時空間特徴を用いた
Youtube動画からの特定動作対応ショットの
自動抽出

DO HANG NGA　　柳井 啓司
電気通信大学　情報工学科

背景

既存の

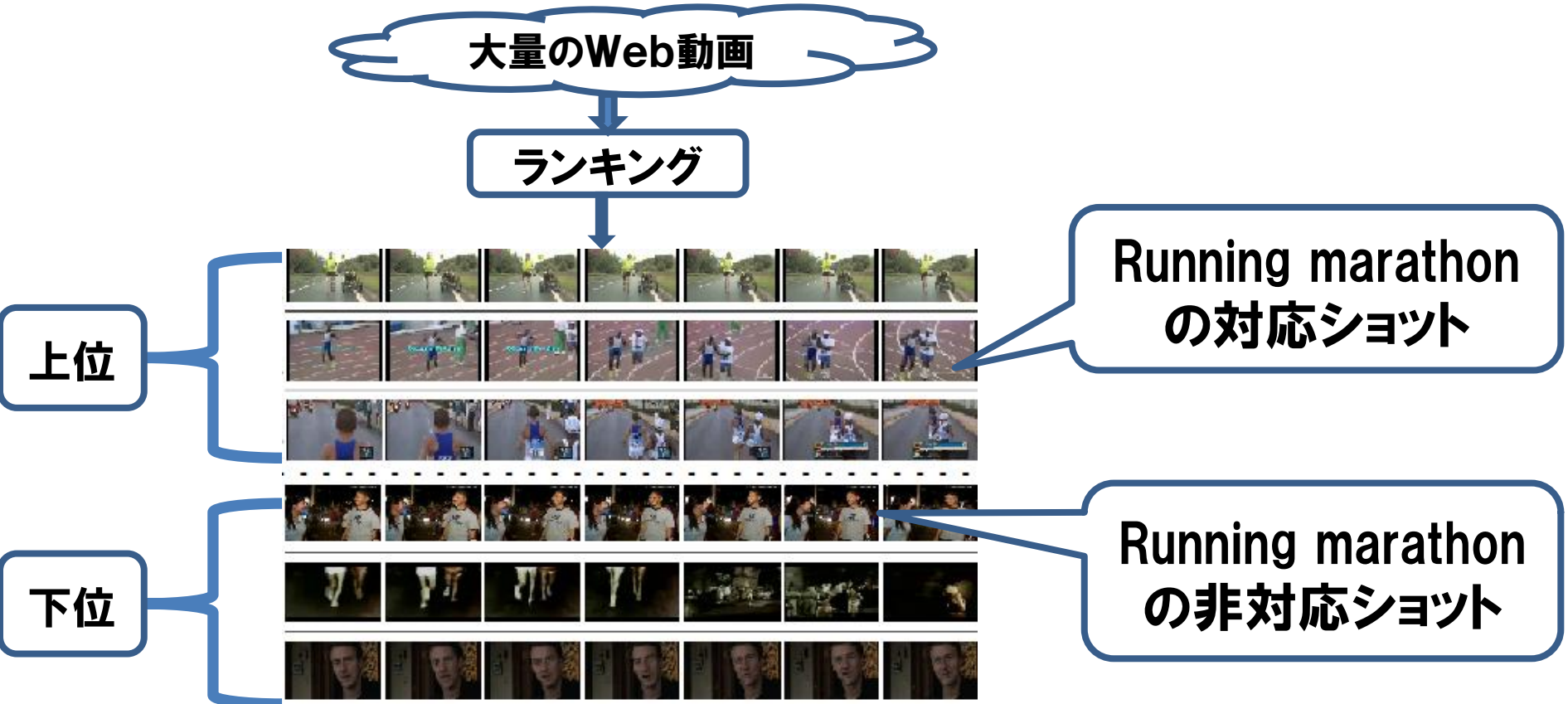
教師信号あり

動画量が少ない

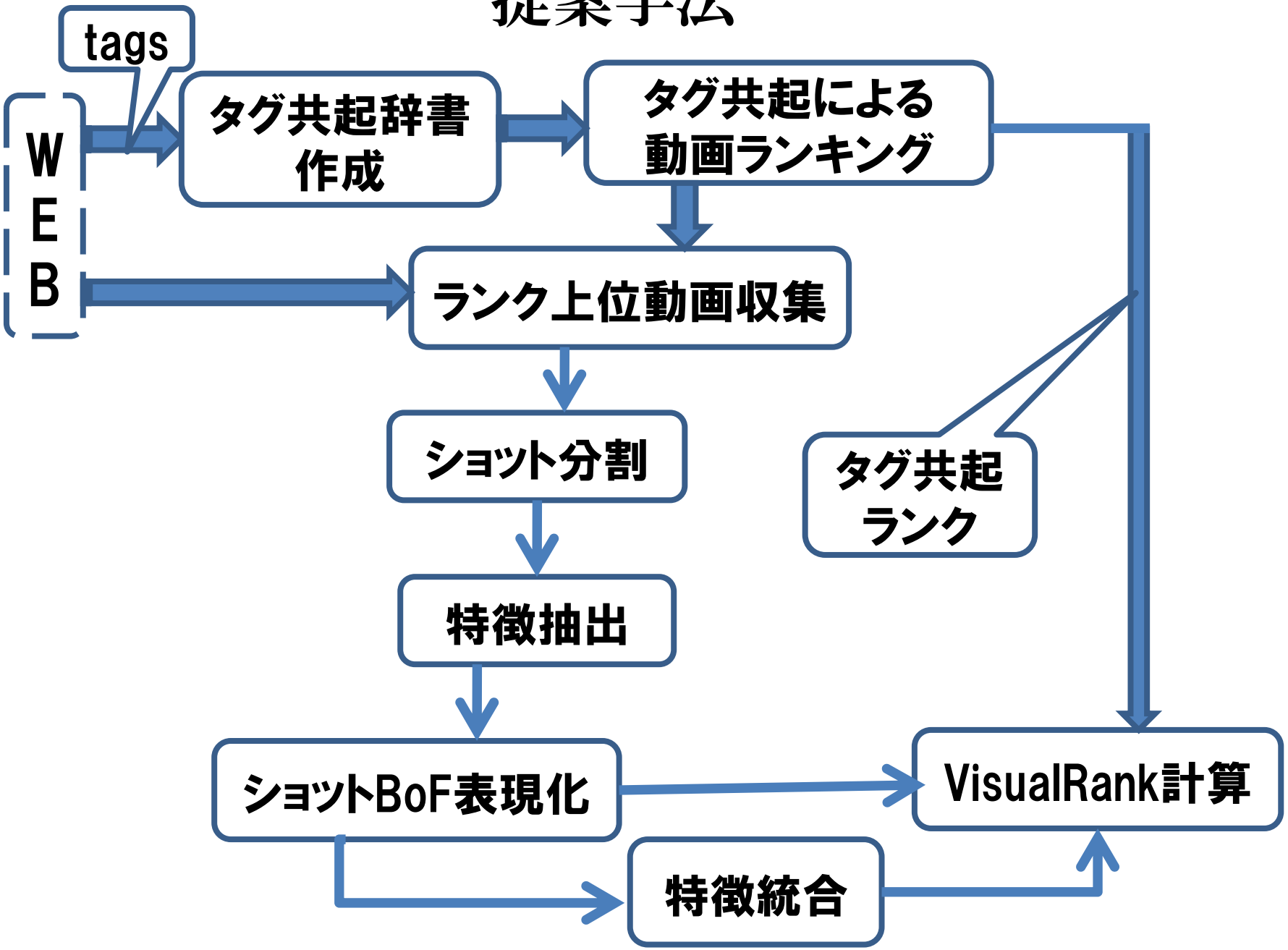
研究の目的

特定動作についてのWeb動画を使用して、

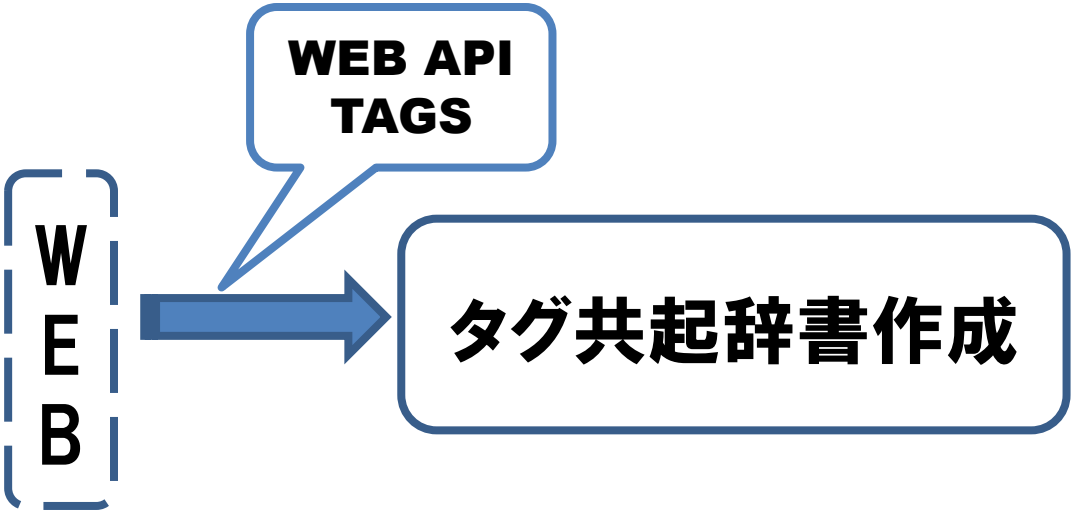
その動作の対応ショットを自動抽出



提案手法


tags
WEB
タグ共起辞書
作成
タグ共起による
動画ランキング
ランク上位動画収集
ショット分割
特徴抽出
タグ共起
ランク
ショットBoF表現化
VisualRank計算
特徴統合

