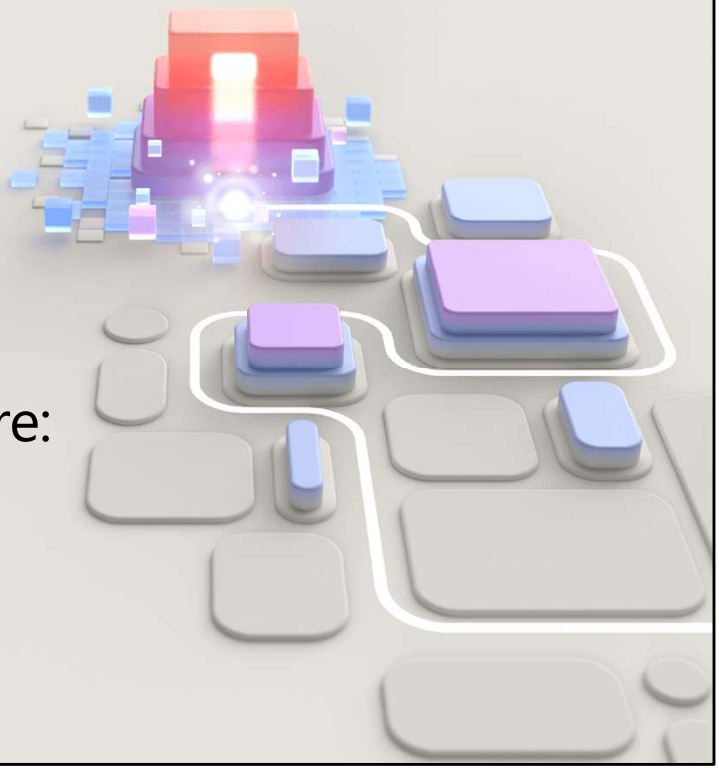
Microsoft

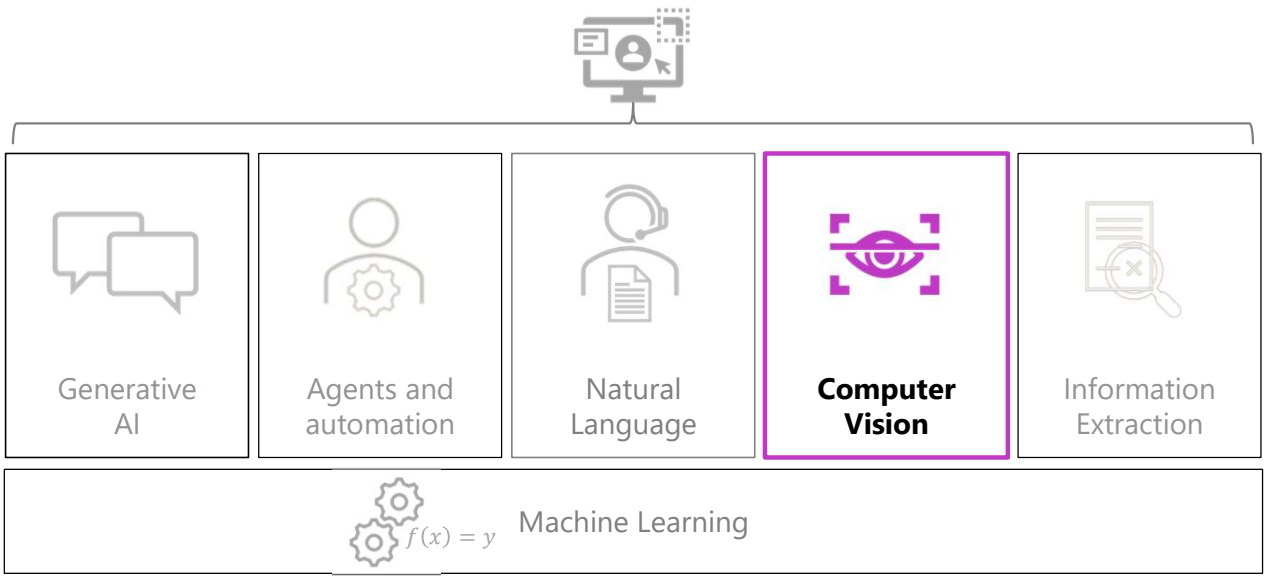


Introduction to AI in Azure:
Computer Vision

References:

