Using Google OCR to digitize Myanmar

Nyein Chan Ko Ko

Tech for Change Team,

Phandeeyar

Phandeeyar

What is digitization?

- Converting information into Digital format .
- Text , Image, Audio ,Documents, Signal etc..

What are the benefits of digitization?

- Searchability (ရှာဖွေလို့ရတယ်)
- Editability (ပြန်ပြင်လို့ရတယ်(
- Accessibility(လူတိုင်း ရယူကြည့်ရှုနိုင်တယ်)
- Storability (သိမ်းထားလို့လွယ်တယ်)
- Backups (အရန်သိမ်းထားလို့ရတယ်)
- Translatability (ဘာသာပြန်လို့ရတယ်)

Phandeeyar

OCR

Optical Character Recognition

(စာလုံးပုံရိပ်များကို သိမှတ်နိုင်သော ကွန်ပျူတာနည်းပညာ)

ပြင်ပတွင်ရှိသော စာအုပ်များ၊ လက်ရေးမူများ၊ ပုံရိပ်များ ကို
ကွန်ပျူတာနားလည်နိုင်သော ကုဒ်များအဖြစ် ပြောင်းလဲခြင်း

Phandeeyar

Library Preservation

Phandeeyar

Google World Brain

Workers at a Google "scanning factory" digitize books for the online library project documented in Google and the World Brain. (PHOTO COURTESY BEN LEWIS)

Phandeeyar

ဘယ်လိုအသုံးပြုသလဲ

- Data entry for business documents, e.g. check, passport, invoice, bank statement and receipt
- Automatic number plate recognition
- In airports, for passport recognition and information extraction
- Automatic insurance documents key information extraction
- Extracting business card information into a contact list[9]
- More quickly make textual versions of printed documents, e.g. book scanning for Project Gutenberg
- Make electronic images of printed documents searchable, e.g. Google Books
- Converting handwriting in real time to control a computer (pen computing)
- Defeating CAPTCHA anti-bot systems, though these are specifically designed to prevent OCR.[10][11][12] The purpose can also be to test the robustness of CAPTCHA anti-bot systems.
- Assistive technology for blind and visually impaired users

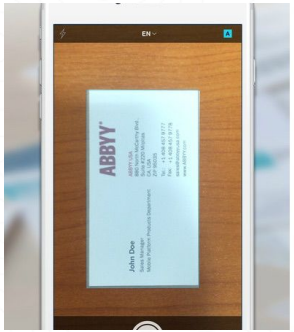
ဘယ်လိုအသုံးပြုသလဲ

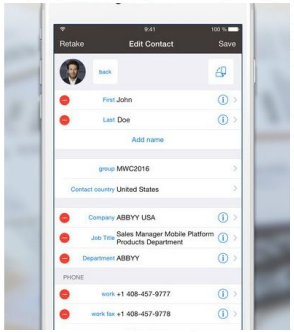
Handwritten text to digital text

Mail sorting machine with OCR

OCR for cars (number plate reader)

Passport recognition (OCR)





Business card reader (OCR)

Text reader for visually impaired.

What else can OCR do?