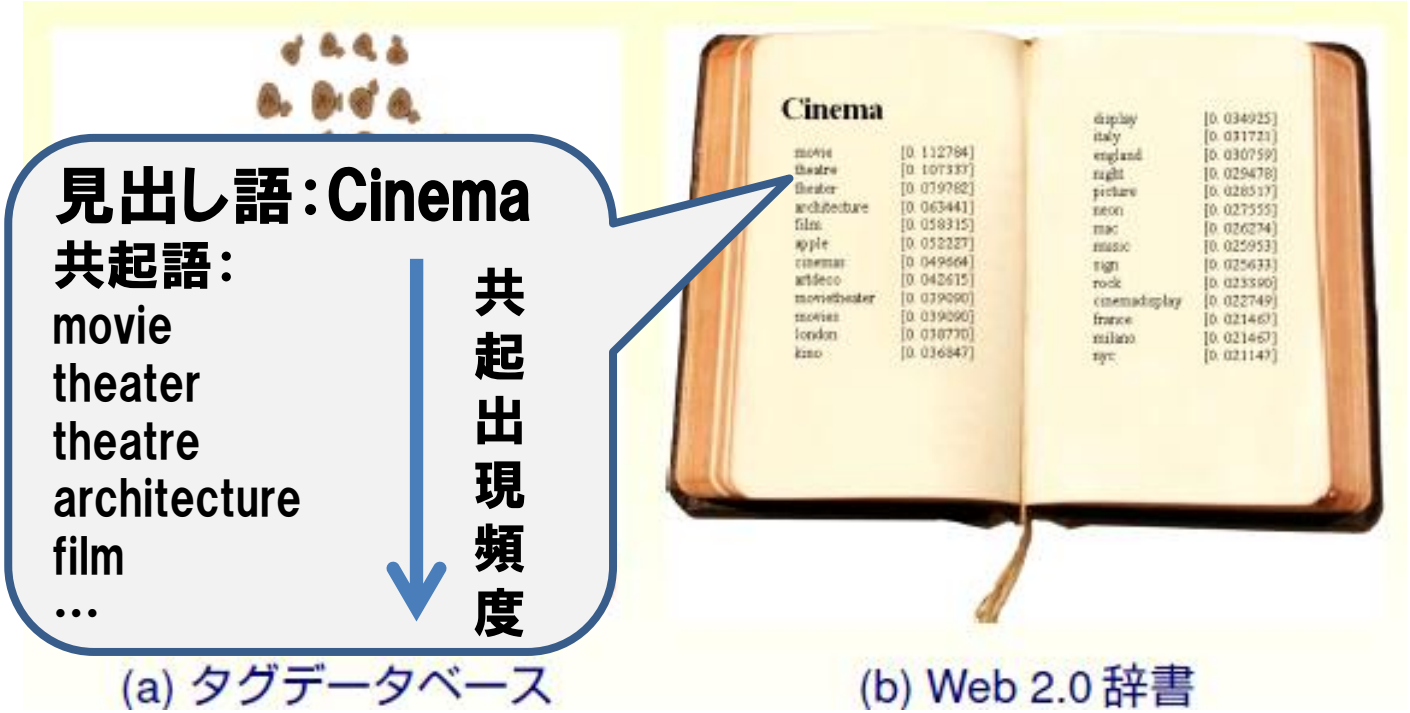
提案手法



タグ共起辞書

タグ共起辞書： Web2.0辞書[10]を適用したもの

Web2.0辞書：共起出現関係により言葉を定義する



[10] Q.Yang, X.Chen, G.Wang. Web2.0 Dictionary CVIR2008

タグ共起辞書作成

ステップ1 各動作について1000動画のタグを収集

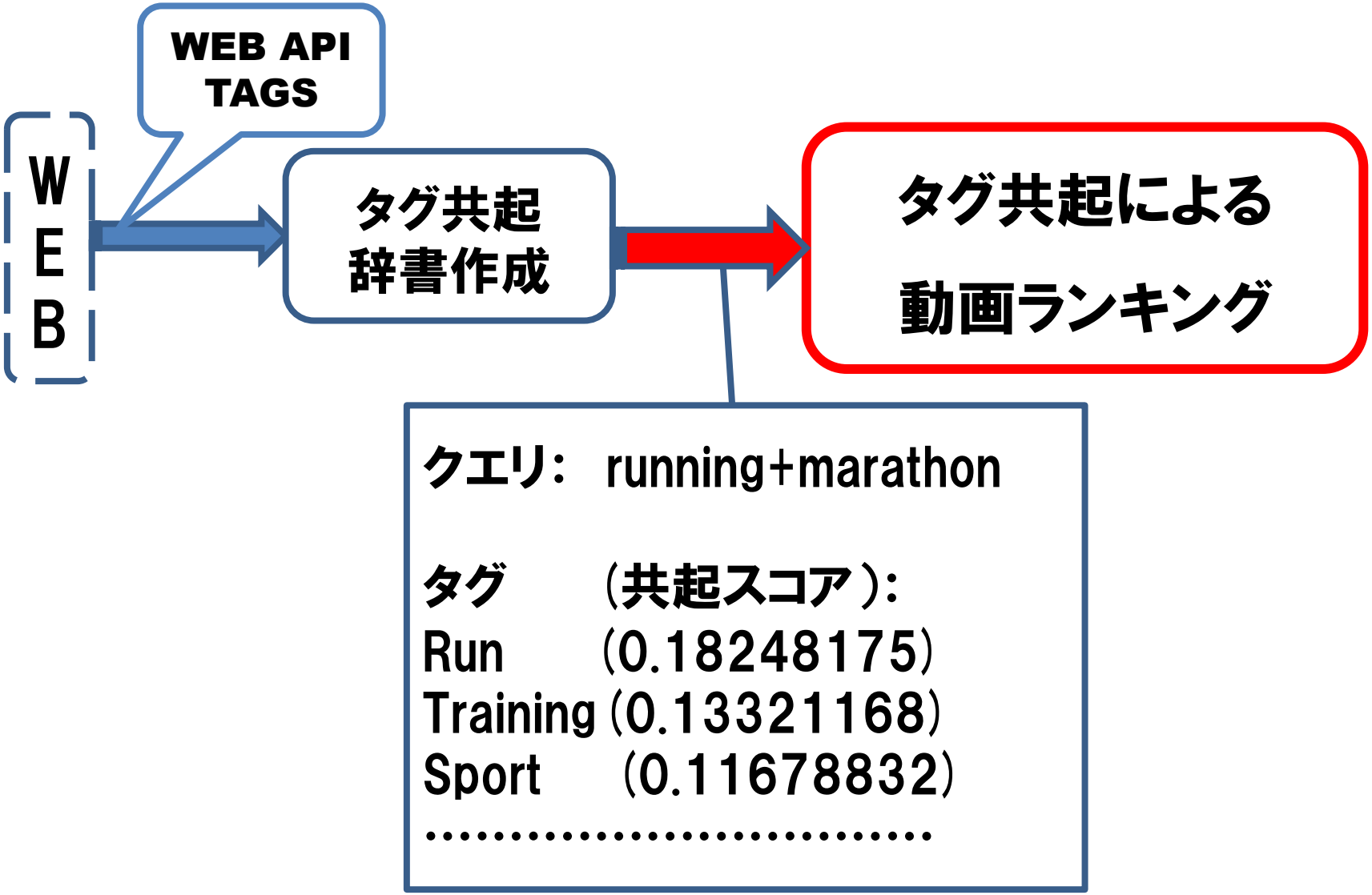
ステップ2 タグを集計し、出現頻度上位2000タグについて、それぞれ1000動画のタグを収集