- https://learn.microsoft.com/training/modules/introduction-computer-vision/
- https://learn.microsoft.com/training/modules/get-started-computer-vision-azure/

Our focus



In this section we will focus on computer vision

Agenda

- Introduction to computer vision concepts
- Get started with computer vision in Microsoft Foundry

Time estimates:
- Introduction to AI-powered information extraction concepts: 30 minutes (including lab exercise)
- Get started with AI-powered information extraction in Microsoft Foundry: 30 minutes (including demo)

Introduction to computer vision concepts

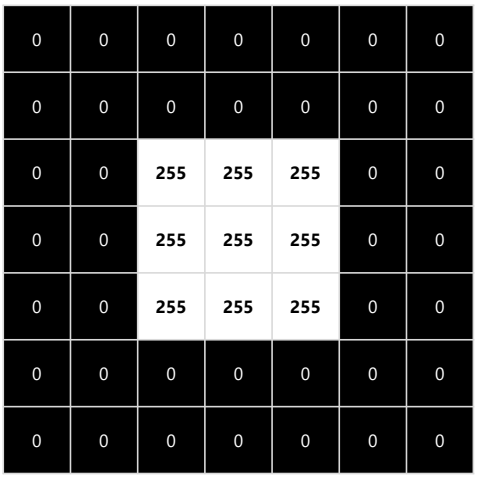
https://aka.ms/mslearn-vision

Use the link on the slide to see the Microsoft Learn learning module from which this section is derived.

Images and image processing

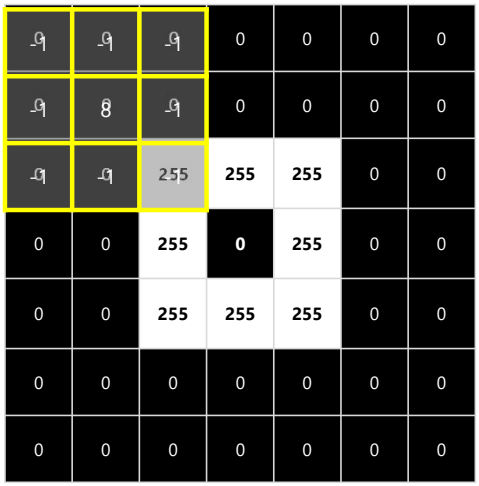
An image is an array of pixel values

Single-channel (monochrome) or multi-channel (color)



Filters are applied based on a *kernel*

Creating structure and extracting features



Denna bild med animation demonstrerar ett grundläggande begrepp inom bildbehandling: hur ett gråskalefoto representeras som en matris av pixelvärden, där 0 motsvarar svart och 255 vitt. Vänstra delen av bilden visar originalbilden – en enkel kvadratform skapad med höga värden (255) omgivet av nollor.

På högra sidan tillämpas ett så kallat filter, i detta fall en **Laplace-filter**.

Filtret representeras av en **kernel**, här en 3x3-matris med vikter. Vid filtrering konvolveras ("sveps") denna kernel över hela bilden. För varje position multipliceras motsvarande pixelvärden med kernelns vikter, och resultaten summeras – detta skapar ett nytt pixelvärde i en ny bildmatris.

Syftet med just Laplace-filtret är att markera kanter i bilden genom att upptäcka plötsliga förändringar i ljusstyrka. Resultatet blir att konturer framträder tydligt – medan homogena ytor (t.ex. mitten av den vita kvadraten) dämpas eller försvinner.