Digitisation + Big Data+AI+ Low Cost Internet = DIGITAL REVOLUTION

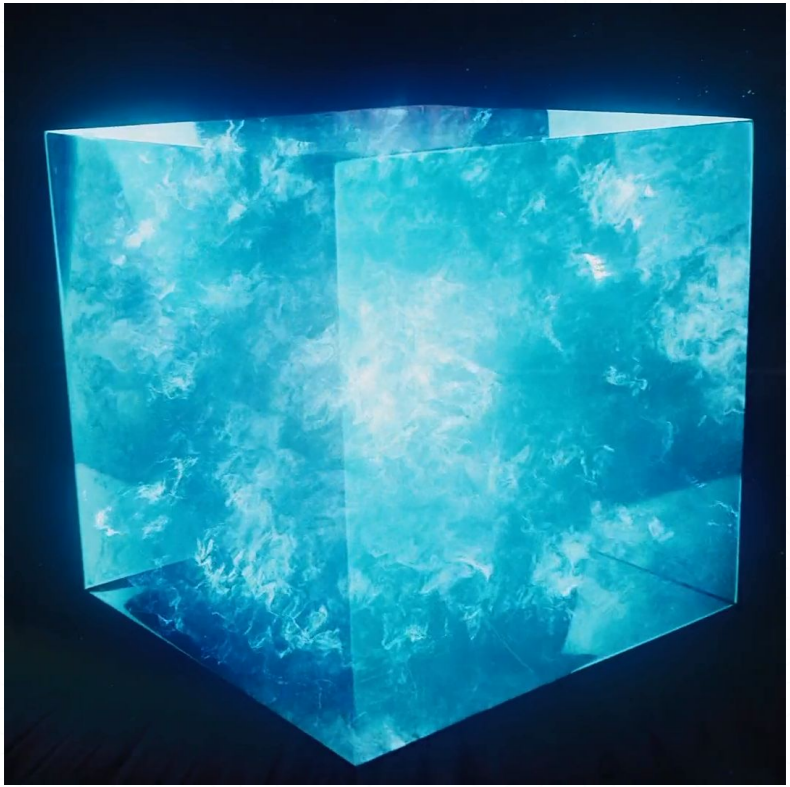
Phandeeyar

Challenges in Myanmar

- No standard for digitization
- Information not accessible.
- Library preservation not strong.
- Lack of tools to digitize objects.

Phandeeyar

- 2006



-2015 **(Burmese)**

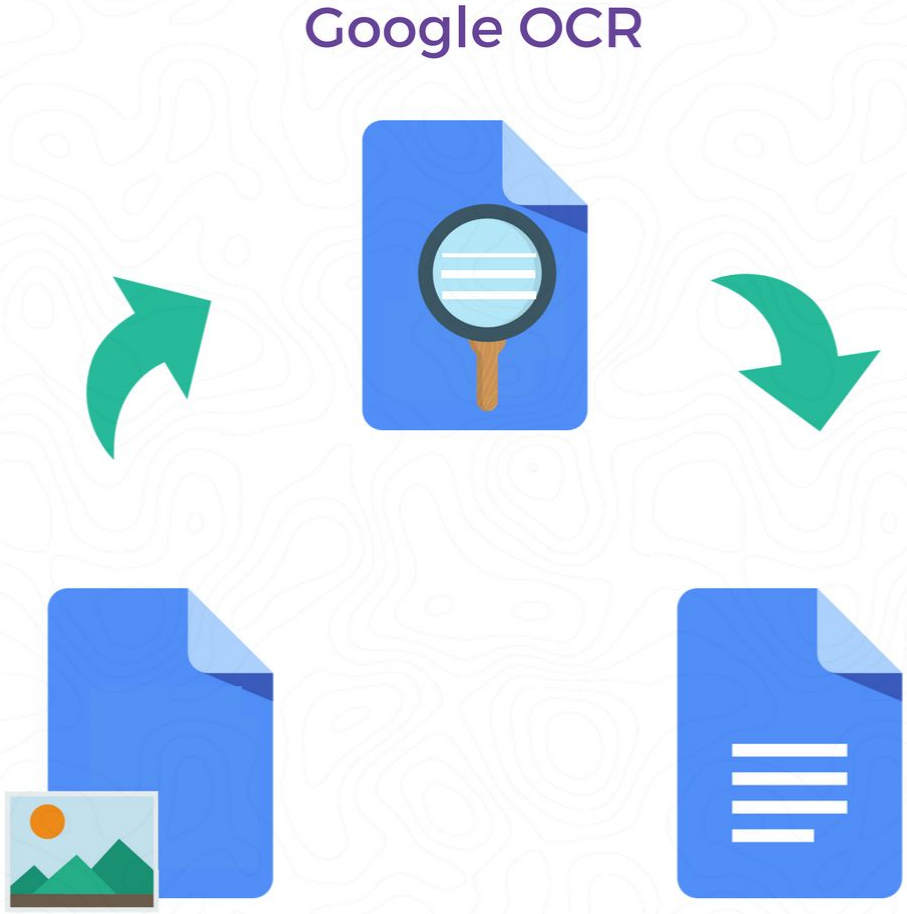
-2018 **(Google Lens)**

Tesseract OCR Engine

Phandeeyar

How does it work ?

