

視覚特徴およびタグ共起を用いた 大規模Webビデオショットランキング

電気通信大学 大学院情報理工学研究科
総合情報学専攻

Do Hang Nga 柳井 啓司

背景

- **Web動画: 無限に存在、無料で取得可能**
 - YouTube, Daily Motion etc.

→ Web動画による動作データ収集
- **ただし、Web上の動画はノイズが多い**

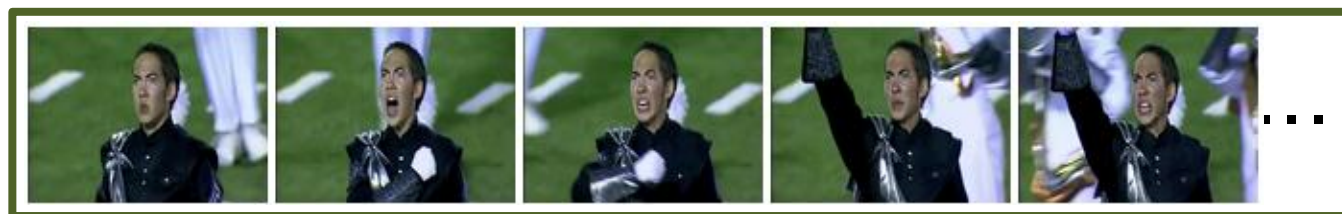
関連動画



You Tube



*Play
trumpet*

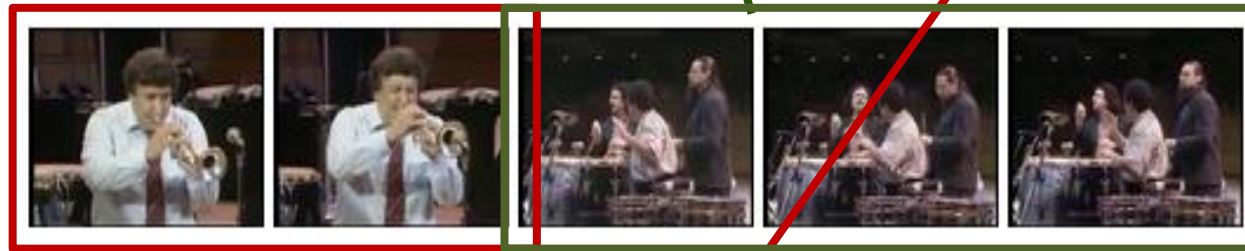


非関連動画

非対応ショット

You Tube

*Play
trumpet*



対応ショット

目的

You Tube

*Play
trumpet*



特定動作
の対応ショット
の自動収集

動作との関連性によるショットランキング

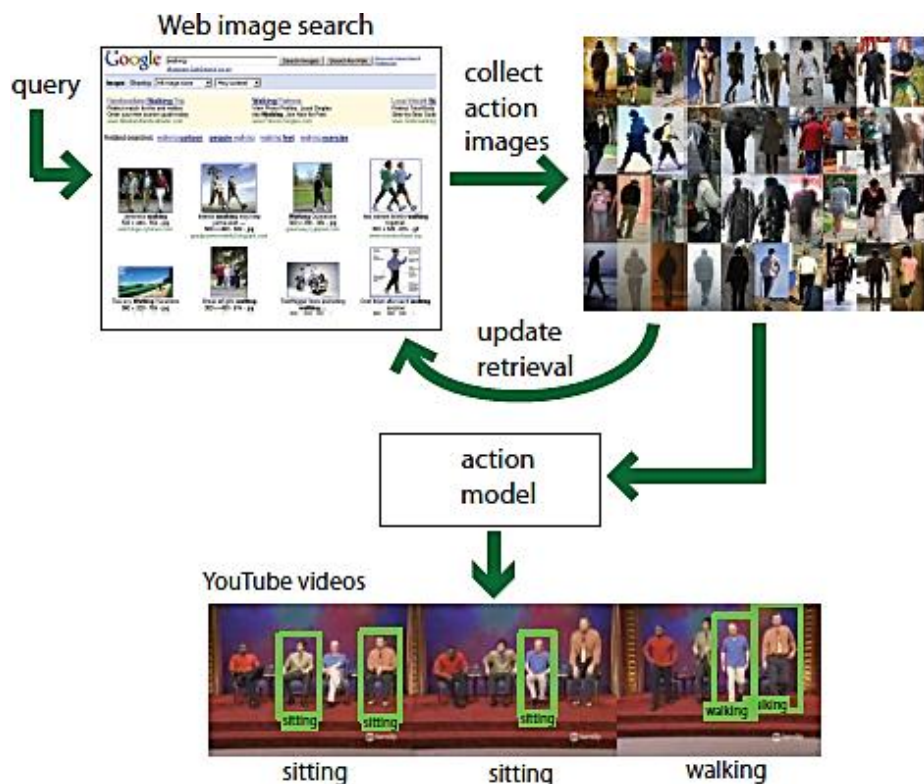
Rank 1

Rank 2



関連研究

- N. I. Cinbins, R. G. Cinbins, and S. Sclaroff.
Learning actions from the web. ICCV'09.
 - Web画像を利用して動作モデルを自動学習
 - 学習したモデルによるWeb動画アノテーション



