テキスト・音声・画像の協調的処理による放送型スポーツ映像におけるハイライト検出とインデクシング

松村晴琉
Haru Matsumura
u320062@st.pu-toyama.ac.jp

富山県立大学 工学部 情報システム工学科

14:50-18:00, Friday, December 19, 2025
N516, Toyama Prefectural University

1. 概要と研究目的

研究目的

- 放送スポーツ映像（特にアメリカンフットボール）からハイライトシーンを自動検出し, インデックス化する手法の提案.
- テキスト (CC), 音声, 画像のマルチモーダル情報を協調的に処理することで, 検出精度を向上させる.

従来の課題

- 既存研究は単一メディア（映像, 音声, テキストのいずれか）に依存するものが多く, ロバスト性に欠けていた.
- 異なる情報源を連携させ, 単一情報源では困難な検出精度向上を目指す.

検出対象のハイライトイベント

- タッチダウン (TD), フィールドゴール (FG) など, 試合の重要な得点シーン.

2. 提案手法の全体像（協調処理のフロー）

提案手法の主要なステップ

1. **テキスト処理 (CC)**:
   - 特定の「インデックス語句」の出現を検出し, ハイライト候補の時間窓 (time window)** を抽出する.
2. **音声処理**:
   - 時間窓内の短時間エネルギー (STE) を分析し, 歓声などの特徴的な音声パターンとマッチングさせる.
3. **画像処理**:
   - 映像をショット分割し, 支配色 (dominant color) の特徴がハイライトに不適切な場合, 候補を棄却 (reject) する.
4. **統合・インデックス化**: 全ての処理結果を統合し, 最終的なハイライトインデックスを作成する.

Closed Caption (CC) の活用

- CC はイベント発生を最も早期に示唆する情報源として, ハイライト候補の抽出に利用される.

インデックス語句の重み付け

- 語句 t の重み $w(t)$ は, ハイライト関連 CC (d_{pos}) と全 CC (d_{all}) での出現頻度に基づき,TF-IDF 類似手法で計算.
- 単語 t の重み $w(t)$:

$$w(t) = tf(t, d_{\mathrm{pos}}) \times \frac{1}{tf(t, d_{\mathrm{all}})} \quad \text{(式 1)}$$

- 2 語ペア $t \to t_2$ の重み $w(t \to t_2)$:

$$w(t \to t_2) = tf(t \to t_2, d_{\mathrm{pos}}) \times \log \frac{1}{tf(t \to t_2, d_{\mathrm{all}})} \quad \text{(式 2)}$$

4. テキスト処理：k-NN 分類器による確率推定

k-近傍法 (k-NN) の利用

- 抽出された語彙特徴ベクトル $\boldsymbol{x}$ がハイライトイベントである確率 $P(\boldsymbol{x})$ を推定する.

重み付き距離計算

- 各語彙 i の重み w_i' を考慮したユークリッド距離 d.

$$d = \sqrt{\sum_{i=1}^{n} w_i'(x_i - x_i{}^e)^2} \quad \text{(式 3)}$$

ハイライト判定

- k 個の最近傍のうち, ハイライトの数 k_{pos} を基に確率 $P(\boldsymbol{x})$ を算出.
- 確率 $P(\boldsymbol{x})$ が閾値 P（例: 1/2）を超えた場合, ハイライトと分類 $C(\boldsymbol{x}) = 1$.

5. 音声・画像による候補の絞り込み

音声処理：短時間エネルギー (STE)

- STE (E_l) は, 歓声などの音響的エネルギーの急激な変化を捉える.
- CC 分類器で得られた候補時刻を基点に,STE の時間変動パターンを解析し, ハイライトに特徴的な音響的証拠とマッチングさせる.

$$E_l = \sqrt{\frac{1}{L}\sum_m [\boldsymbol{X}(m)\boldsymbol{W}(l-m)]^2} \quad \text{(式 5)}$$

画像処理：支配色による棄却

- 映像をショットに分割し, 各ショットの支配的な色を分析する.
- スポーツ映像特有の背景色（アメフトの緑のフィールドなど）と著しく異なるショットを検出することで, ハイライトとは無関係なシーン（例: CM, 観客席）を棄却し, 誤検出を減らす.

6. 実験結果と考察

実験設定

- 対象: NFL (National Football League) の試合映像.
- 比較: CC のみの手法 (**CC**) と、CC + 音声・画像の協調処理 (**CC+**).

表 1: ハイライト検出結果

手法	TD	FG	平均検出率
CC	23/32 (72%)	15/26 (58%)	66%
CC+ (協調処理)	**23/28 (82%)**	**15/19 (79%)**	**81%**

- **結論**: 音声と画像情報が,CC のノイズ除去と時間区間の正確な特定に大きく貢献した.

7. 結論と今後の展望

結論

- 放送スポーツ映像のハイライト検出において、テキスト, 音声, 画像の協調処理が有効であることを実証した.
- 特に,CC の候補を音声（STE）と画像（支配色）で絞り込むアプローチにより, 高い検出率とインデックス化精度を達成した.

今後の展望

- 野球（ホームラン検出）など, 他のスポーツへの汎用性の検証.
- 処理速度の最適化と, より詳細なイベント分類への拡張.