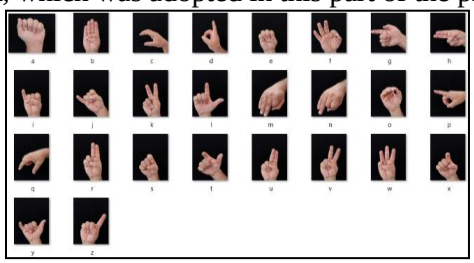

Figure 2. Sample of sign language characters: Manual alphabet

The second part of the work deals with communication from the natural person to the physically disabled (deaf and dumb). The available libraries (open source) have been used within Google, IBM, and Microsoft, and the Google Library has been approved for its dependence on cloud computing. The main structure of this part consists of the audio-receiving unit, converting it into written text, and then converting the text into the corresponding visual form, according to the approved dictionary of signs. The part of converting audio to written text is shown in Figure 3.

The Application Program Interface (API) has been linked with an appropriate program that was written to break the text into its main parts of terms, words and different characters.

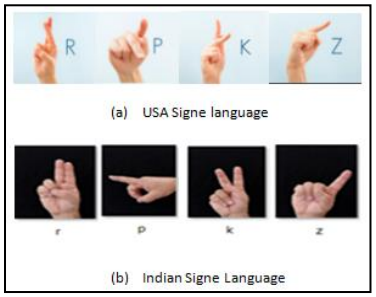

Figure 3. Converting audio to written text & hand gesture images

As for the last step, it is to convert these parts to the pictorial and video forms after creating a pictorial and video dictionary of terms and characters. It is appropriate to mention that it is easy to add any abbreviations and any number of abbreviations to be used without resorting to the blindness of learning the neural network, where the greatest weight lies in distinguishing the sound and transferring it to the used library.

3.1. Indian Sign Language (ISL)

We are all aware that communication is a critical aspect of expressing one's emotions or conveying information to another. While the ability to listen and speak is the most critical component of communication, many of us are unfortunate because we were not born with this God-given ability. These individuals were deaf and illiterate. However, sign language is used by only 2.5 million to 5.0 million people, limiting the number of people with whom they can easily communicate (Mitchell et al., 2012).

ISL is similar to English Sign Language in those persons who understand one may communicate effectively in the other (NIDCD, 2019). Apart from the four gesture characters, their sign languages contain distinctions, as there are variances between Indian sign languages, US sign languages, and Australian sign languages (K, P, R, Z), as shown in Figure 4.

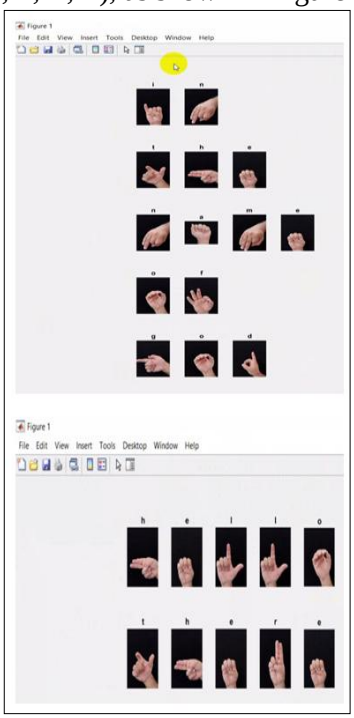

Figure 4. (a) USA sign language characters, (b) Indian sign language characters

3.2. Collection of Data and Abbreviations

Because of the lack of a uniform dataset for all countries/subcontinents, a dataset for letter Sign Language was produced. The current conditions necessitate the development of a large vocabulary dataset. Additionally, this information is incomplete, necessitating the incorporation of additional hand movements. In the future, this research will assist other researchers in creating their own datasets tailored to their specific needs (Sahoo et al., 2014).

This dataset comprises 26 characters ranging from A to Z with various abbreviations. Using both hands, I captured 1200 photos for each alphabet and abbreviation. It was saved in a separate file for letters or abbreviations, and each of these files was named after the letter or abbreviation stored. Figure 5 shows a dataset alphabet images and many abbreviations.