ステップ3 約200万動画のタグのうち、5回以上出現したタグの共起頻度を集計

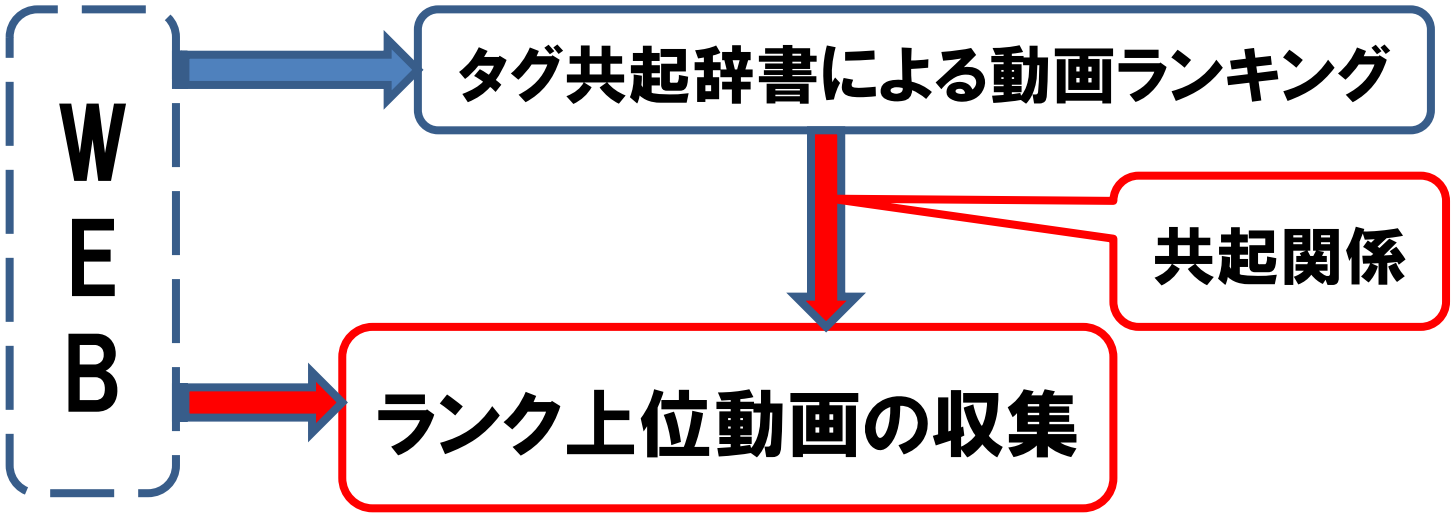
$$P(b \mid a) = \frac{count(a,b)}{count(a)}$$

where $\begin{cases} count(a,b): a, b\text{の共起出現回数} \\ count(a): \quad a\text{の出現回数} \end{cases}$

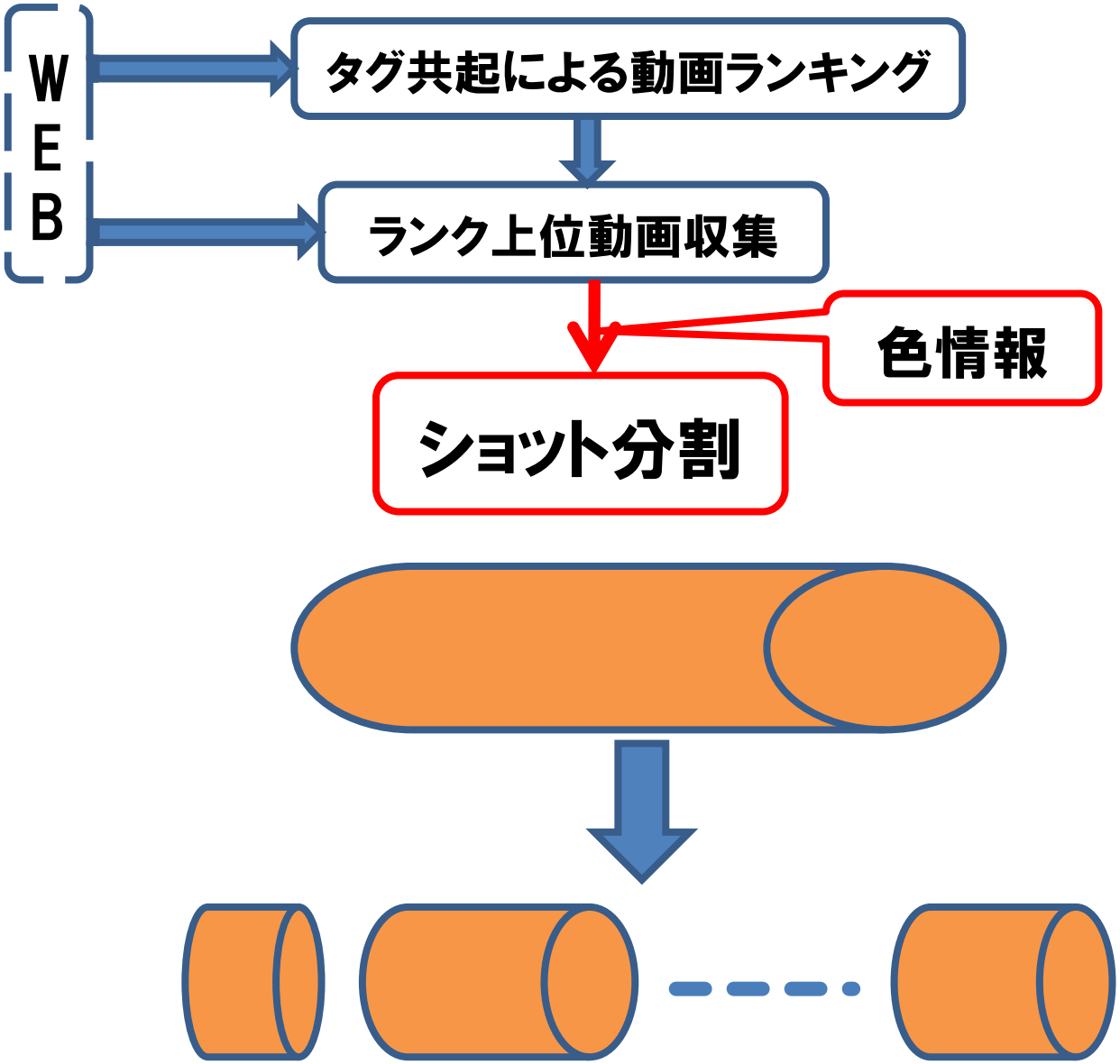
提案手法


WEB API
TAGS
WEB
タグ共起
辞書作成
タグ共起による
動画ランキング
クエリ: running+marathon
タグ (共起スコア):
Run (0.18248175)
Training (0.13321168)
Sport (0.11678832)
……