*Cinbis*ら

Web画像、静的特徴

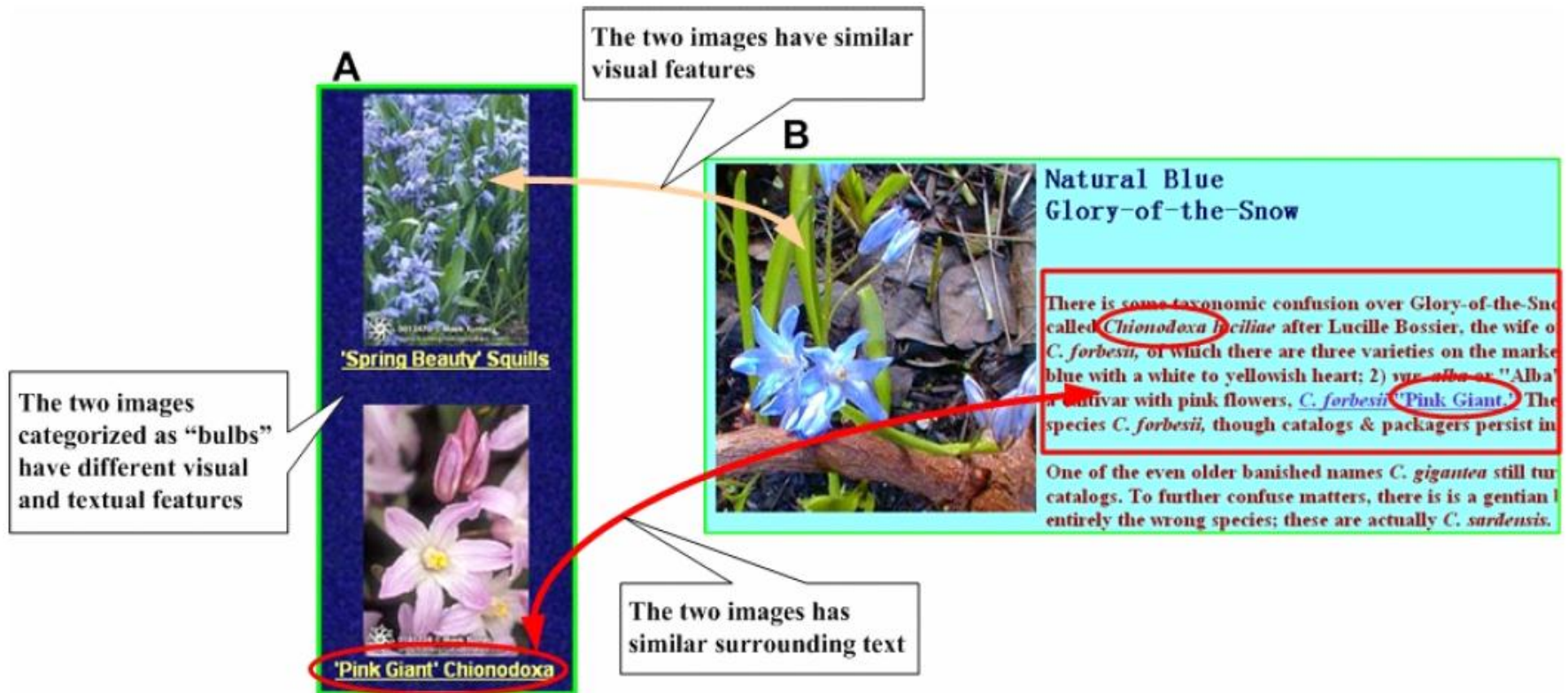
本研究



Web動画、Web画像、
時空間特徴、テキスト

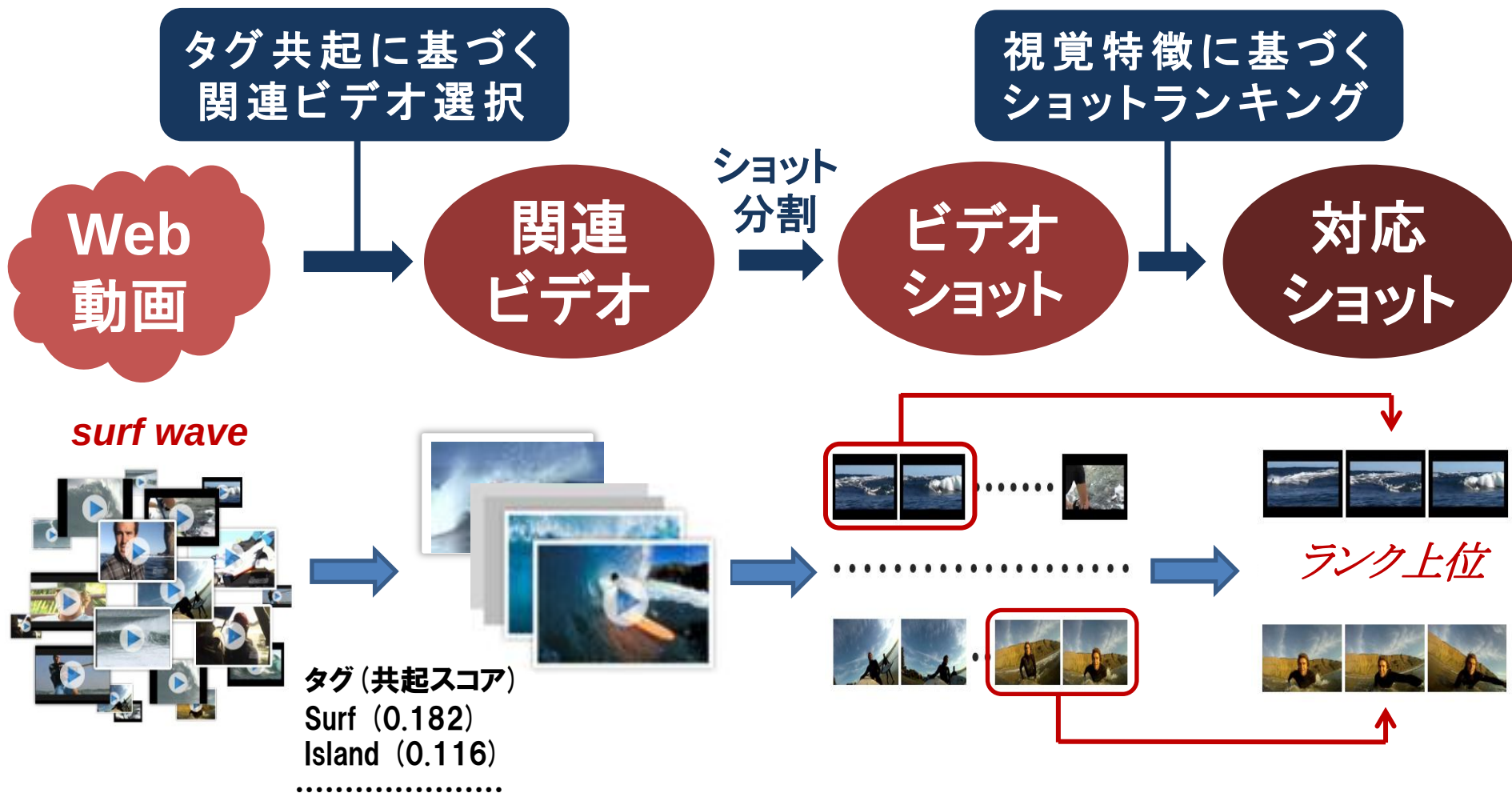
関連研究

- X. J. Wang, W. Y. Ma, G. R. Xue, X. Li.
Multi-model similarity propagation and its application for web image retrieval. ACM MM'04.
 - Web画像とそれらのテキストのリンク構造による類似画像検索



既存手法

既存手法[ICCV'11]



[ICCV'11] Do Hang Nga and Keiji Yanai: Automatic Construction of an Action Video Shot Database using Web Videos. ICCV2011.