Bilden ger en konkret förståelse för hur filter används i allt från fotoförbättring till maskininlärning och datorseende.

En gråskale-bild är bara en matris av pixelvärden mellan 0 (svart) och 255 (vitt). Genom att använda ett filter – en liten matris med vikter – kan vi hitta mönster, som kanter eller former, i bilden.

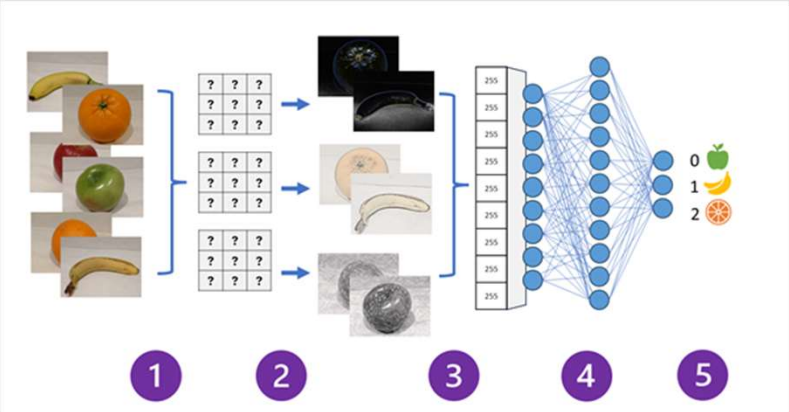
Exemplet här använder ett **Laplace-filter** som hittar **kanter**, alltså där bilden förändras snabbt. Sådana filter används i t.ex. **ansiktsigenkänning**, **självkörande bilar** (för att se vägkanter) eller **medicinska bilder** (för att hitta tumörer eller strukturer). Det är grunden för mycket datorseende och AI.

This is an animated slide. Practice the animation to make the following points.

- A grayscale image consists of an array of pixel values between 0 (black) and 255 (white) (a color picture consists of three arrays called *channels*, each containing pixel values for red, green, and blue hues)
- Most image processing involves the application of filters, which consist of *kernels* that define an array of weight values.
- The filter kernel is *convolved* (passed across and down) over the image, calculating values for a new array by using a *weighted sum* of the original pixel values multiplied by the corresponding kernel weights.
- The new array represents the filtered image, usually with some kind of visual effect. In this case, a filter names a *Laplace* filter has been used to highlight the edges of shapes in the original image.

Convolutional Neural Networks

1. Labeled mages are used to train the model
2. Filter layers extract *feature maps* from each image
3. The feature maps are flattened
4. The feature values are fed into a fully connected neural network
5. The output layer produces a probability value for each possible class label



- During training, the filter kernels start with random weights. These weights are iteratively adjusted to improve the accuracy of the predictions based on the known labels.
- The trained model uses learned weights to extract features from new images and predict their class.

Bilden förklarar hur ett Convolutional Neural Network (CNN) fungerar – en typ av artificiellt neuralt nätverk som används för att känna igen bilder, till exempel om en bild visar ett äpple, en banan eller en apelsin.

Här är processen steg för steg:

Inlärning med märkta bilder:
För att nätverket ska kunna lära sig måste det först tränas med många bilder där vi redan vet vad de föreställer, till exempel "det här är en banan".

Filtrering – hitta mönster (feature maps):
Filtreringslager använder så kallade filter (små matrisformade mallar) för att söka efter mönster i bilderna, som kanter, kurvor eller färgförändringar.

Platta ut mönstren:
De mönster eller "feature maps" som hittas omvandlas till en lång lista med siffror, så att informationen kan matas vidare in i nästa steg.

Mata in i ett neuralt nätverk:
De utplattade siffrorna skickas in i ett vanligt neuralt nätverk – alltså ett system av "noder" eller konstgjorda neuroner som bearbetar informationen.

Resultat – vad visar bilden?:
Nätverket räknar ut sannolikheter för varje klass (t.ex. 70 % chans att det är en apelsin). Den klass med högst sannolikhet väljs som svaret.