The photos were flipped from right to left using a code. While the height and width ratios vary considerably, the average image size is roughly 227x227 pixels.

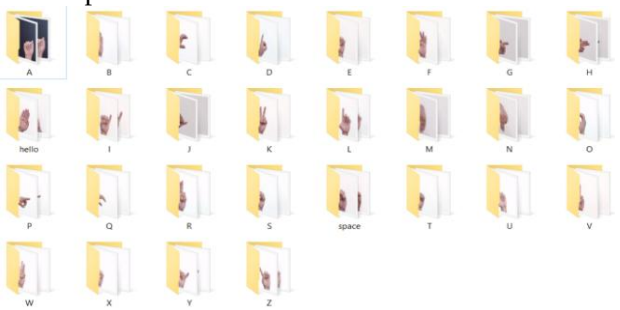

Figure 5. Files with a dataset of the alphabet images and abbreviations

Over 31,200 photos of letters images taken using a web camera are included in the collection. Additionally, there were a few abbreviations. The way to add the images of abbreviations to this dataset is shown in Figure 6.

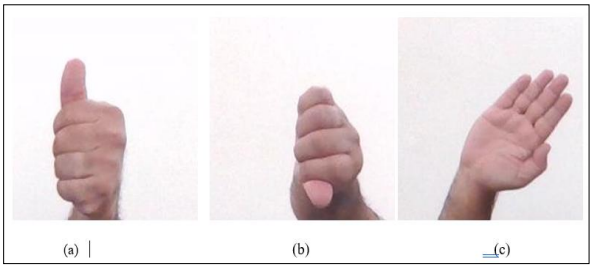

Figure 6. Images of abbreviations: (a) yes, (b) no, and (c) hello abbreviations

4. Results

Using our self-generated dataset, we achieved to detect hand gestures for 26 letters using the YOLO Network, as shown in Figure 7. Additionally, 99.70% accuracy was achieved for the trained data with one epoch, iteration (1020) and hardware resource single (GPU), as shown in Figure 8.

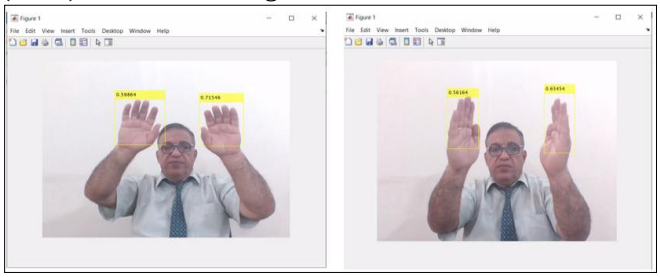

Figure 7. An application performing real-time hand gesture detection

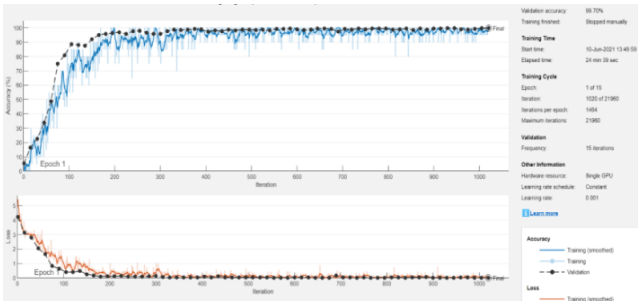

Figure 8. Accuracy achieved for the trained data

The current approach is capable of identifying static gestures. Moreover, this method is capable of classifying any individual's distinctive sign gestures. We classified hand motions to text using the AlexNet algorithm. Figure 9 shows the application working and detecting hand gestures for other sign alphabet characters. Real-time testing was performed under different lighting conditions and normal backgrounds.

The second stage (character classifier) uses a CNN network similar to the ALEX network to categorize an all-sign language dictionary (Indian dictionary) that includes the signs for space and end that are used to split phrases and end sentences. Generally, the structure of the used net is well-known, and it was changed to address the issue at hand. Such networks require time in the initial stage and the effective network parameters in the initial stage (training sets and epoch number).