タグ共起による動画選択

- ビデオVのタグ共起スコア:

$$Sc(V/t) = \frac{1}{|\mathcal{T}_V|} \sum_{t_i \in \mathcal{T}_V} \log_2 F(t, t_i)$$

$F(t, t_i)$: t と t_i の両方が付けられたビデオの数

$\mathcal{T}_V$: ビデオVのタグの集合

VisualRank^(*)によるショットランキング

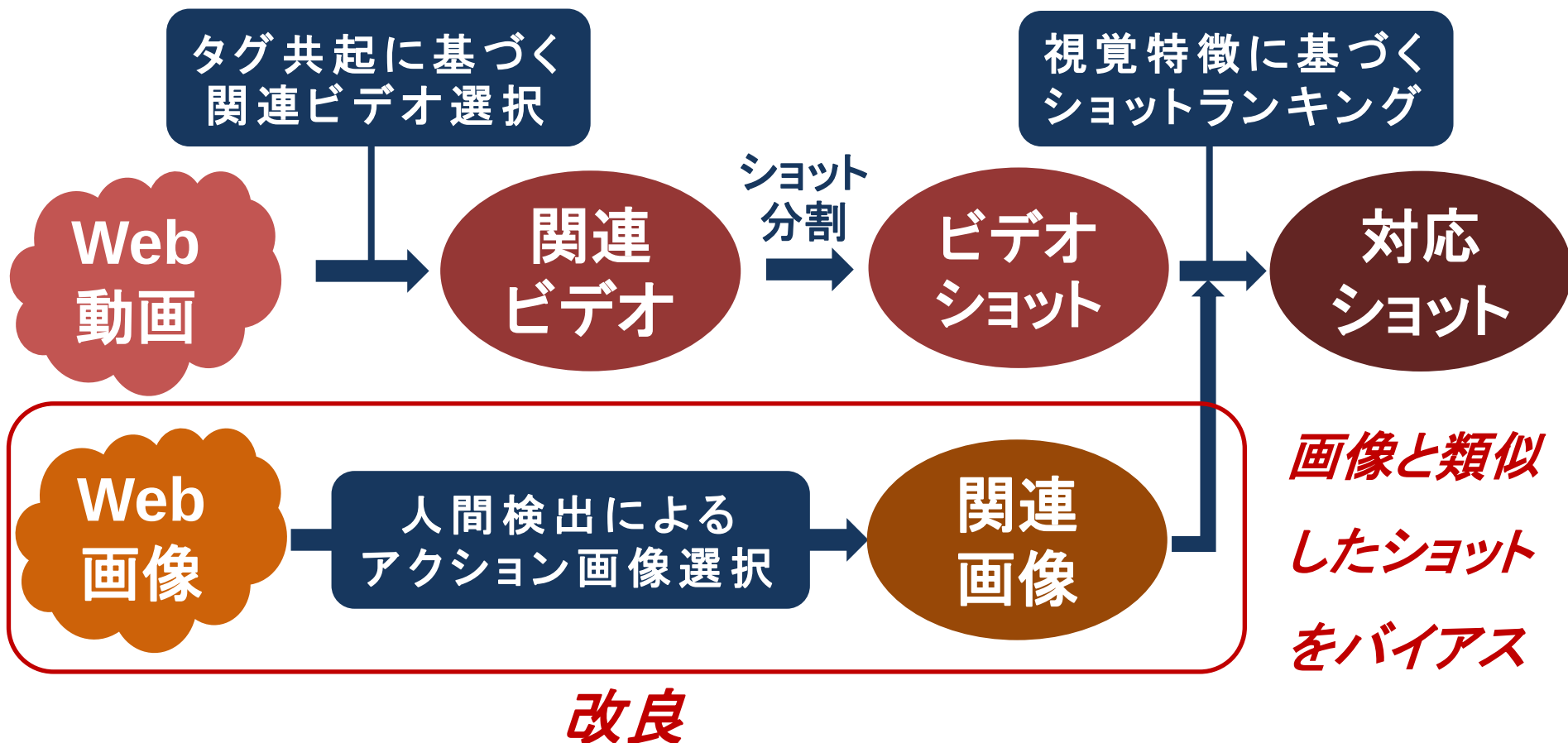
$$\underset{\substack{\uparrow \\ \text{ランクベクトル}}}{r} = d \underset{\substack{\uparrow \\ \text{類似度行列}}}{S^*} r + (1 - d) \underset{\substack{\uparrow \\ \text{補正ベクトル}}}{p}$$

S_{ij} = ショット*i*とショット*j*の時空間特徴の
ヒストグラムのインターセクション

$$p_i = \begin{cases} \frac{1}{k} & (i < k) \\ 0 & (i \geq k) \end{cases} \quad \begin{array}{l} \text{タグ共起スコア上位}k\text{ショット} \\ \text{はバイアスされる} \end{array}$$

(*) Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *PAMI*, 30(11):1870–1890, 2008.