- စာရွက်တစ်ခုခုကို ဖုန်း (သို့) စကန်နာ ဖြင့် စကန်ဖတ်ပါ။
- ပြီးရင် Google drive ပေါ်ကို တင်ပါ။
- တင်ထားတဲ့ File ကို Open as Google Doc ဖြင့်ဖွင့်ပါ။
- ထွက်လာသော စာများကို စစ်ဆေးပြီး လိုအပ်သလို Format ပြင်ပါ။



Phandeeyar

Google Lens

- Google Photos Mobile app ထဲတွင်ပါဝင်ပါသည်။
- ကိုယ် ပြောင်းချင်သော ပုံကိုဓာတ်ပုံရိုက်ပြီး Google Photos ထဲတွင်ဖွင့်ပါ။
- Len ပုံလေးကို နှိပ်ပြီး မိမိလိုချင်ရာစာကို ဆွဲထုတ်နိုင်ပါပြီ။

Phandeeyar

ရှိပြီးသား pdf များကို doc ဖိုင် ပြောင်းလဲခြင်း

- စာရိုက်ပြီး PDF အဖြစ် ပြောင်းထားသော ဖိုင်များကို Google Doc တွင် တန်းဖွင့်၍ doc ဖိုင် ပြန်ယူလို့မရပါ ။ မှားယွင်းသော အက္ခရာများသာ ထွက်လာပါမည်။

- ထိုသို့ သော အခက်အခဲကို ကျော်လွှားနိုင်သော နည်းတစ်နည်းမှာ PDF ကို JPEG ဖိုင် အဖြစ်ပြောင်းလဲ ပြီး ။ အဆိုပါ JPEG ကို PDF ပြန်လည် ပြုလုပ်ခြင်းဖြစ်သည် ။ ရရှိလာသော PDF ဖိုင်ကို Google Drive ပေါ်တင်ပြီး Open as Google docs ဆိုပြီး digitize လုပ်နိုင်သည်။ အဆိုပါ လုပ်ဆောင်ချက်ကို အောက်ပါ website များတွင် အလွယ်တကူသွားရောက်လုပ်နိုင်ပါသည်။

  PDF မှာ JPEG ပြောင်းရန် - https://pdftoimage.com/

  JPEG မှ PDF ပြောင်းရန် - https://imagetopdf.com/

Phandeeyar

Troubleshooting

1. ပြောင်းလိုက်တဲ့ စာတွေကို ဖတ်လို့မရဘူး။
2. စာတွေက အမှားအရမ်းများတယ်။
3. Table တွေကို ပြောင်းလို့မရဘူးလား
4. Pdf အဖြစ် upload ပြောင်းတင်တာမရဘူး။
5. လက်ရေးကို ကော မဖတ်ဖူးလား။

Phandeeyar

ပြောင်းလိုက်တဲ့ စာတွေကို ဖတ်လို့မရဘူး။

Google OCR က ဘာသာစကား ၁၀၀ ကျော်ကို
ကျောသားရင်သားမခွဲထောက်ပံ့ဖို့အတွက် ယူနီကုဒ် နည်းစနစ်ကို သုံးတယ်လို့
ဆိုထားပါတယ်။
ဒါကြောင့် ကျွန်တော်တို့ သုံးတဲ့ဇော်ဂျီနဲ့ဆိုရင် အမှန်မမြင် ရနိုင်ပါဘူး။
ဒါကိုဖြေရှင်းဖို့ အတွက် Zawgyi to Unicode Converter တွေကိုသုံးရပါတယ်။
လူသုံးများတဲ့ ဖောင့်ပြောင်းတဲ့ tool တွေကတော့ အောက်မှာပါ။
http://www.rabbit-converter.org/Rabbit/

Phandeeyar

စာတွေ အရမ်းများတယ်။

အရမ်းများရတဲ့အကြောင်းအရင်းကတော့
အောက်ပါအချက်တွေကြောင့်ဖြစ်နိုင်ပါတယ်၊

1. စကန်ဖတ်စဉ်မှာ မလိုအပ်တဲ့ စာတွေပါဝင်နေခြင်း
2. ဖတ်မယ့်စာရွက်ပေါ်မှာ အလင်းကောင်းစွာမရခြင်း
3. ညစ်ပတ်နေခြင်း ၊
4. ပါဝင်သော စာများမထင်ရှားခြင်း
5. Google OCR မှ မသိနိုင်သော လက်ရေးမူများ၊ ဖောင့်များကိုသုံးထားခြင်း။