Träningen:
I början gissar nätverket ganska dåligt. Men genom att jämföra med de rätta svaren och justera vikterna (värdena i filtren) lär det sig vad som kännetecknar till exempel ett äpple. Detta upprepas tusentals gånger tills träffsäkerheten blir bra.

Efter träning:
När modellen är färdigtränad kan den känna igen nya bilder och säga vad det troligen är – även om den aldrig sett just den bilden förut.

Detta är grunden till tekniker som ansiktsigenkänning, självkörande bilar och bildsökning i telefoner.

One of the most common machine learning model architectures for computer vision is a *convolutional neural network* (CNN). CNNs use filters to extract numeric feature maps from images, and then feed the feature values into a deep learning model to generate a label prediction. For example, in an *image classification* scenario, the label represents the main subject of the image (in other words, what is this an image of?). You might train a CNN model with images of different kinds of fruit (such as apple, banana, and orange) so that the label that is predicted is the type of fruit in a given image.

During the *training* process for a CNN, filter kernels are initially defined using randomly generated weight values. Then, as the training process progresses, the models predictions are evaluated against known label values, and the filter weights are adjusted to improve accuracy. Eventually, the trained fruit image classification model uses the filter weights that best extract features that help identify different kinds of fruit.

The following diagram illustrates how a CNN for an image classification model works:

1.Images with known labels (for example, 0: apple, 1: banana, or 2: orange) are fed into the network to train the model.

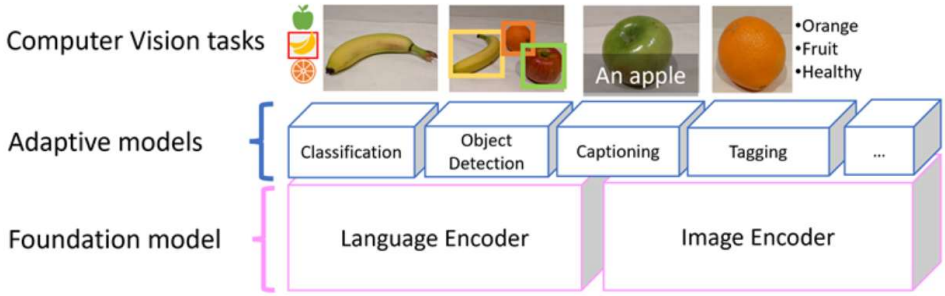
2.One or more layers of filters is used to extract features from each image as it is fed through the network. The filter kernels start with randomly assigned weights and generate arrays of numeric values called *feature maps*.

3.The feature maps are flattened into a single dimensional array of feature values.

4.The feature values are fed into a fully connected neural network.

5.The output layer of the neural network uses a *softmax* or similar function to produce a result that contains a probability value for each possible class, for example [0.2, 0.5, 0.3].

Multi-modal models



- A newer approach to modeling involves combining language and vision models that encode image and text data
- The model encapsulates semantic relationships between features extracted from the images and text extracted from related captions.
- A multi-modal model can be used as a *foundation* model for more specialized *adaptive* models.

Den här bilden förklarar hur **multi-modala modeller** fungerar – alltså AI-modeller som kombinerar både **bild** och **text**.

Traditionellt har modeller antingen tolkat text (språkmodeller) eller bilder (bildmodeller). En multi-modal modell kombinerar båda och kan på så vis förstå **sambandet mellan vad man ser och vad man säger** om det.

I grunden finns en så kallad **foundation model** som består av två delar:

En **image encoder** som tolkar bilden och gör om den till siffror (funktioner), och

En **language encoder** som gör samma sak med text (som bildtexter eller etiketter).

