サムネイルの自動生成手法の提案

奈良女子大学　高田研究室

YouTubeとは

- 2005年に開始された動画共有サービス
- 月間アクティブユーザ数：世界2位（2022年1月時点）

YouTubeで収入を得る条件

- 動画の総再生時間：**4,000時間以上**
- チャンネル登録者数：**1,000人以上**

→長時間の動画再生・視聴者の獲得が必要

サムネイルの影響

- YouTubeアプリで表示される情報：タイトル、サムネイル、チャンネル名、再生回数、投稿時間
- 投稿者自身が設定できるもの：タイトル、サムネイル

→サムネイルの影響：大

- 魅力的なサムネイルの作成→再生数（クリック数）の増加

現在存在する手法

YouTubeによるサムネイルの自動生成

- 動画内から3枚抜粋
- テキスト配置などの機能：×

『意味的に要約された動画用サムネイルの生成』−研究報告コンピュータグラフィックスとビジュアル情報学（2024）

- 意味的に重要と判断したオブジェクトを配置
- テキスト配置などの機能：×

目的：テキストを配置しているサムネイルの自動生成

手順

①動画の内容を要約

・動画から音声を抽出

・音声をテキストに変換

②サムネイルに配置するテキスト生成

・要約文を生成

③YouTubeから画像候補を取得

④テキストを画像内に自動配置

音声抽出技術

●投稿する予定の動画から音声を抽出

→FFmpeg、Gstreamer

・LinuxやWindows、macOSなどで動作

・オープンソースのマルチメディアフレームワーク

・機能：動画から音声を抽出、複数動画の結合、動画の切り出しなど

テキスト変換技術

●音声データからテキストデータへ変換

→Whisper、Speech-to-Text API、Watson Speech to Text

- Whisper（OpenAI）：68万時間分のデータセットで学習
- Speech-to-Text API（Google）：Googleのディープラーニングニューラルネットワークアルゴリズムを利用
- Watson Speech to Text（IBM）：実際コールセンターなどで活用、最大6人まで認識可能

要約文生成技術

●テキストデータから要約文を生成

→Azure AI language、Text Summarization API

- Azure AI language（Microsoft）：言語検出など多数の機能有
- Text Summarization API（Recruit）：最大文字数200文字かつ最大文章数10の制約有、要約後の文章数を指定可