既存手法[LSVSM'12]



[LSVSM'12] Do Hang Nga and Keiji Yanai: Automatic Collection of Web Video Shots Corresponding to Specific Actions using Web Images.

CVPR Workshop on Large-Scale Video Search and Mining (LSVSM). Jun. 2012.

画像導入によるショットバイアス

Play
trumpet

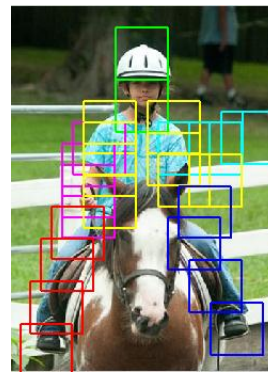
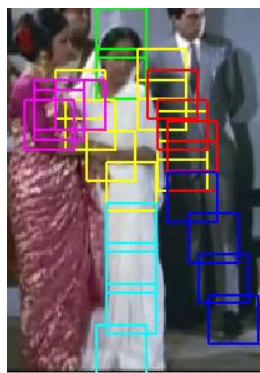
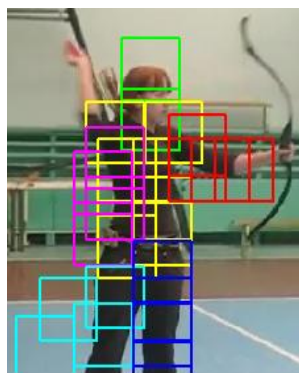