Dessa funktioner samlas i en gemensam representationsyta där modellen kan lära sig att t.ex. "den här formen + färgen = apelsin".
Ovanpå denna grundmodell bygger man specialiserade **adaptive models** för olika syften:

Klassificering (vilken frukt är det?)
Objektdetektion (var finns frukten i bilden?)
Bildtextgenerering (skriv: "en grön äpple")
Taggning (etiketter som "hälsosam", "frukt")

Fördelen med detta tillvägagångssätt är att det blir möjligt att lösa flera uppgifter med **en och samma modell**, som dessutom kan generalisera till nya uppgifter snabbare. Det används bland annat i AI-tjänster som kan förstå både bilder och språk samtidigt – som t.ex. GPT-4 med bildfunktioner.

Modern vision models are trained with huge volumes of captioned images from the internet and include both a language encoder and an image encoder. Often users will interact with and adapted *foundation* models. Foundation models are pre-trained general models on which you can build multiple *adaptive* models for specialist tasks. For example, you can adapt a foundation model to perform:

Image classification: Identifying to which category an image belongs.

Object detection: Locating individual objects within an image.

Captioning: Generating appropriate descriptions of images.

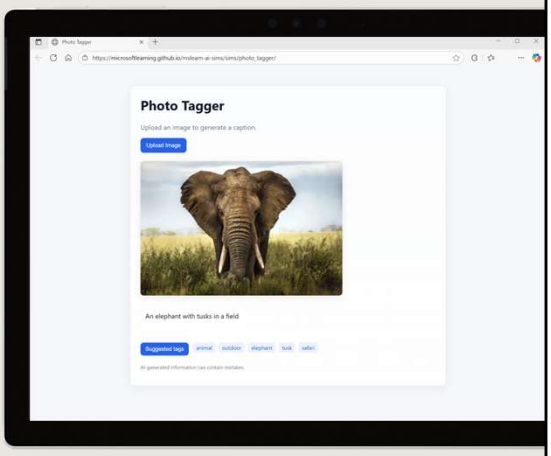
Tagging: Compiling a list of relevant text tags for an image.

Exercise

Explore computer vision scenarios

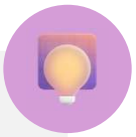
In this exercise, you'll explore computer vision capabilities in an application.

Start the exercise at:
https://go.microsoft.com/fwlink/?linkid=2334316



Note: The application used in this exercise is a *simulation* - there's no actual AI computer vision service behind it. However, it's based on real capabilities you can implement with Microsoft Foundry; and in particular, Azure Vision .

Knowledge check



1 **Computer vision is based on the manipulation and analysis of what kinds of values in an image?**

- ☐ Timestamps in photograph metadata
- ☑ Pixels
- ☐ Image file names

2 **What is the primary role of filters in a convolutional neural network (CNN) used for image classification?**

- ☐ To apply visual effects to enhance image appearance
- ☑ To extract numeric features from images for use in a neural network
- ☐ To compress image size for faster processing

3 **What is the primary function of a multi-modal model in computer vision?**

- ☐ To generate random captions for unlabeled images
- ☐ To replace CNNs entirely in all vision tasks
- ☑ To combine image features with natural language embeddings for richer understanding

Här är korta svar med max 10 ord per motivering:

1. Vad analyserar datorseende i bilder?

❌ **Fel:** Metadata används inte för visuell analys.
✅ **Rätt:** Bilder representeras som numeriska pixelvärden.
❌ **Fel:** Filnamn innehåller ingen visuell information.

2. Filtrens roll i CNN för bildklassificering

❌ **Fel:** Filter skapar inte visuella effekter.
✅ **Rätt:** Identifierar mönster som kan läras av nätverket.
❌ **Fel:** Komprimering är inte filtrens huvudsyfte.