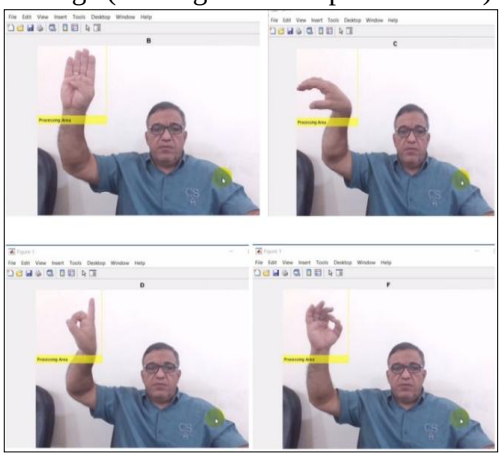

Figure 9. The application working and detecting hand gestures for sign (B, C, D, F) alphabet character

The training set for each sign character consists of 1200 images covering different scale, orientations, and reflections and a total of more than 32000 training images, shown in Table (1) summaries the training procedure, using ELEX Network for classification performance of the trained data using the classifier.

Table 1. Alex network training schedule

Epoch	Iteration	Time (h:m:s)	Mini batch RMSE	Mini batch LOSE	Base Learning Rate
1	1	00:00:02	5.86	34.3	0.0010
1	10	00:00:4	0.95	0.9	0.0010
Continuation of Table 1					
2	20	00:00:07	0.86	0.7	0.0010
2	30	00:00:09	0.71	0.5	0.0010
3	40	00:00:12	0.75	0.6	0.0010
3	50	00:00:15	0.71	0.5	0.0010
4	60	00:0019	0.59	0.4	0.0010
5	70	00:00:22	0.59	0.4	0.0010
5	80	00:00:24	0.60	0.4	0.0010
6	90	00:00:27	0.57	0.3	0.0010
6	100	00:00:30	0.51	0.3	0.0010
7	110	00:00:36	0.51	0.3	0.0010
8	120	00:00:38	0.46	0.2	0.0010
8	130	00:00:41	0.50	0.2	0.0010
9	140	00:00:44	0.46	0.2	0.0010
9	150	00:00:47	0.39	0.1	0.0010
10	160	00:00:49	0.44	0.2	0.0010

5. Conclusion

The basic purpose of interactive virtual environments is to enable natural, efficient and, adaptive communications between the user and computer. Human gestures, including the finger positions and movements, hands, and arms, are one of the most abundant non-verbal communication modalities that enable natural interaction between humans and their virtual environment. Sign gestures can be static, in which the human adopts a particular position, or dynamic, in which the hands and fingers move in unison. Using YOLO and AlexNet Networks, this study proposes a real-time system that consists of two modules: hand detection and gesture identification.

The results demonstrate that the system is capable of achieving acceptable real-time performance and classification accuracy of 100 percent for letters with varying scale, orientation, and lighting, as well as a complex background. Three critical factors determine the system's accuracy: the quality of the webcam used to capture photos for the dataset, the number of training images, and the choice of the Conventional Neural Network model.

References

[1] BENITEZ-GARCIA, G., PRUDENTE-TIXTECO, L., CASTRO-MADRID, L.C., TOSCANO-MEDINA, R., et al. (2021) Improving Real-Time Hand Gesture Recognition with Semantic Segmentation. *Sensors*, 21(2), 356. https://doi.org/10.3390/s210203

[2] BHAME, V., SREEMATHY, R., & DHUMAL, H. (2014a) Vision Based Calculator for Speech and Hearing-Impaired using Hand Gesture Recognition. *International Journal of Engineering Research & Technology*, 3(6), 632-635.

[3] BHAME, V., SREEMATHY, R., & DHUMAL, H. (2014b) Vision Based Hand Gesture Recognition Using Eccentric Approach for Human Computer Interaction. *Proceedings of the 2014International Conference on Advances in Computing.*

Communications and Informatics (ICACCI), 949-953, https://doi.org/10.1109/ICACCI.2014.6968545.