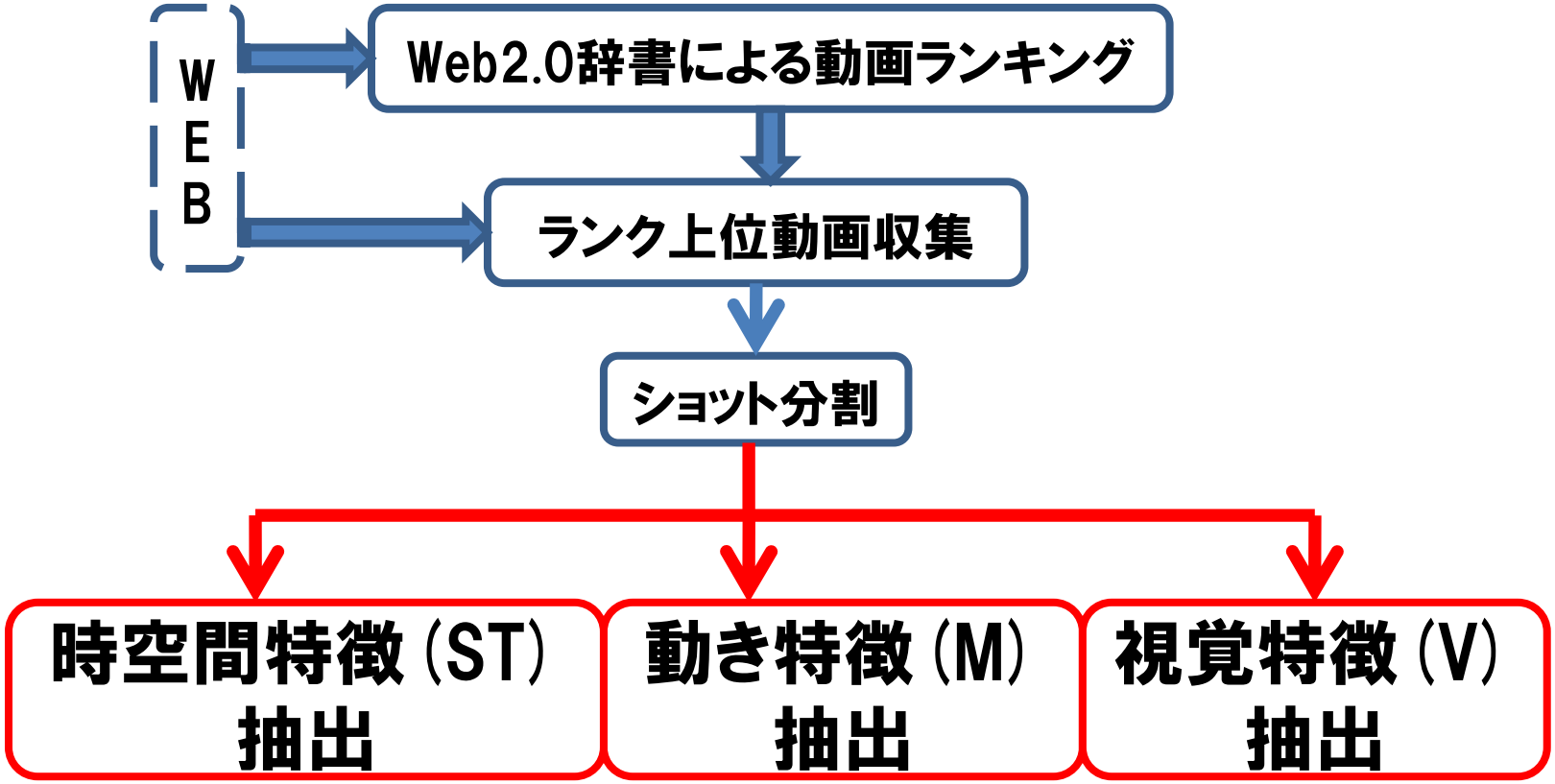
提案手法



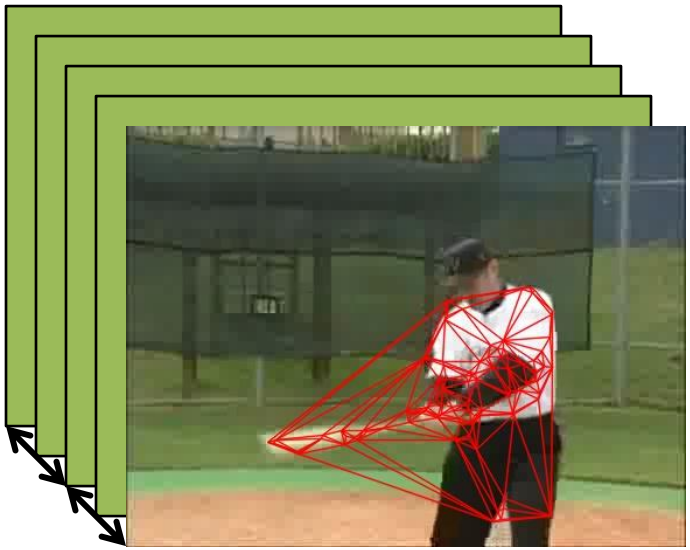
提案手法



提案手法


WEB
Web2.0辞書による動画ランキング
ランク上位動画収集
ショット分割
時空間特徴(ST)抽出
動き特徴(M)抽出
視覚特徴(V)抽出

時空間特徴[5]



①. 5フレームを1ユニットとする

②. SURFを抽出、オプティカルフローを計算

③. 動きがある点：特徴点

④. ドロネー三角形を作成
以降三点で一組の特徴と考える

⑤. ユニットを更に区切り,それぞれの
インターバルから動き特徴を抽出

⑥ 視覚特徴と動き特徴を統合し、
特徴をヒストグラム化する

[5]: A.Noguchi and K.Yanai: A SURF-based Spatio-Temporal Feature for feature-fusion-based action recognition