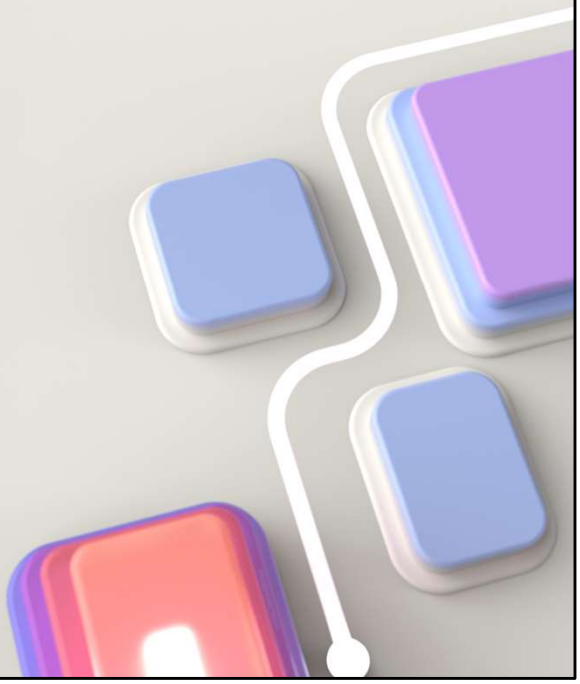
3. Funktion hos multimodala modeller i datorseende

❌ **Fel:** Skapar inte slumpmässiga bildtexter.
❌ **Fel:** Ersätter inte CNN:er i alla tillämpningar.
✅ **Rätt:** Kombinerar bild och språk för bättre förståelse.

Allow students a few minutes to think about the questions, then reveal the correct answers.

Get started with computer vision in Microsoft Foundry

https://aka.ms/mslearn-azure-vision



Use the link on the slide to see the Microsoft Learn learning module from which this section is derived.

Computer vision in Microsoft Foundry

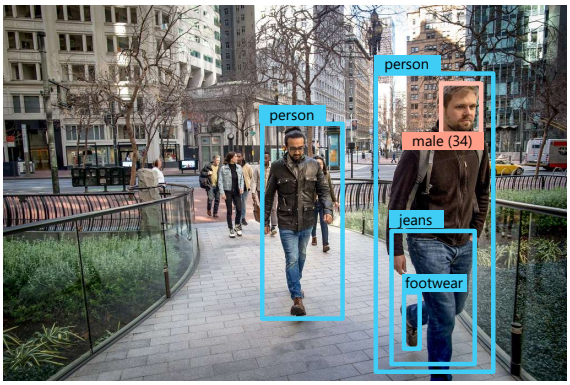
 **Azure Vision in Foundry Tools**	 **Face API**
• Image Analysis: • Image tagging, captions, model customization, and more. • Optical Character Recognition (OCR) • Video analysis	• Face detection • Face recognition

We'll cover some of the key capabilities of these computer vision related services in the rest of this module.
For more details of the capabilities of these Foundry tools, see the documentation at:
- https://learn.microsoft.com/azure/ai-services/computer-vision/
- https://learn.microsoft.com/azure/ai-services/computer-vision/overview-identity

Image analysis with *Azure Vision*

Capabilities include:

- Model customization
- Read text from images
- Detect people in images
- Generate image captions
- Detect objects
- Tag visual features
- Smart crop

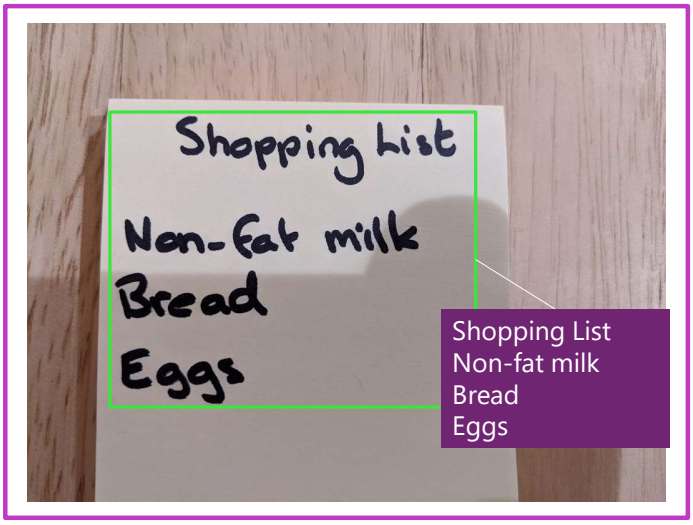


Caption: A group of people walking on a sidewalk
Tags: Building, jeans, street, outdoor, jacket, city, person

The Azure Vision Image Analysis service can extract a wide variety of visual features from your images. For example, it can determine whether an image contains adult content, find specific brands or objects, or find human faces.