Phandeeyar

Table တွေကို ကော ပြောင်းလို့မရဘူးလား

ဒါကလည်း အမေးများတဲ့မေးခွန်းတစ်ခုဖြစ်ပါတယ်။ လက်ရှိအချိန်ထိ Table ထဲကစာတွေကို Format တကျ ဆွဲထုတ်ဖို့က မဖြစ်နိုင်သေးပါဘူး။
နောက်ပိုင်းတော့ရလာမယ်လို့မျှော်လင့်ရပါတယ်။

Phandeeyar

စကန်ဖတ်ပြီး ရလာတဲ့ပုံတွေကို pdf ပြောင်းတင်တာမရဘူး။

များသောအားဖြင့် တော့ ဒီလို PDf ပြောင်းတင်လို့ရပါတယ်။ အကယ်၍ ကိုယ်က pdf ကနေ Google doc ပြောင်းလဲလို့ ဘာစာမှ မပေါ်ဘူးဆိုရင်တော့ အောက်ပါအချက်တွေကြောင့် ဖြစ်နိုင်ပါတယ်။

1. Pdf ထဲတွင်ပါဝင်သော image size ကြီးလွန်းခြင်း
2. ဘာကြောင့်မှန်းမသိခြင်း :D

ဒီလိုဖြစ်ခဲ့ရင်တော့ PDF အစား JPEG ဖိုင်တွေ ပြုလုပ်ပြီးတင်နိုင်ပါတယ်။

Phandeeyar

လက်ရေးကိုကော မဖတ်ဘူးလား

လက်ရေးကို ဖတ်နိုင်တဲ့အဆင့်ထိ Google OCR က မရောက်သေးပါဘူး.
အခြားဘာသာစကားတွေ အတွက်တော့ ရကောင်းရပါလိမ့်မယ်။

လက်ရှိမှာ မြန်မာဘာသာအတွက်ကတော့ မြန်မာဉတို့၊ ဇော်ဂျီတို့လို
စာလုံးဝိုင်းဝိုင်းလေးတွေကိုပဲ Standard စာလုံးအဖြစ် သိနိုင်ပါတယ်။
ဒါကြောင့်ဖောင့်အလှတွေ လက်ရေးအတွန့်တွေအတွက်တော့ သုံးလို့မရနိုင်သေးပါဘူး။

Phandeeyar

Great Tips

- Camscanner Application (သို့မဟုတ်) Microsoft Office Lens app ကိုအသုံးပြုပါ။
- ဖုန်းဖြင့် ဓာတ်ပုံရိုက်ပြီး scan ဖတ်စဉ် လက်ကို ငြိမ်အောင်ထားရှိပါ။ မလိုအပ်သော စာများ၊ မပါအောင် auto crop လုပ်ပေးသည့် Function ကိုအသုံးပြုပါ။
- တစ်ခုထက်ပိုသော ပုံရိပ်များကို PDF သို့မဟုတ် JPEG အဖြစ် Google drive ပေါ်သို့ တင်နိုင်ပါသည်။ JPEG ဖြင့်တင်သော ပုံများကို OCR အဖြစ်ပြောင်းပါက ပိုကောင်းပါသည်။ ပုံနှင့်စာ တွဲပြသော ကြောင့် ဖြစ်သည်။
- OCR ဖတ်၍ထွက်လာသော Google Doc များအားလုံးကို တစ်ခုစီစစ်မည့်အစား Doc တစ်ခုထဲတွင် ကော်ပီကူးထည့်၍ စစ်ခြင်းက အလုပ်ပိုတွင်ကျယ်စေကြောင်းတွေ့ရှိရသည်။