統合特徴

- 統合特徴＝時空間特徴＋動き特徴＋視覚特徴

局所動き特徴＋動的視覚特徴

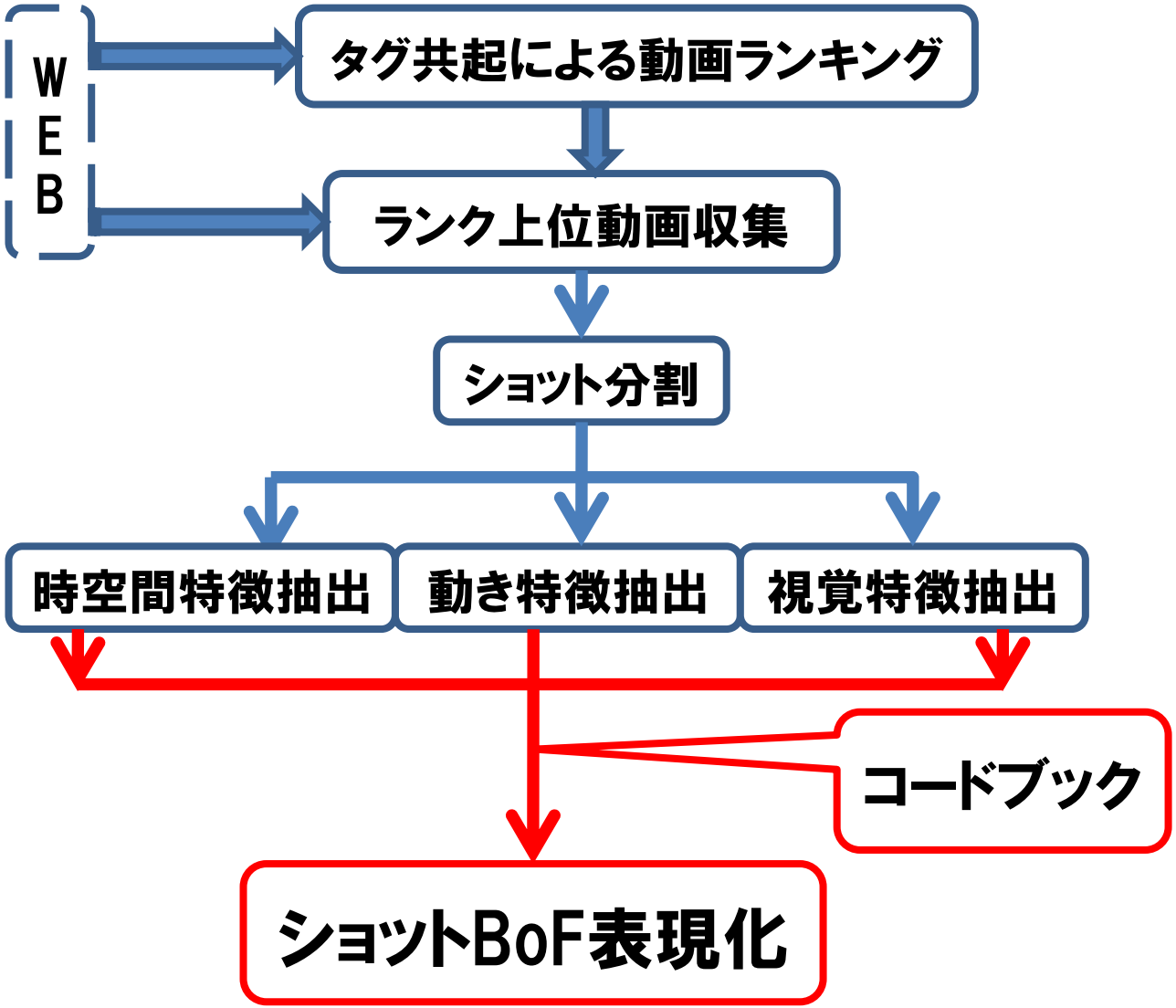
全体的

静的

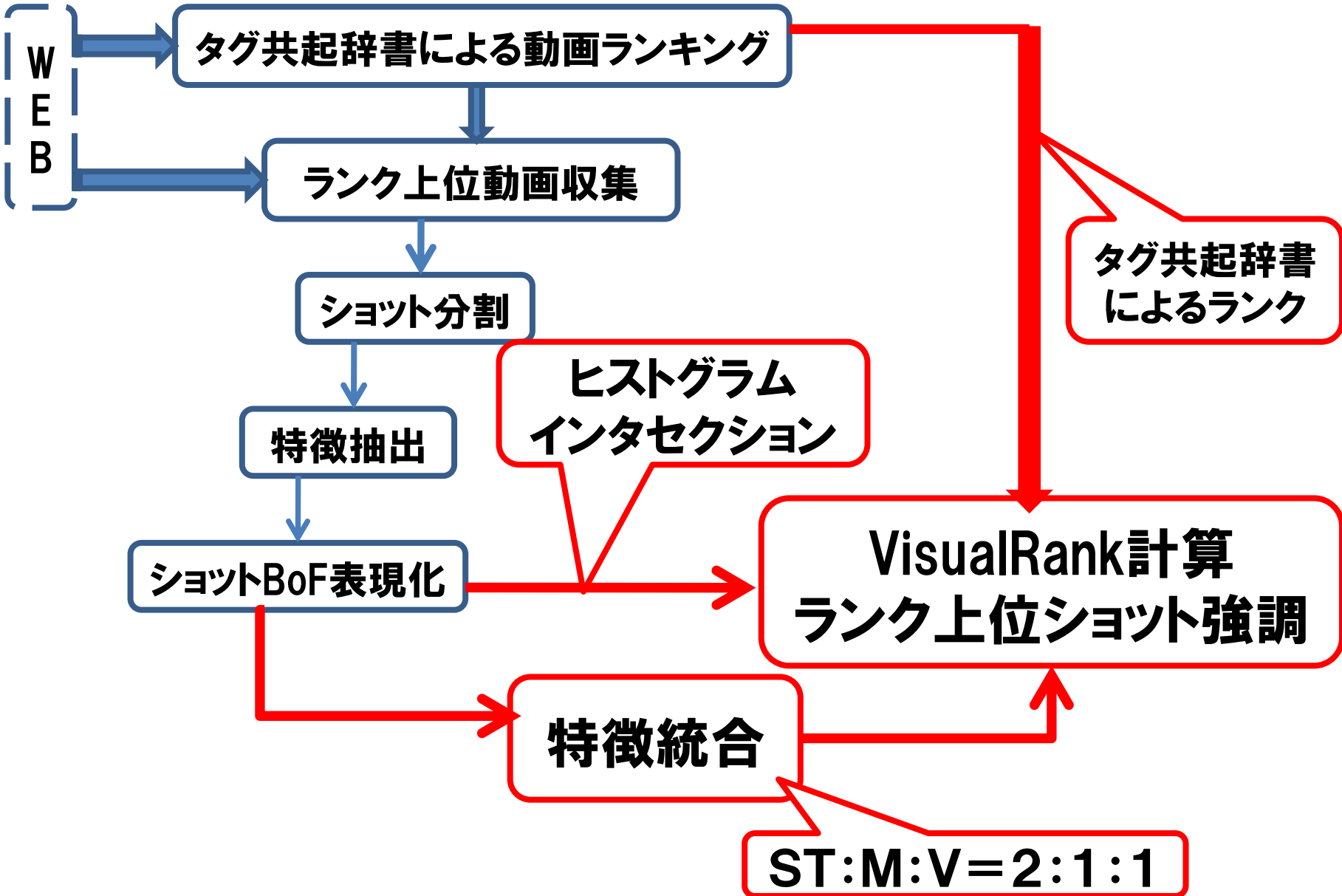
- 動き特徴
  - Lucas-Kanade法によりグリッドで抽出
  - 8方向7段階の強さのヒストグラムに投票
- 視覚特徴
  - 画像を20*20のグリッドに分割
  - 6方向4周期のガボールフィルタを利用

20*20*6*4次元のヒストグラム

提案手法


WEB
タグ共起による動画ランキング
ランク上位動画収集
ショット分割
時空間特徴抽出
動き特徴抽出
視覚特徴抽出
コードブック
ショットBoF表現化

提案手法


WEB
タグ共起辞書による動画ランキング
ランク上位動画収集
ショット分割
特徴抽出
ショットBoF表現化
ヒストグラム
インタセクション
タグ共起辞書
によるランク
VisualRank計算
ランク上位ショット強調
特徴統合
ST:M:V＝2:1:1

ショットのVisualRankの計算

- VisualRank[12]計算：

$$\mathbf{r} = dS^* \mathbf{r} + (1-d)p \quad \text{where} \begin{cases} \mathbf{r}: & \text{ランク値ベクトル} \\ S^*: & \text{正規化した類似度行列} \\ d: & \text{補正パラメータ} \\ p: & \text{補正ベクトル} \end{cases}$$

$$s(H_1, H_2) = \sum_{i=1}^{|H|} min(H_{1i}, H_{2i})$$

- **特徴統合を適用：**

$$S_{combined} = w_{st} \times S^*_{st} + w_m \times S^*_m + w_v \times S^*_v$$

$$where \quad w_{st} = \frac{1}{2}, w_m = \frac{1}{4}, w_v = \frac{1}{4}$$

VisualRank計算に共起スコア適用

- 共起スコア適用なし：

$$p = \left[\frac{1}{n}\right]_{nx1}$$

- 共起スコア適用あり(共起スコアの高いショットに強調)：

$$p = v_j = \begin{cases} \frac{1}{m}, 1 \leq j \leq m \\ 0, m < j \leq n \end{cases}$$

実験設定： $n \approx 2000, m = 1000$

実験データ

- データセット：NoguchiらのYoutubeデータセット

walking

shoot

jumping

eating

実験データ

動作	動画数	ショット数
Batting	144	2000
Eating+ramen	133	1630
Jumping+trampoline	142	1650
Running+marathon	166	2000
Shooting+football	149	2000
Walking+street	139	1772
平均	145	1843

実験

- **結果評価法**:ランキング後のランク上位1～100ショットについての**適合率**を利用
- **目的**:*タグ共起辞書*と*特徴統合*の適用の有効性の検討

タグ共起辞書適用の有効性の検討

行った実験の一覧
(使用特徴：時空間特徴)