ビデオショット

アクション画像

p_i ~ ショット i の画像との類似度

ポーズマッチングによる類似度計算



ポーズ
推定

ボディパート
配置

オリエンテーション
計算

ユークリッド
距離

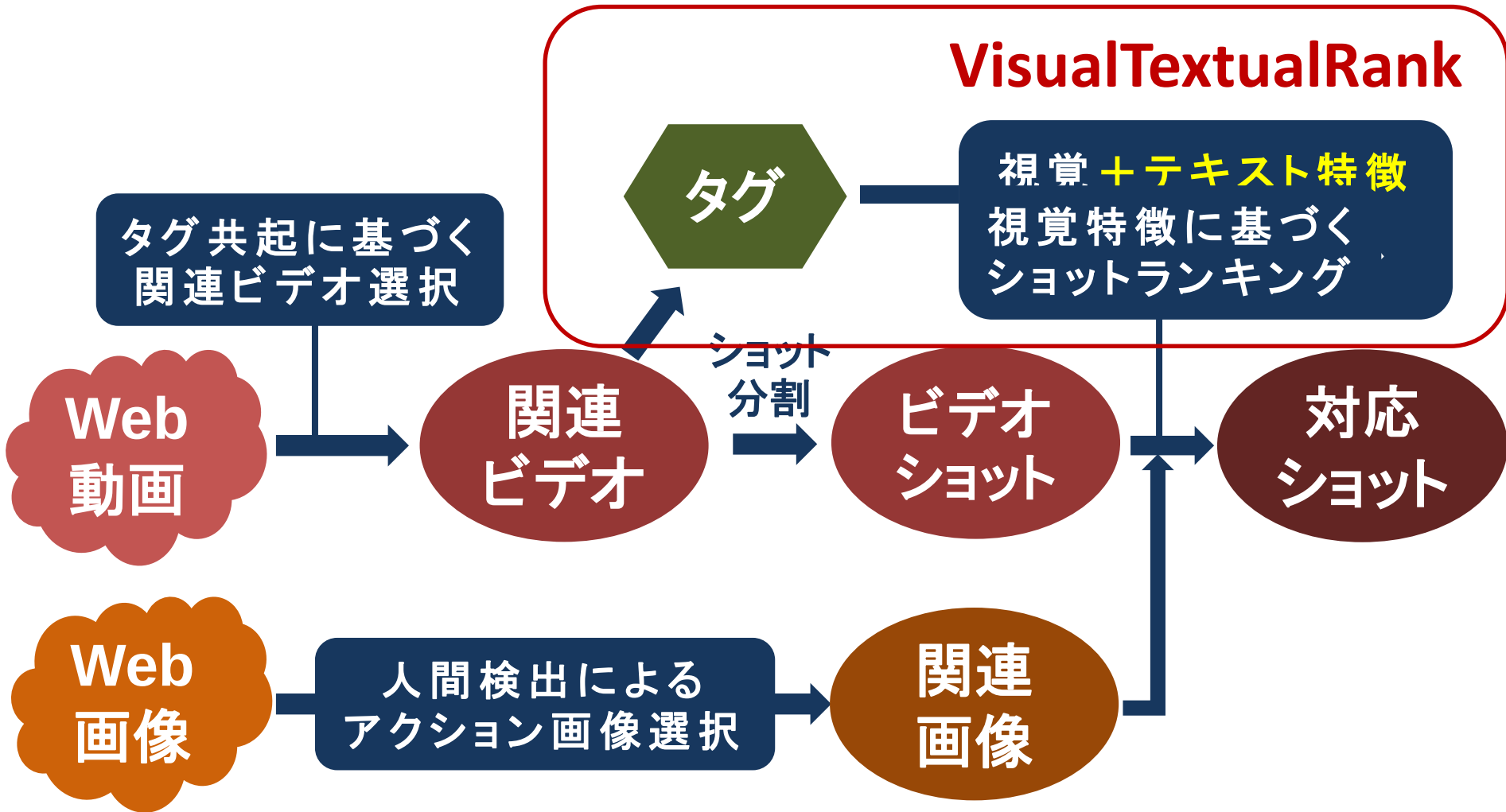
類似度
計算

Y. Yang, D. Ramanan: Articulated pose estimation
using flexible mixture of parts. CVPR 2011.

提案システム

提案システム

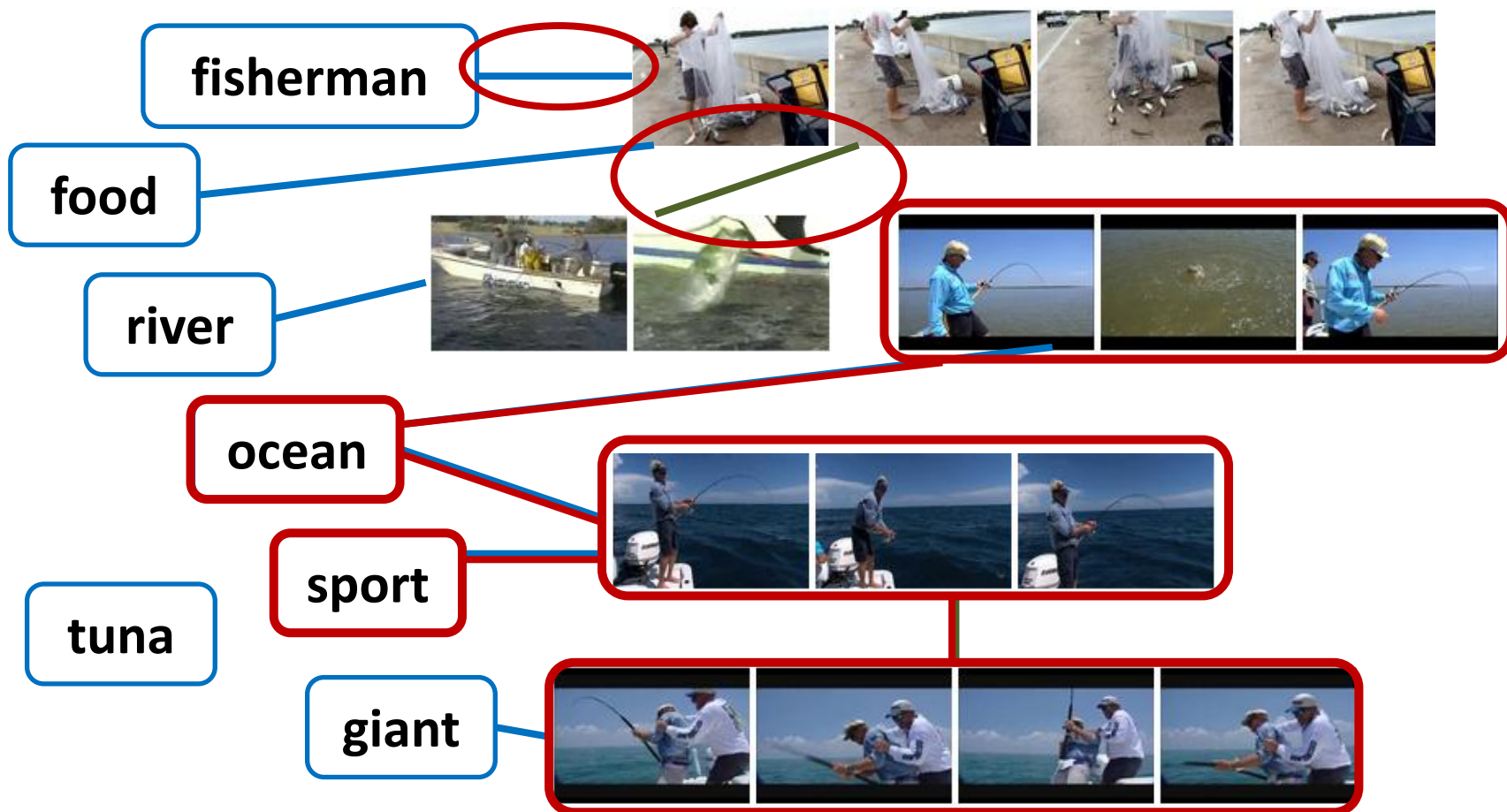
改良



VisualTextualRank: VisualRankの拡張

- アイデア

➤例: "catch+fish"



VisualTextualRank

ショット
ランク

RS_k

ショット
類似度行列

SM^*

ショットータグ
類似度行列

SC^*

RT_k

$$= \alpha \times SM^* \times SC^* \times RT_k + (1 - \alpha)p$$

RT_{k+1}

$k+1$

$$= (SC')^* \times RS_k$$

タグランク

タグーショット類似度行列

ショット S とタグ t の類似度 =

S のビデオの t に対するタグ共起スコア

VisualTextualRank

- 同時にショットランキングとタグランキングを行なう
- タグランキング: コンテンツ情報も考慮することによってタグノイズの影響は減らす
- ショットランキング: テキスト情報を導入することによって視覚類似のショットしかもらわないことがない

実験

- 評価: Precision@100
 - 上位100ショットのなかの関連ショットの割合
 - 手動で評価
- 実験: 以下の4つの手法を比較
 - 既存手法1: VisualRank
 - 既存手法2: VisualRank+Web画像とのポーズマッチング
 - Exp.1: VisualTextualRank
 - Exp.2: VisualTextualRank+Web画像とのポーズマッチング
- データ: 既存手法で精度が悪かった13人間動作

実験結果

Action	既存手法 1	既存手法 2	Exp. 1	Exp. 2	Action	既存手法 1	既存手法 2	Exp. 1	Exp. 2
blow+candle	29	35	44	35	wash+dishes	15	40	39	39
wash+clothes	15	31	39	39	tie+necktie	11	24	28	27
eat+sushi	12	15	23	15	row+dumbell	16	30	30	33
jump+rope	26	39	49	47	play+guitar	28	38	41	43
catch+fish	28	45	59	54	serve+tennis	11	24	27	30
drive+car	28	35	40	34	roll+makizushi	24	26	26	27
slap+face	20	36	45	44	平均	20.2	32.1	37.7	35.9