画像取得技術

既存手法

●YouTubeのアルゴリズム

・動画内から3枚自動抜粋

・テキスト挿入などの機能：×

→元画像として使用

テキスト自動配置における処理

- テキストの位置算出
- テキストの装飾処理

→文字の大きさ調整

→色調整

→縁取り装飾処理

テキスト位置算出

●元画像にテキストを自動配置

- 背景や被写体に考慮する必要有

- 先行研究：『**バナー制作のための背景を考慮した自動テキスト配置**』-人工知能学会全国大会論文集第34回（2020年）

テキスト位置算出の条件

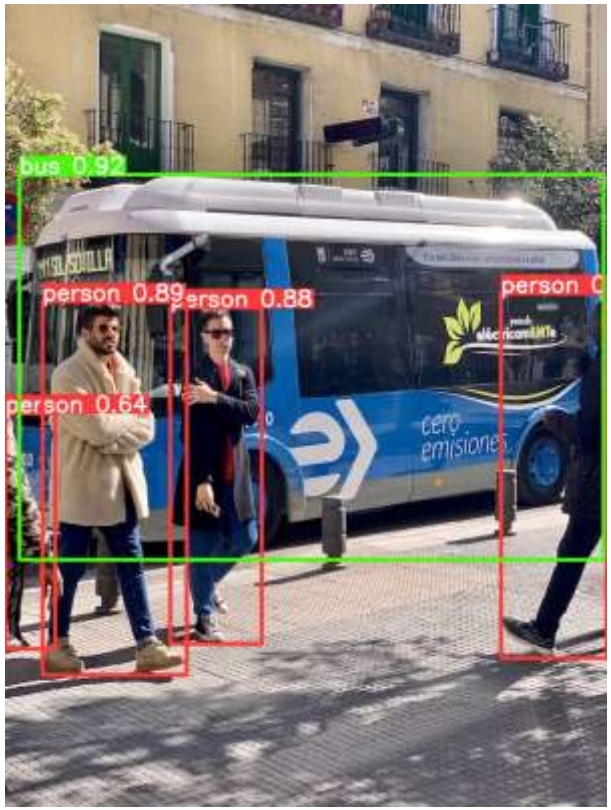
- 被写体の重なりを最小限
- 全体構図のバランスを配慮

位置算出の手順

①バウンディングボックスを取得

②物体マップの生成

③テキストの位置を算出



※バウンディングボックス：物体検出に用いられる長方形。画像内の物体の位置、クラス分類とその確率が格納

バウンディングボックスの取得

- 入力画像からN個の物体についてバウンディングボックスbnを取得
- 使用した物体検出器：YOLOv8（2023年1月公開）

YOLO

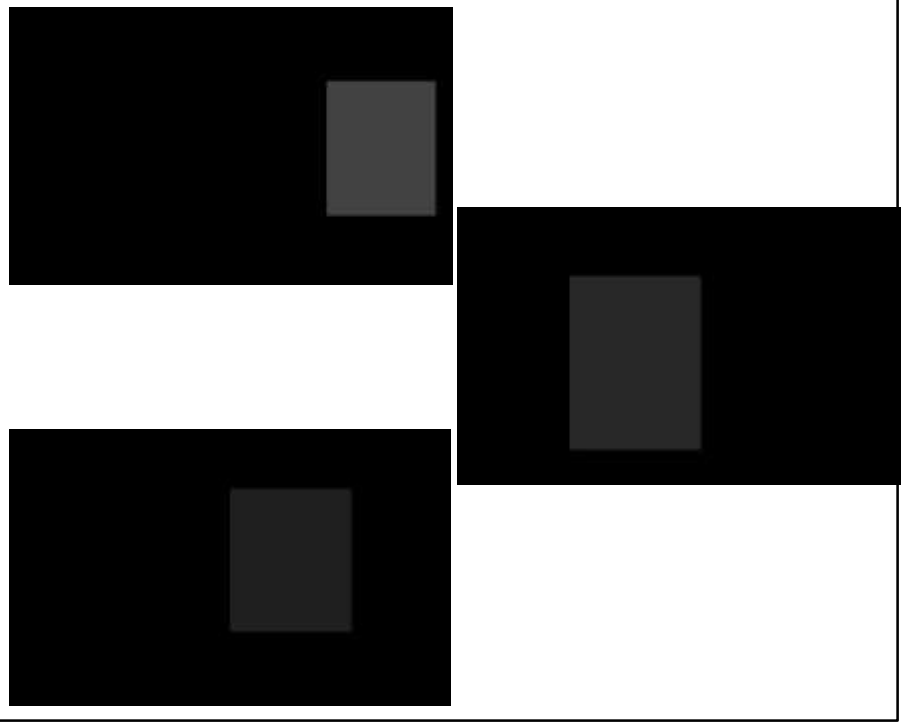
- You Only Look Onceの略称（2016年発表）
- 処理速度が速く、誤検出を大幅に抑制

座標(xy)（左上）	座標(xy)（右下）	確信度（0~1）
(1146,268)	(1538,753)	0.66
(406,250)	(879,876)	0.40
(800,217)	(1239,735)	0.31

物体マップの生成

- 入力画像と同じ大きさのマップBnを生成

座標(xy)（左上）	座標(xy)（右下）	確信度（0~1）
(1146,268)	(1538,753)	0.66
(406,250)	(879,876)	0.40
(800,217)	(1239,735)	0.31



物体マップの生成

- N個のBnを合成
- 各画素：バウンディングボックスbnの確信度（物体が検出された領域＞検出されなかった背景領域）

物体マップO

テキスト位置決定

- 目標：注目すべき被写体を避けた背景領域にテキスト配置
- 被写体の中心＝物体マップOの重心座標C

- $C_x = \frac{\sum_x \sum_y x\, O(x,y)}{\sum_x \sum_y O(x,y)}$
- $C_y = \frac{\sum_x \sum_y y\, O(x,y)}{\sum_x \sum_y O(x,y)}$

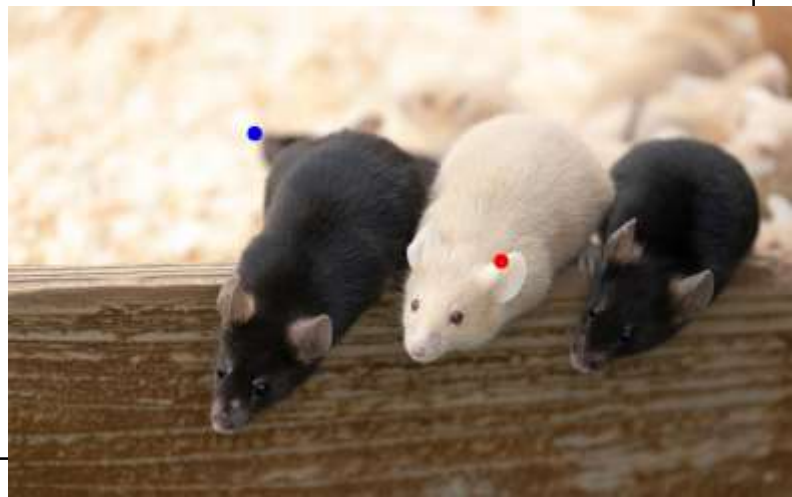
テキスト位置決定

- 重心座標Cに対して一番離れた四隅を直線で結び、中点にテキストを配置

→構図のバランスのため余白を考慮

<u>画像左上を原点とした場合</u>

- $P_x = \begin{cases} I_{width} - (C_x/2), (I_{width} > 2C_x) \\ C_x/2, (otherwise) \end{cases}$
- $P_y = \begin{cases} I_{height} - (C_y/2), (I_{height} > 2C_y) \\ C_y/2, (otherwise) \end{cases}$