Reading text with Optical Character Recognition (OCR)

- Detect the location and characters of **printed** and **handwritten** text
- Options for quick text extraction from images, or asynchronous analysis of larger scanned documents



We've already discussed the Computer Vision service and its use in image analysis. It also provides OCR capabilities to detect and extract text from images.

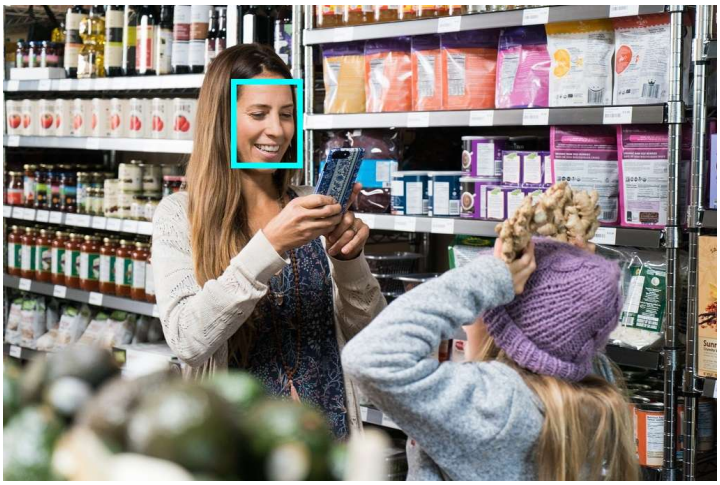
Detecting faces with *Face API*

Everyone can use the Face service to detect:

- Blur
- Exposure
- Glasses
- Head pose
- Noise
- Occlusion

Only Managed Microsoft customers can access facial recognition capabilities:

- Similarity matching
- Identity verification



Azure Face Service används för att känna igen och analysera ansikten i bilder. Alla användare kan upptäcka tekniska faktorer som suddighet (blur), ljussättning (exposure), glasögon, huvudposition, brus (noise) och skymd sikt (occlusion). Dessa hjälper till att avgöra bildkvalitet och hur lätt ett ansikte är att analysera.

Mer avancerade funktioner, som likhetsjämförelse (t.ex. är det samma person i två bilder?) och identitetsverifiering (vem är personen på bilden), är endast tillgängliga för godkända kunder hos Microsoft. Tjänsten används i allt från fotohantering till säkerhetslösningar, men med tydliga regler för ansiktsigenkänning.

While Azure Vision provides some basic face detection and analysis features, the Face service offers additional capabilities. All users can use the Face service to detect:

- **Blur**: how blurred the face is (which can be an indication of how likely the face is to be the main focus of the image)
- **Exposure**: aspects such as underexposed or over exposed and applies to the face in the image and not the overall image exposure
- **Glasses**: if the person is wearing glasses
- **Head pose**: the face's orientation in a 3D space
- **Noise**: refers to visual noise in the image. If you have taken a photo with a high ISO setting for darker settings, you would notice this noise in the image. The image looks grainy or full of tiny dots that make the image less clear
- **Occlusion**: determines if there may be objects blocking the face in the image

Responsible AI use

To support Microsoft's [Responsible AI Standard](https://blogs.microsoft.com/on-the-issues/2022/06/21/microsofts-framework-for-building-ai-systems-responsibly/), a new [Limited Access policy](https://aka.ms/AAh91ff) has been implemented for the Face

service and Computer Vision service.