実験	動画利用の優先	補正ベクトルの強調	上位100適合率
1	×	×	33.7%
2	○	×	41%
3	○	○	47.3%

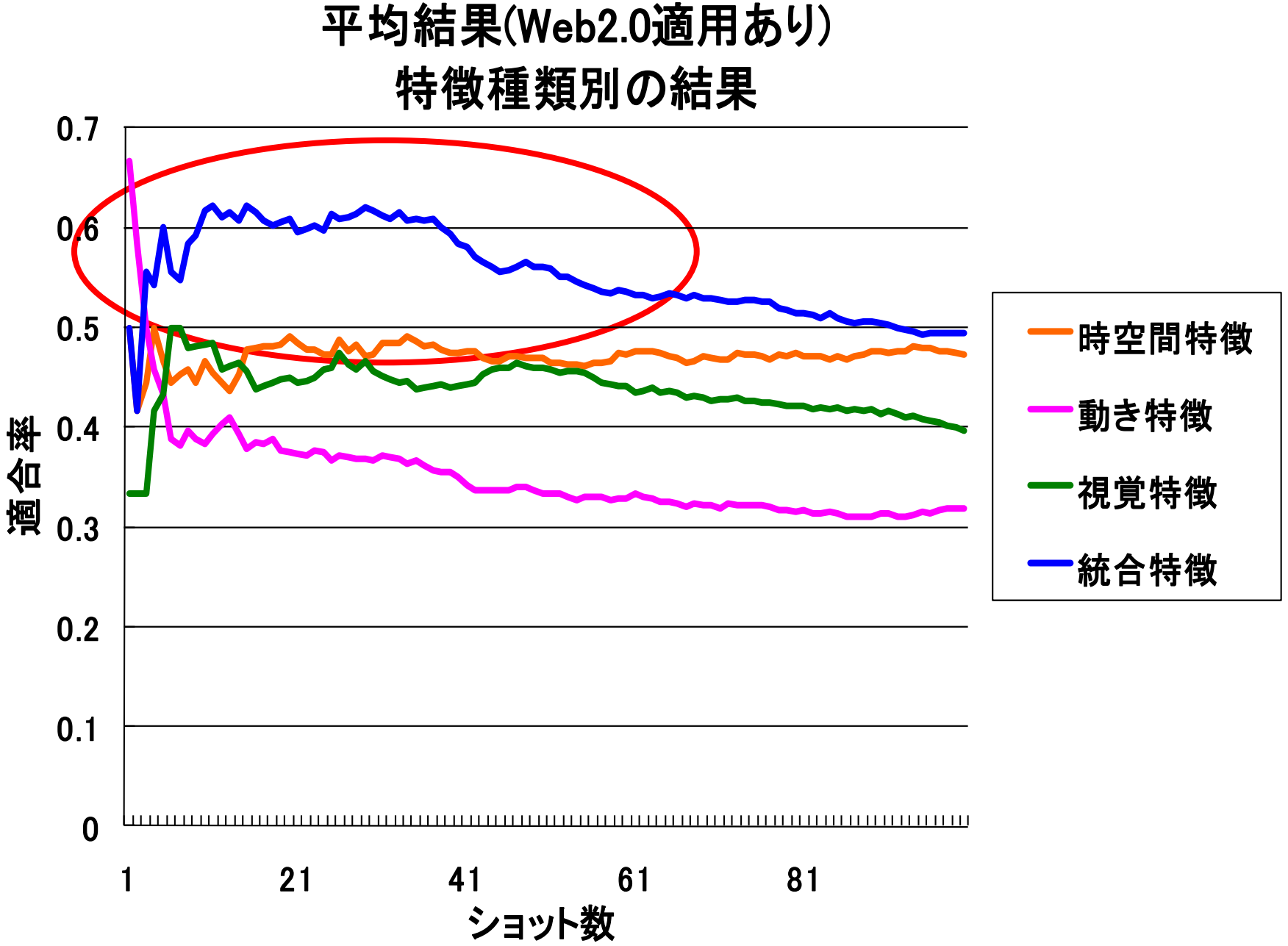
7.3%↑

6.3%↑

特徴統合の有効性の検討

行った実験の一覧

実験	使用特徴	上位100の適合率
3	**時空間特徴**	**47.3%**
4	動き特徴	**31.8%**
5	視覚特徴	**39.7%**
6	**統合特徴**	**49.5%**


平均結果(Web2.0適用あり)
特徴種類別の結果
適合率
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
21
41
61
81
ショット数
時空間特徴
動き特徴
視覚特徴
統合特徴

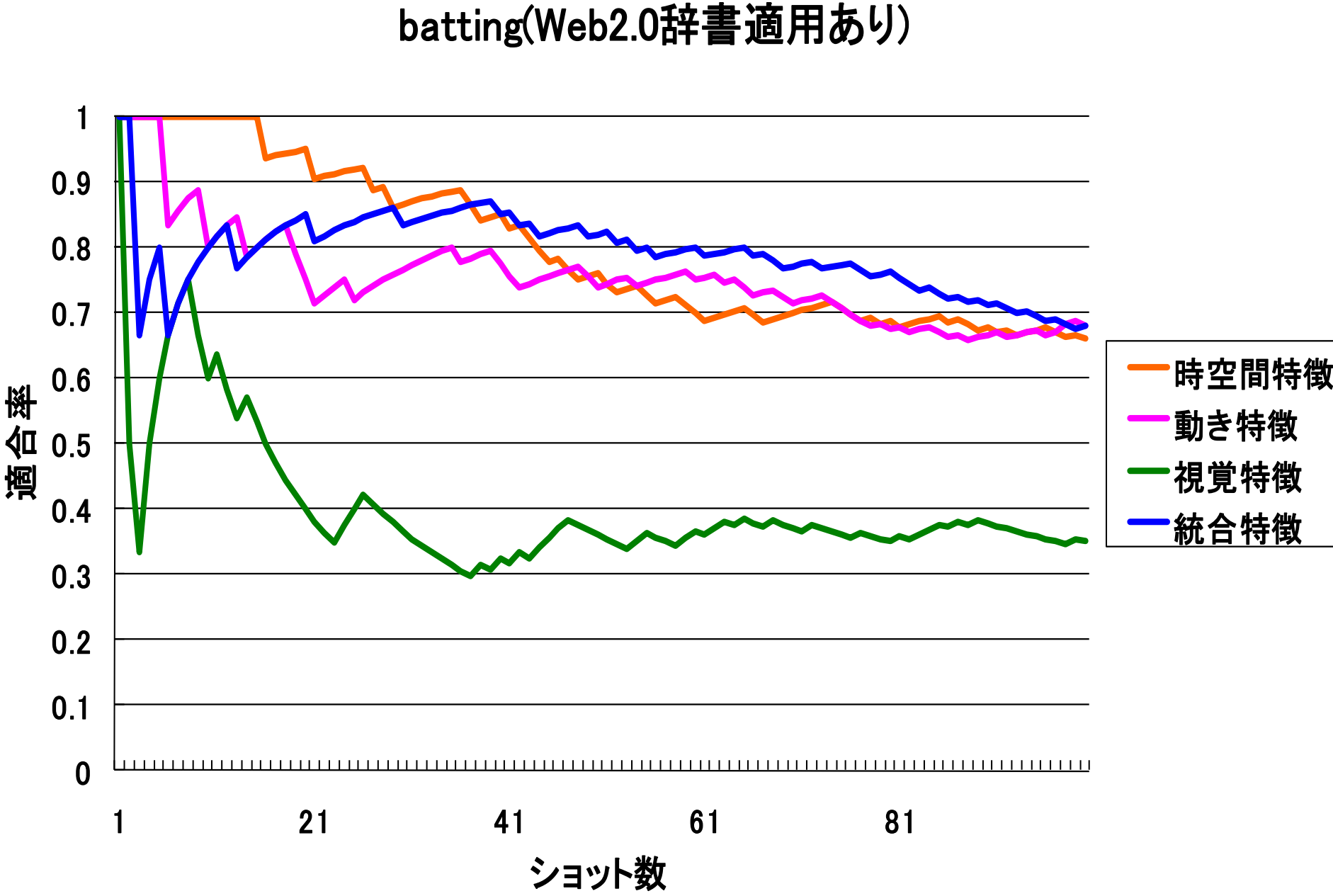
結論

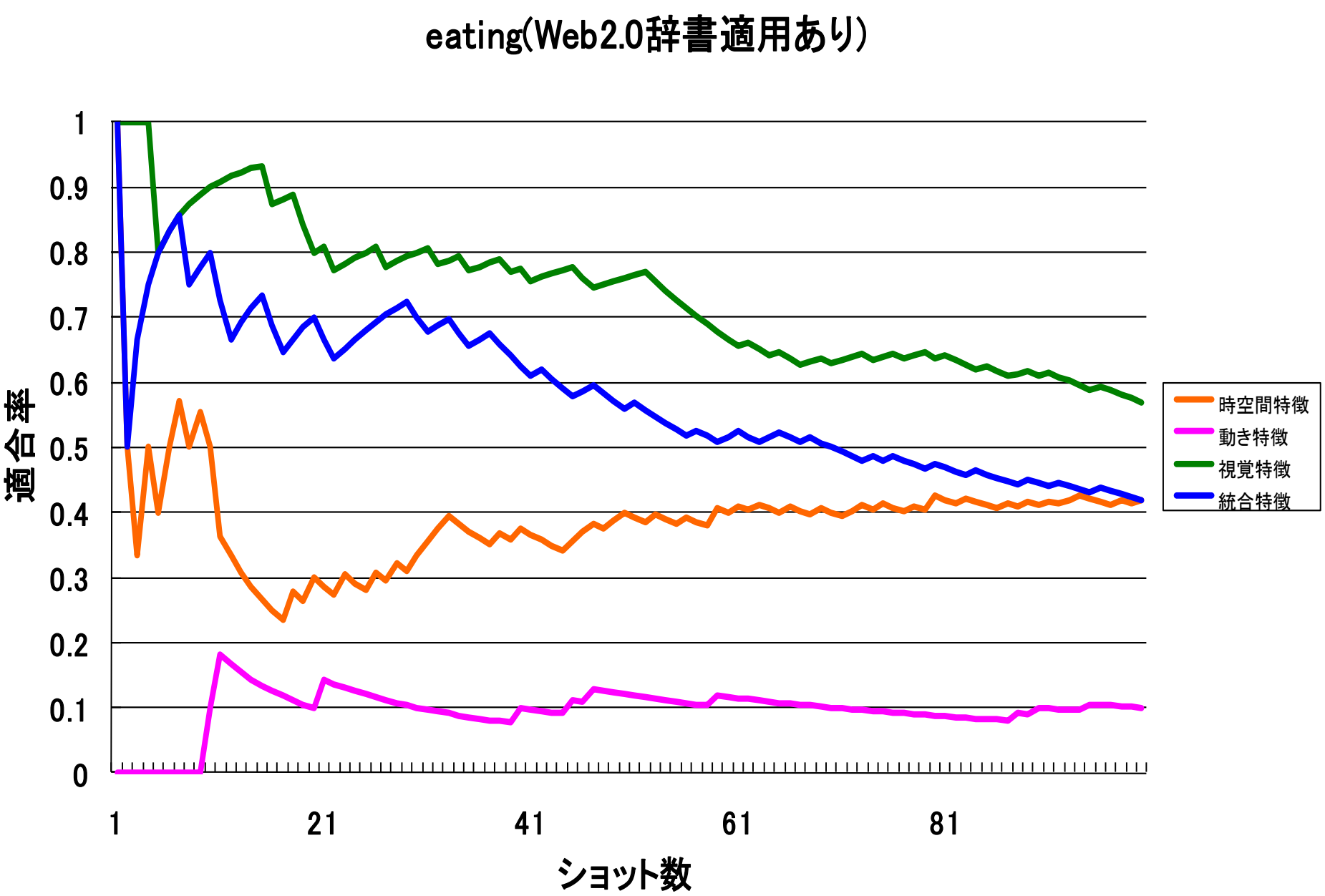
- Web動画を利用して、動作の自動学習フレームワークを提案した
- タグ共起辞書の適用により適合率が約14%改善できた
- 特徴統合によって上位のほうにより多く対応したショットをランキングできた