17.5% ↑

画像導入の有効性

serve+tennis



row+dumbbell



blow+candle



eat+sushi



drive+car



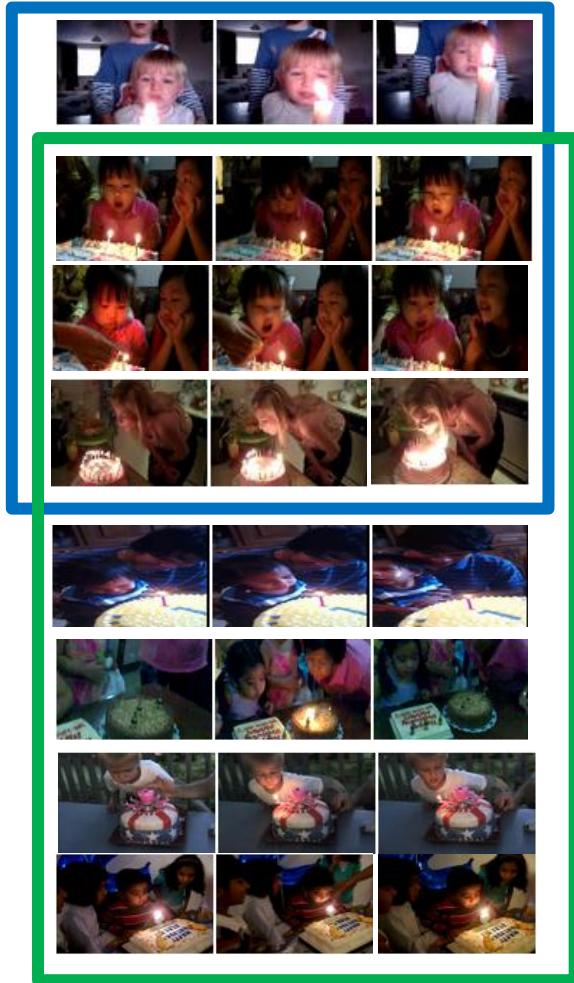
ポーズ
マッチング
が有効の
場合

ポーズ
マッチング
が失敗の
場合

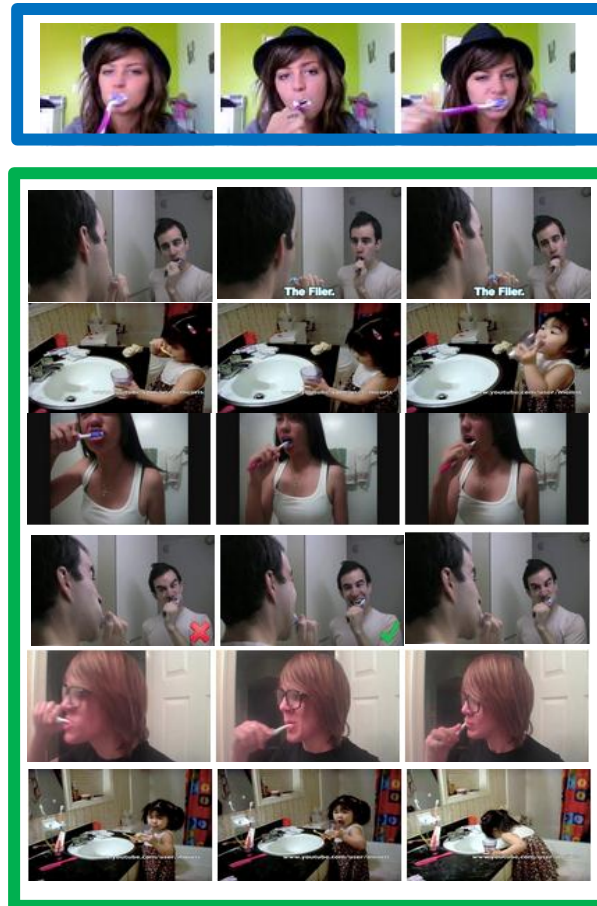
実験結果例

上位15ショットのなかの関連ショット

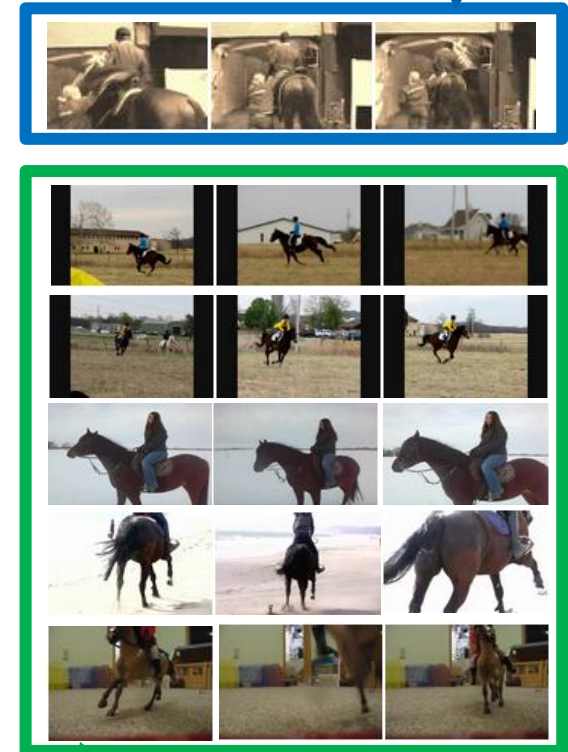
従来



blow candle



brush teeth



ride horse

提案

VisualTextualRankの有効性

- さらに実験
 - データ: 既存手法で**精度が良かった10人間動作**
 - 実験: 既存手法1 & Exp.1 (VisualTextualRank)

Action	既存手法1	Exp.1	Action	既存手法1	Exp.1
harvest+rice	49	46	shave+mustache	31	30
play+trumpet	41	59	dance+flamenco	45	53
ski	49	60	clean+floor	31	38
dance+hiphop	43	68	pick+lock	30	28
play+drum	40	45	swim+crawl	36	49
			平均	39.5	47.6

↓ 従来



 **提案**

play trumpet

結論

- 視覚特徴およびタグ共起を用いた大規模Web動画ショットランキングフレームワークを提案した
- 視覚特徴のみを適用した既存フレームワークを改良できた