Phandeeyar

Shop

Games

Family

My account

My Play activity

My wishlist

Redeem

Parent Guide

My Play activity

My wishlist

Redeem

CamScanner -Phone PDF Creator

Top Developer

INTSIG Information Co.,Ltd **Productivity**

★★★★½ 728,010

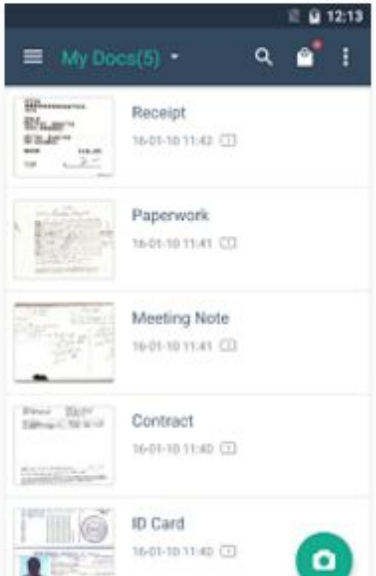
3+

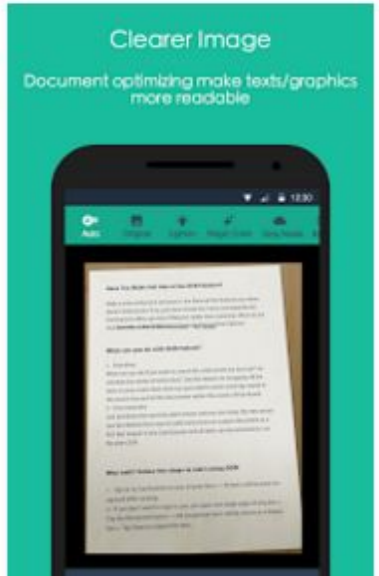
Offers in-app purchases

✕ This app is incompatible with all of your devices.

Add to Wishlist

Install





Microsoft Office Lens - PDF Scanner

Microsoft Corporation Productivity

★★★★½ 354,146

3+

This app is compatible with some of your devices.

Add to Wishlist Install

Easy Neat Powerful Orga

Google Photos

Google LLC Photography

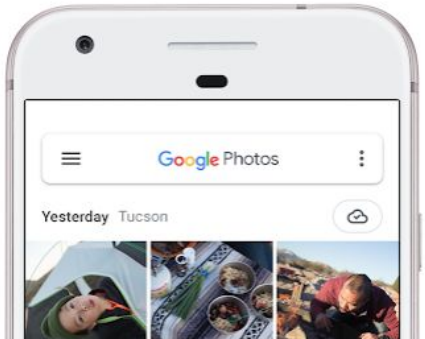
16,391,640

3+

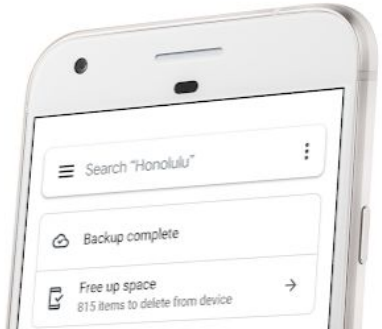
This app is compatible with all of your devices.

Installed

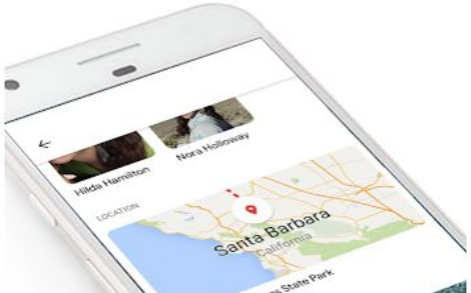
Back up your photos & access them anywhere