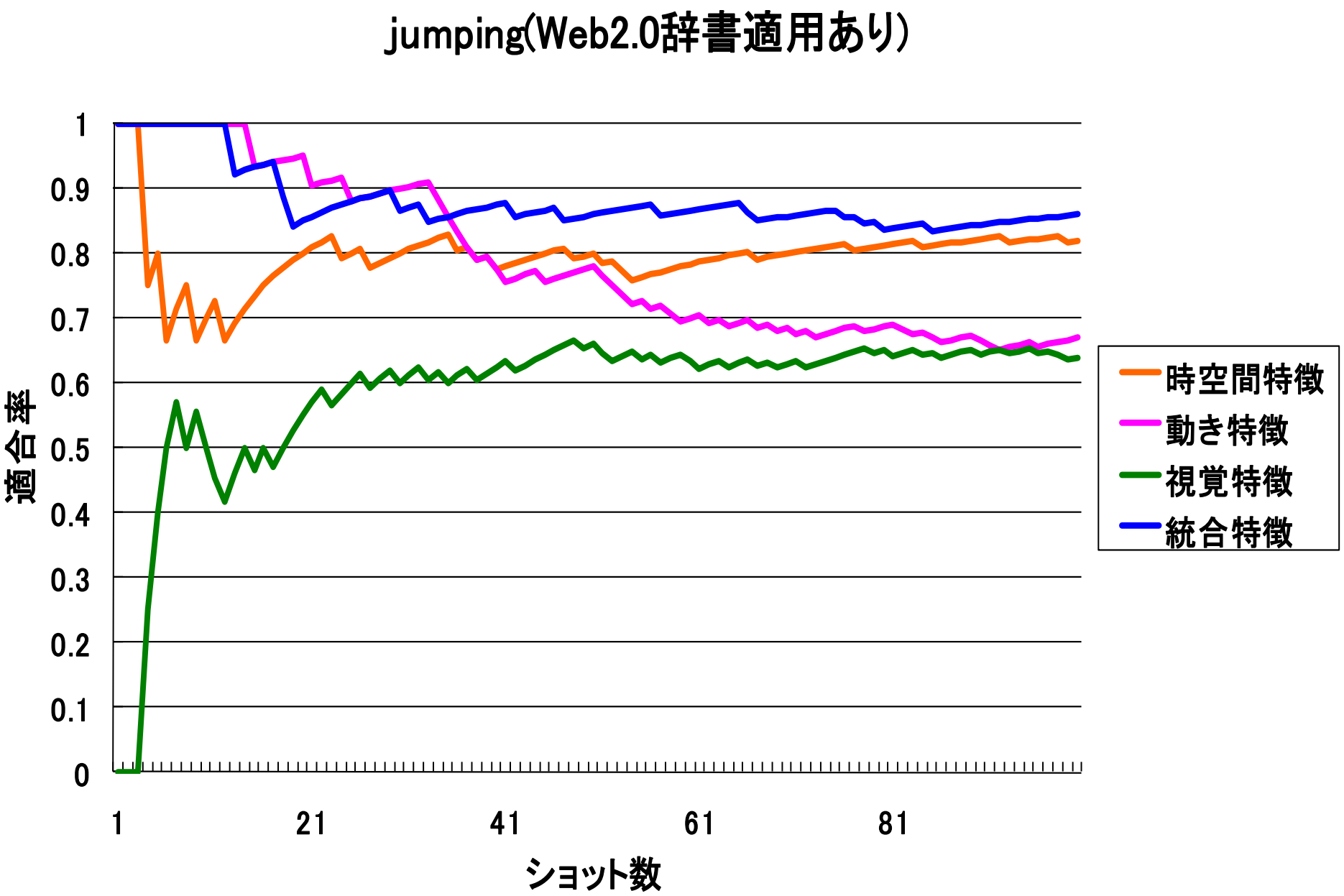
今後の課題

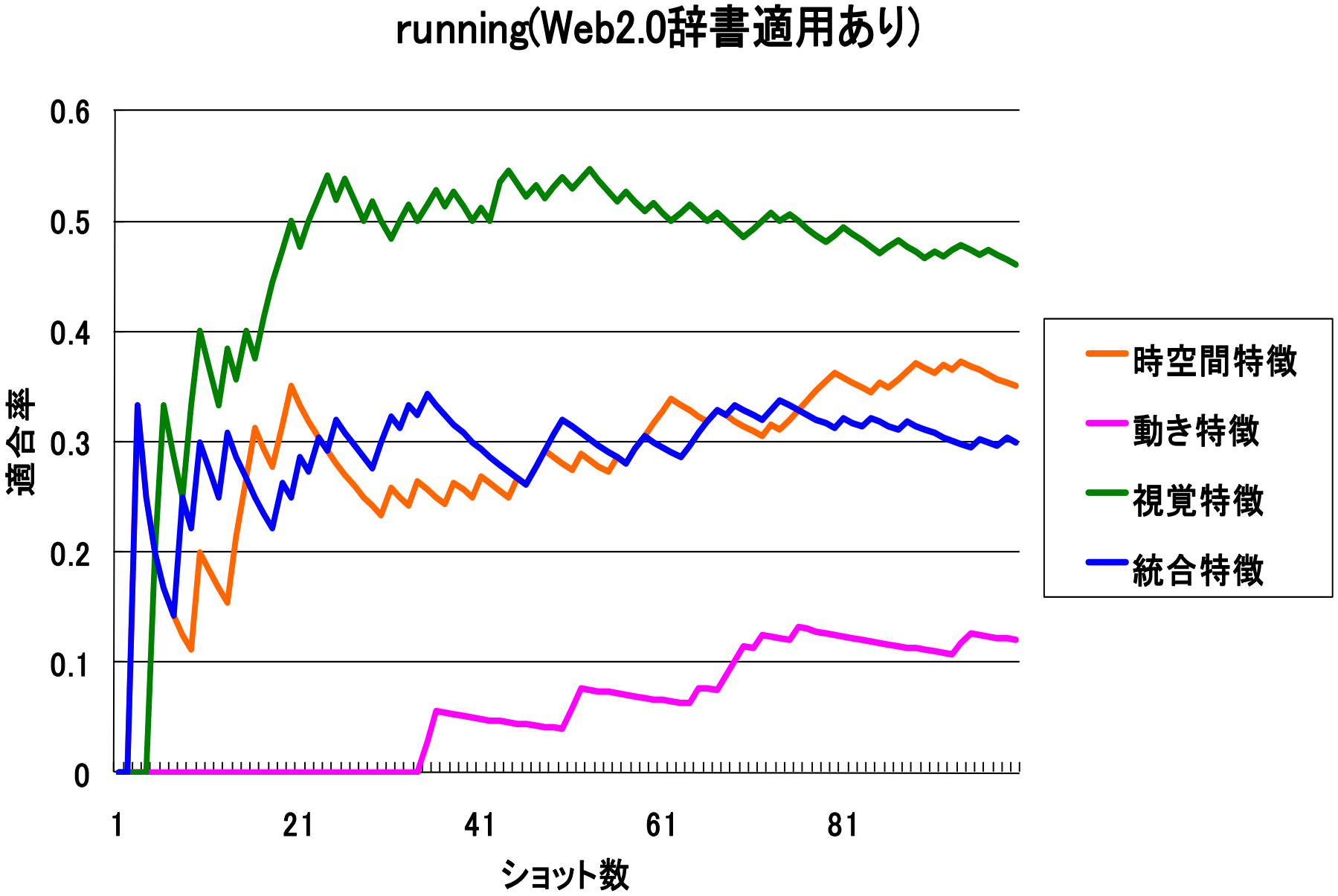
- VisualRankの補正ベクトルの設定の検討
  - －上位ショットの数の検討
  - －スコアや順位に応じたバイアスの値の設定
- 大規模な実験の実現