[4] EMOND, A., RIDD, M., SUTHERLAND, H., ALLSOP, L., et al. (2015) The current health of the signing Deaf community in the UK compared with the general population: a cross-sectional study. *BMJ Open*, 5(1), article ID e006668.

[5] FARZI, A., and TARJOMANNEJAD, A. (2015) Prediction of phase equilibria in binary systems containing acetone using artificial neural network. *International Journal of Scientific & Engineering Research*, 6(9), 358-363.

[6] MITCHELL, R. E., YOUNG, T. A., BACHELDA, B., & KARCHMER, M. A. (2006) How Many People Use ASL in the United States?: Why Estimates Need Updating. *Sign Language Studies*, 6(3), 306-335.

[7] NATIONAL INSTITUTE OF DEAFNESS AND OTHER COMMUNICATION DISORDERS. (2019) American Sign Language. U.S. Department of Health and Human Services; Available from https://www.nidcd.nih.gov/health/american-sign-language:

[8] SAHOO, A., MISHRA, G., & RAVULAKOLLU, K. (2014) Sign language recognition: State of the art. *ARPN Journal of Engineering and Applied Sciences*, 9, 116-134.

[9] SAWANT, S.N. (2014) Sign Language Recognition System to aid Deaf-dumb People Using PCA. *International Journal of Computer Science & Engineering Technology*, 5(05), 570-574.

[10] WHO calls on private sector to provide affordable hearing aids in developing world. (2001) *Indian Journal of Medical Sciences*, *55*(9), 511–513. Available from https://www.ncbi.nlm.nih.gov/pubmed/11887302.

参考文:

[1] BENITEZ-GARCIA, G.、PRUDENTE-TIXTECO, L.、CASTRO-MADRID, L.C.、TOSCANO-MEDINA, R. 等。(2021) 通过语义分割改进实时手势识别。传感器，21（2），356。https://doi.org/10.3390/s210203

[2] BHAME, V., SREEMATHY, R., & DHUMAL, H. (2014a) 使用手势识别的语音和听力障碍视觉计算器。国际工程研究与技术杂志，3(6)， 632-635。

[3] BHAME, V., SREEMATHY, R., & DHUMAL, H. (2014b) 使用偏心方法进行人机交互的基于视觉的手势识别。2014 年国际计算进步会议论文集。通信和信息学(国际商会)， 949-953，https://doi.org/10.1109/ICACCI.2014.6968545。

[4] EMOND, A., RIDD, M., SUTHERLAND, H., ALLSOP, L. 等。(2015) 与普通人群相比，英国聋人社区目前的健康状况：一项横断面研究。英国医学杂志公开赛，5(1)，文章 ID e006668。

[5] FARZI, A. 和 TARJOMANNEJAD, A. (2015) 使用人工神经网络预测含有丙酮的二元系统中的相位平衡。国际科学与工程研究杂志，6(9)，358-363。

[6] MITCHELL, R. E., YOUNG, T. A., BACHELDA, B., 和 KARCHMER, M. A. (2006) 在美国有多少人使用手语？：为什么估计需要更新。手语研究，6(3)，306-335。

[7] 国家耳聋和其他交流障碍研究所。（2019）美国手语。美国卫生与公众服务部；可从 https://www.nidcd.nih.gov/health/american-sign-language 获得：

[8] SAHOO, A.、MISHRA, G. 和 RAVULAKOLLU, K. (2014) 手语识别：最先进的技术。ARPN 工程与应用科学杂志，9，116-134。

[9] SAWANT, S.N. (2014) 手语识别系统帮助聋哑人使用主成分分析。国际计算机科学与工程技术杂志，5(05)，570-574。

[10] 世卫组织呼吁私营部门在发展中国家提供负担得起的助听器。(2001) 印度医学科学杂志，55(9)，511-513。可从 https://www.ncbi.nlm.nih.gov/pubmed/11887302 获得。