Never run out of storage again



Smarter albums for all your adventures



Find your pl - no taggir

Google Drive

Editors' Choice

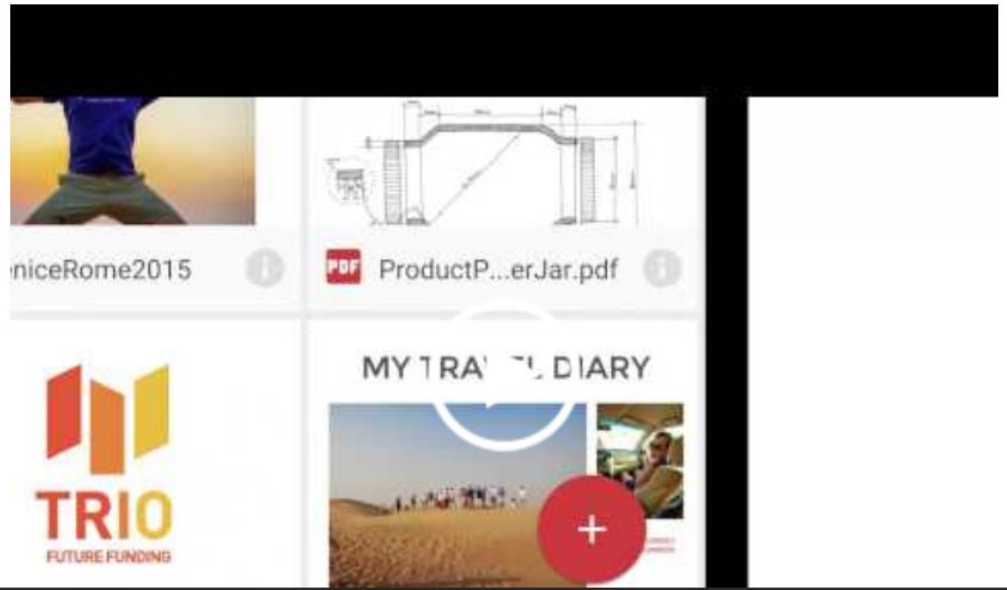
Google LLC Productivity

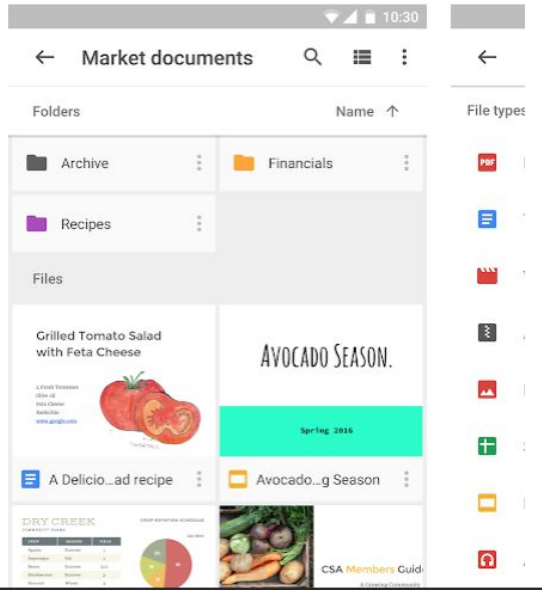
3,717,636

3+

This app is compatible with all of your devices.

Installed






LET'S GET STARTED!

၁၅ မိနစ်

1. ပေးထားသော Wifi network နှင့်ချိတ်ဆက်ပါ။(ကွန်ပျူတာကော ၊ ဖုန်းကော)
2. Google Drive app နှင့် Office Lens or Cam Scanner App ကို ဖုန်းတွင် ဒေါင်းလုဒ်ဆွဲပါ ။
3. Google Drive app ကို မိမိ Google acc ဖြင့် login ဝင်ပါ။
4. ပေးထားသော စာရွက်ကို Office lens app ဖြင့် Scan ဖတ်ပါ။
5. Jpeg ဖိုင်အဖြစ်သိမ်းပါ။
6. ပြီးနောက် Gallery ထဲကိုဝင်ပြီး Office folder ထဲကနေ စောစော က scan ဖတ်ခဲ့သော ပုံကို သွားကြည့်ပါ။
7. Google Drive ထဲကို Share ပြီးပို့လိုက်ပါ။
8. ကွန်ပျူတာကနေ Google Drive ထဲကို ဝင်ပါ။
9. စောစော က ဖိုင် ရောက်မရောက်ကြည့်ပါ
10. ရောက်လျှင် open as Google doc ဖြင့်ဖွင့်ပါ။

Phandeeyar

၁၅ မိနစ်

11. ရလာသော စာကို ဖတ်မရပါက rabbit convert website တွင် ဇော်ဂျီဖောင့်ပြောင်းပါ။
12. ရလာသော စာကို ပြန် copy ကူး၍ မူရင်း Google doc ထဲpaste လုပ်ပါ။
13. မိမိလိုချင်သော ဖောမက်အတိုင်း ပြန်ပြင်ယူပါ။
14. ပြီးသွားလျှင် nyeinchankk@phandeeyar.org သို့ ရှယ်ပေးပါ။

Phandeeyar

Links to learn

ORC apps you don't know

How to use Google OCR (Burmese)

Using Google OCR in Spreadsheet

THIS PRESENTATION IS FREELY AVAILABLE AT

http://bit.ly/GoogleOCRMM

Phandeeyar


ပျော်ရွှင်ဖွယ်
နှစ်သစ် ဖြစ်ပါစေ