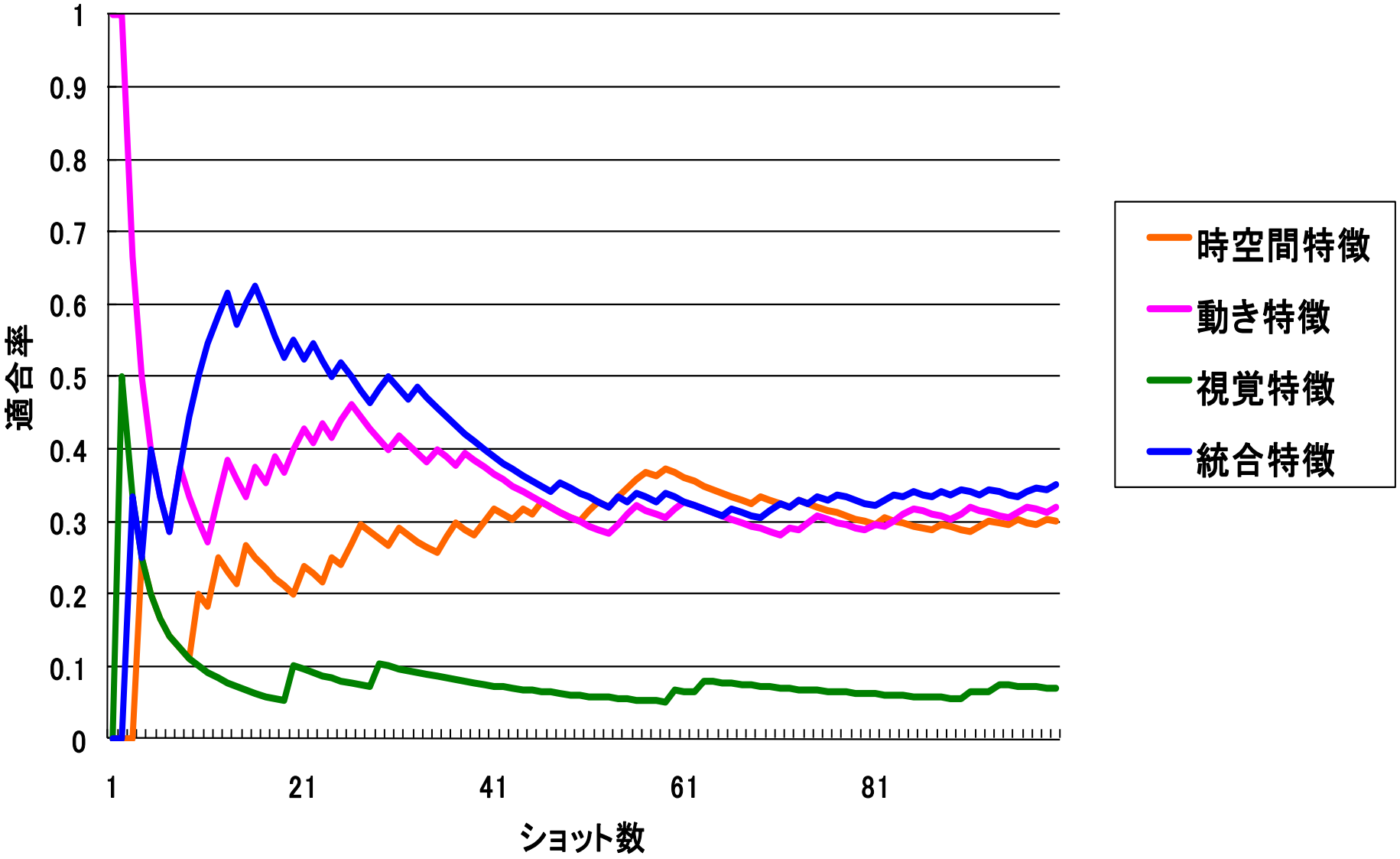
http://img.cs.uec.ac.jp/dohang/vrank_result/flv100/index.html

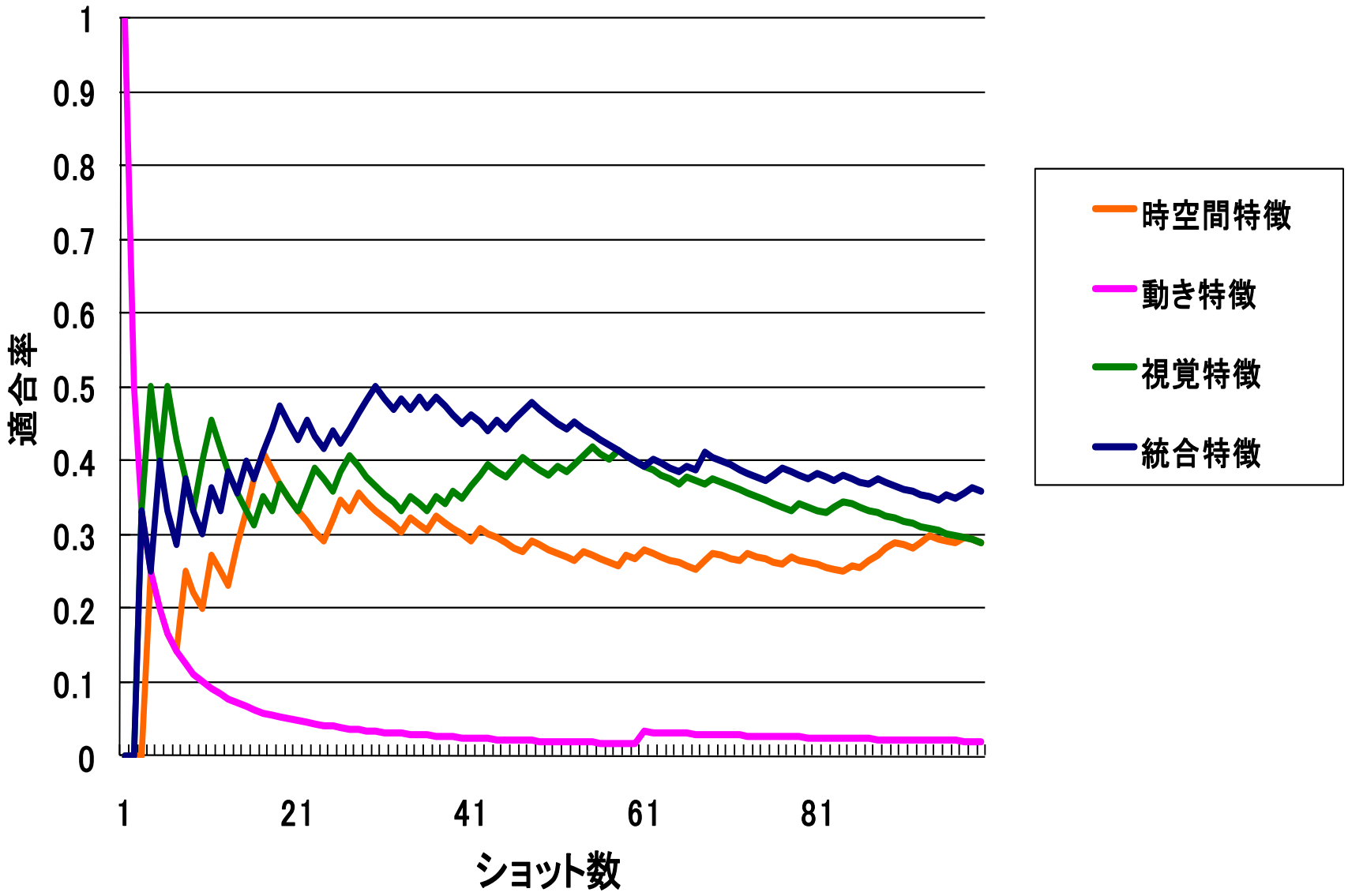

batting(Web2.0辞書適用あり)
適合率
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
21
41
61
81
ショット数
時空間特徴
動き特徴
視覚特徴
統合特徴


eating(Web2.0辞書適用あり)
適合率
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
21
41
61
81
ショット数
時空間特徴
動き特徴
視覚特徴
統合特徴


jumping(Web2.0辞書適用あり)
適合率
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
21
41
61
81
ショット数
時空間特徴
動き特徴
視覚特徴
統合特徴


running(Web2.0辞書適用あり)
適合率
0.6
0.5
0.4
0.3
0.2
0.1
0
1
21
41
61
81
ショット数
時空間特徴
動き特徴
視覚特徴
統合特徴


shooting(Web2.0辞書適用あり)
適合率
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
21
41
61
81
ショット数
時空間特徴
動き特徴
視覚特徴
統合特徴


walking(Web2.0辞書適用あり)
適合率
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
21
41
61
81
ショット数
時空間特徴
動き特徴
視覚特徴
統合特徴