Features

The sections below highlight the features and capabilities of the Google Video Intelligence API.

Supported video formats

The Video Intelligence API supports common video formats, including `.MOV`, `.MPEG4`, `.MP4`, `.AVI`, and the formats decodable by ffmpeg (https://ffmpeg.org/).

Label detection

Label detection annotates a video with labels (tags) for entities that are detected in a video or video segments and returns the following:

- A list of video segment annotations where an entity is detected.
- A list of frame annotations where an entity is detected.
- If specified in the request, a list of shots where an entity is detected. For details, see Shot change detection (#shot-change).

For example, for a video of a train at a crossing, the Video Intelligence returns labels such as "train", "transportation", "railroad crossing", and so on. Each label includes a time segment with the time offset (timestamp) for the entity's appearance from the beginning of the video. Each annotation also contains additional information including an entity id that you can use to find more information about the entity in the Google Knowledge Graph Search API (https://developers.google.com/knowledge-graph/).

Each entity returned can also include associated category entities in the `categoryEntities` field. For example the “Terrier” entity label has a category of “Dog”. Category entities have a hierarchy. For example, the “Dog” category is a child of the “Mammal” category in the hierarchy. For a list of the common category entities that the Video Intelligence uses, see entry-level-categories.json (https://cloud.google.com/video-intelligence/docs/entry-level-categories.json).

To detect labels in a video, call the `annotate` (https://cloud.google.com/video-intelligence/docs/reference/rest/v1/videos/annotate) method and specify `LABEL_DETECTION` (https://cloud.google.com/video-intelligence/docs/reference/rest/v1/videos#Feature) in the `features` field.

For examples, see Analyzing Videos for Labels (https://cloud.google.com/video-intelligence/docs/analyze-labels) and Label Detection Tutorial (https://cloud.google.com/video-intelligence/docs/label-tutorial).

Shot change detection

By default the Video Intelligence examines a video or video segments by frame. That is, each complete picture in the series that forms the video. You can also have the Video Intelligence annotate a video or video segment according to each shot (scene) that it detects in the input video.

Shot change detection annotates a video with video segments that are selected based on content transition (scenes) as opposed to the individual frames. For example, a golf video following two players across the golf course with some panning to the woods for background may produce two shots: "players" and "woods," giving the developer access to the most relevant video segments showing the players for highlights.

To detect shot changes in a video, call the `annotate` (https://cloud.google.com/video-intelligence/docs/reference/rest/v1/videos/annotate) method and specify `SHOT_CHANGE_DETECTION` (https://cloud.google.com/video-intelligence/docs/reference/rest/v1/videos#Feature) in the `features` field.

For examples, see Analyzing Videos for Shot Changes (https://cloud.google.com/video-intelligence/docs/analyze-shots) and Shot Change Detection Tutorial (https://cloud.google.com/video-intelligence/docs/shot_detection).

Explicit content detection

Explicit Content Detection detects adult content within a video. Adult content is content generally appropriate for 18 years of age and older, including but not limited to nudity, sexual activities, and pornography (including cartoons or anime).

Explicit content detection annotates a video with explicit content annotations (tags) for entities that are detected in the video or video segments provided. The response returns a video frame timestamp where the explicit content is detected.

To detect explicit content in a video, call the `annotate` (https://cloud.google.com/video-intelligence/docs/reference/rest/v1/videos/annotate) method and specify `EXPLICIT_CONTENT_DETECTION` (https://cloud.google.com/video-intelligence/docs/reference/rest/v1/videos#Feature) in the `features` field.

For an example, see Analyzing Videos for Explicit Content (https://cloud.google.com/video-intelligence/docs/analyze-safesearch).

Regionalization

You can use the `location_id` parameter in your AnnotateVideoRequest (https://cloud.google.com/video-intelligence/docs/reference/rest/v1/videos#resource-annotaterequest), to specify the Google Cloud Platform region (https://cloud.google.com/about/locations/) where annotation is performed. The following regions are currently supported:

- us-east1
- us-west1
- europe-west1
- asia-east1

If no region is specified, the region is determined based on the video file location.