Anyone can use the Face service to:

* Detect the location of faces in an image
* Determine if a face is wearing glasses
* Determine if there's occlusion, blur, noise, or over/under exposure for any of the faces
* Return the head pose coordinates for each face in an image

The Limited Access policy requires customers to [submit an intake form](https://aka.ms/facerecognition) to access additional Face service capabilities including:

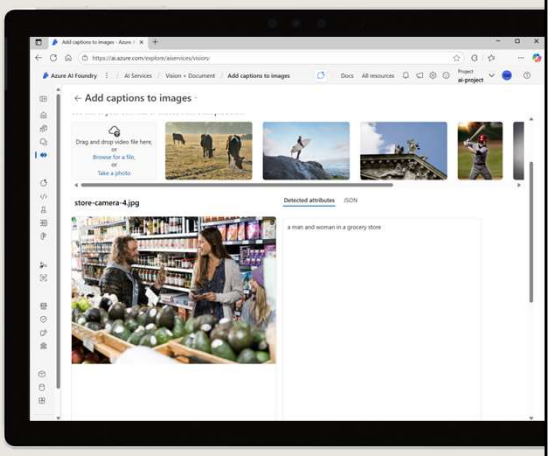
* The ability to compare faces for similarity
* The ability to identify named individuals in an image

Exercise – If time permits

Analyze images in Microsoft Foundry

In this exercise, you'll use Microsoft Foundry to explore Azure Vision capabilities.

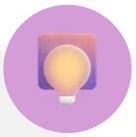
Start the exercise at:
https://go.microsoft.com/fwlink/?linkid=2250145



You require an Azure subscription to perform this exercise – which may be provided by an Authorized Lab Hoster.

The exercise can take a substantial amount of time. Our recommendation in one-day deliveries is for the instructor to demonstrate the core tasks in this exercise; completing the exercise ahead of time (https://go.microsoft.com/fwlink/?linkid=2250145) so you have a project already created.

Knowledge check



1 **You want to use the Face detection service to identify faces in images. What can be identified using the Face detection service?**

- ☐ Faces that cannot be seen because the person has turned their back.
- ☑ Partially obscured faces.
- ☐ Faces that are obscured by another object.

2 **You want to use Azure Vision and Azure Language. You also want developers to require only one key and endpoint to access all your services. What kind of resource should you create in Azure?**

- ☑ Foundry
- ☐ Language
- ☐ Vision

3 **Which capabilities are part of Azure Vision?**

- ☐ Face detection and speech recognition
- ☑ Optical Character Recognition and face detection
- ☐ Document Intelligence and speech recognition

Här är korta svar med max 10 ord per motivering:

1. Vad kan identifieras med Face detection?

❌ **Fel:** Ansikten som inte syns kan inte identifieras.
✅ **Rätt:** Tjänsten klarar delvis synliga ansikten.
❌ **Fel:** Helt blockerade ansikten kan inte detekteras.

2. En nyckel och endpoint för Vision och Language

✅ **Rätt:** Foundry ger gemensam åtkomst till flera AI-tjänster.
❌ **Fel:** Language täcker inte Vision-tjänster.
❌ **Fel:** Vision täcker inte Language-tjänster.

3. Funktioner som ingår i Azure Vision

❌ **Fel:** Taligenkänning tillhör Azure Speech.
✅ **Rätt:** OCR och ansiktsdetektion är Vision-funktioner.
❌ **Fel:** Speech ingår inte i Azure Vision.

Allow students a few minutes to think about the questions, then reveal the correct answers.

Summary

Computer vision concepts

- Computer vision is based on analysis and manipulation of pixels
- Convolutional neural networks (CNNs) apply deep learning model architectures to images
- Multi-modal models combine image processing models with language models

Get started with computer vision in Microsoft Foundry

- Azure Vision enables image analysis
- Use the Face API to detect faces in images
- Read text with optical character recognition (OCR)