文字の大きさ

- テキストサイズが小さい：視認性の低下

→大きくする必要有

- 上下左右の余白を配慮

→余白を画像の5%分確保

改行処理

BudouX（Google）

- オープンソースのライブラリ
- 次の文字で区切るか否かの二値分類問題

文字色調整

手順

① 背景色を取得
② 補色を計算
③ コントラスト比を計算
④ 明暗調整
⑤ 縁取り装飾

※補色：色相環で反対に位置する色の組み合わせ

※コントラスト比：最も明るい白と最も暗い黒の明るさの比率

背景色

・計算方法：各画素におけるRGB値の平均値

元画像

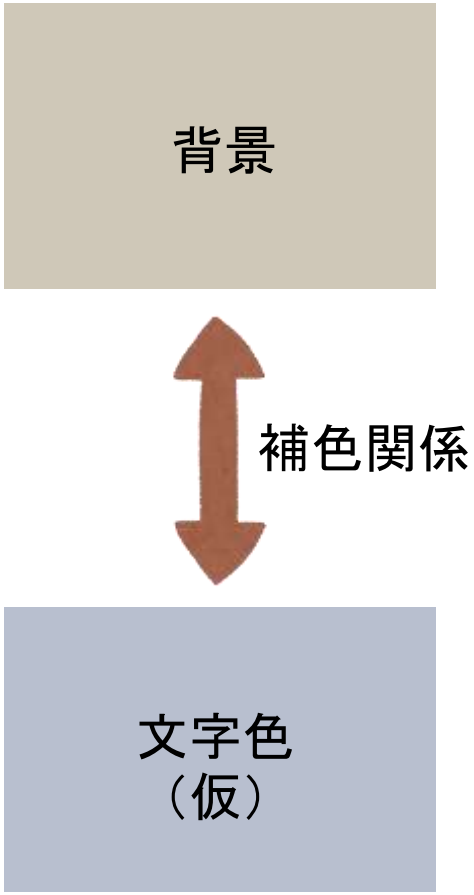
テキストbbox

平均値

補色計算

- 背景色の補色を計算

- $R = (\max(R, G, B) + \min(R, G, B)) - R$
- $G = (\max(R, G, B) + \min(R, G, B)) - G$
- $B = (\max(R, G, B) + \min(R, G, B)) - B$



コントラスト比

基準値
4.5:1

W3C勧告WCAG2.1（2023/9/21公開）

計算式

$$(L_1 + 0.05)/(L_2 + 0.05)$$

- L_1, L_2：相対輝度。必ず$L_1 > L_2$
- 相対輝度：最も暗い黒を0, 最も明るい白を1に正規化した相対的な明るさ
- 相対輝度の計算式

$$L = R * 0.2126 + G * 0.7152 + B * 0.0722$$

縁取り装飾

- テキストの強調効果
- 文字色が暗い→縁取り色：白
- 文字色が明るい→縁取り色：黒

縁取り前

縁取り後

実験概要

- 物体検出器：YOLOv8

<u>入力画像</u>

- ４枚
- 画像サイズ：任意
- 写真素材

- テキスト：手動入力（８文字以上）

実験：位置算出処理

- 橙色の点：重心座標（被写体の中心）
- 青色の点：テキストの位置座標

cat

foods

tea

egg

実験：改行処理

改行前

改行前

改行前

改行前

改行後

改行後

改行後

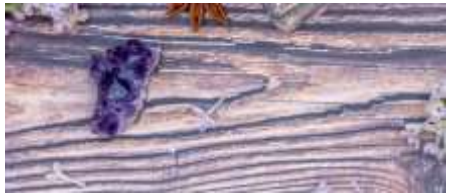
改行後

実験：補色導出処理

cat

foods

tea



egg

実験：コントラスト比

補色計算後

補色計算後

補色計算後

補色計算後

明暗調整後

明暗調整後

明暗調整後

明暗調整後

実験：縁取り装飾

縁取り前

縁取り前

縁取り前

縁取り前

縁取り後

縁取り後

縁取り後

縁取り後

まとめ

●サムネイルの自動生成手法の提案

手順

●動画の内容を要約

・動画から音声を抽出

・音声をテキストに変換

●画像に配置するテキスト生成

・要約文を生成

●YouTubeから画像候補を取得

●テキストを画像内に自動配置：実装済

まとめ

- テキストを画像内に自動配置
- テキスト位置算出：背景と周囲の余白を考慮
- 文字の大きさ：自動改行処理
- 文字色調整：背景に考慮、縁取り装飾