Speech Transcription

Speech Transcription transcribes spoken word audio in a video or video segment into text and returns blocks of text for each portion of transcribed audio.

To transcribe speech from a video, call the `annotate` (https://cloud.google.com/video-intelligence/docs/reference/rest/v1/videos/annotate) method and specify `SPEECH_TRANSCRIPTION` (https://cloud.google.com/video-intelligence/docs/reference/rest/v1/videos#Feature) in the `features` field.

You can use the following features when transcribing speech:

- **Alternative words**: Use the `maxAlternatives` option to specify the maximum number of options for recognized text translations to include in the response. This value can be an integer from 1 to 30. The default is 1. The API returns multiple transcriptions in descending order based on the confidence value for the transcription. Alternative transcriptions do not include word-level entries.
- **Profanity filtering**: Use the `filterProfanity` option to filter out known profanities in transcriptions. Matched words are replaced with the leading character of the word followed by asterisks. The default is false.
- **Transcription hints**: Use the `speechContexts` option to provide common or unusual phrases in your audio. Those phrases are then used to assist the transcription service to create more accurate transcriptions. You provide a transcription hint as a SpeechContext (https://cloud.google.com/video-intelligence/docs/reference/rest/v1/videos#SpeechContext) object.
- **Audio track selection**: Use the `audioTracks` option to specify which track to transcribe from multi-track audio. Users can specify up to two tracks. Default is 0.
- **Automatic punctuation**: Use the `enableAutomaticPunctuation` option to include punctuation in the transcribed text. The default is false.
- **Multiple speakers**: Use the `enableSpeakerDiarization` option to identify different speakers in a video. In the response, each recognized word includes a `speakerTag` field that identifies which speaker the recognized word is attributed to.

For best results, provide audio recorded at 16,000Hz or greater sampling rate.

For an example, see Speech Transcription (https://cloud.google.com/video-intelligence/docs/transcription).

Object Tracking

Object tracking tracks multiple objects detected in an input video or video segments and returns labels (tags) for detected entities along with the location of the entity in the frame. For example, a video of vehicles crossing a traffic signal may produce labels such as “car” , “truck”, “bike”, “tires”, “lights”, “window” and so on. Each label includes a series of bounding boxes showing the location of the object in the frame. Each bounding box also has an associated time segment with a time offset (timestamp) that indicates the duration offset from the beginning of the video. The annotation also contains additional entity information including an entity id that you can use to find more information about the entity in the Google Knowledge Graph Search API (https://developers.google.com/knowledge-graph/).

Object tracking differs from label detection (https://cloud.google.com/video-intelligence/docs/analyze-labels) in that label detection provides labels for the entire frame (without bounding boxes), while object tracking detects individual objects and provides a label along with a bounding box that describes the location in the frame for each object.

To make an object tracking request, call the `annotate` (https://cloud.google.com/video-intelligence/docs/reference/rest/v1p2beta1/videos/annotate) method and specify `OBJECT_TRACKING` (https://cloud.google.com/video-intelligence/docs/reference/rest/v1p2beta1/videos#Feature) in the `features` field.

Note: There is a limit on the size of the detected objects. Very small objects in the video might not get detected.

For an example, see Object Tracking (https://cloud.google.com/video-intelligence/docs/object-tracking).

Text Detection

Text Detection performs Optical Character Recognition (OCR) to detect visible text from frames in a video or video segments and returns the detected text along with information about where the text was detected in the video.

Text detection is available for all of the languages (https://cloud.google.com/vision/docs/languages) supported by the Cloud Vision API.

To detect visible text from a video or video segments, call the `annotate` (https://cloud.google.com/video-intelligence/docs/reference/rest/v1p2beta1/videos/annotate) method and specify `TEXT_DETECTION` (https://cloud.google.com/video-intelligence/docs/reference/rest/v1p2beta1/videos#Feature) in the `features` field.

For an example, see Text Detection (https://cloud.google.com/video-intelligence/docs/text-detection).
