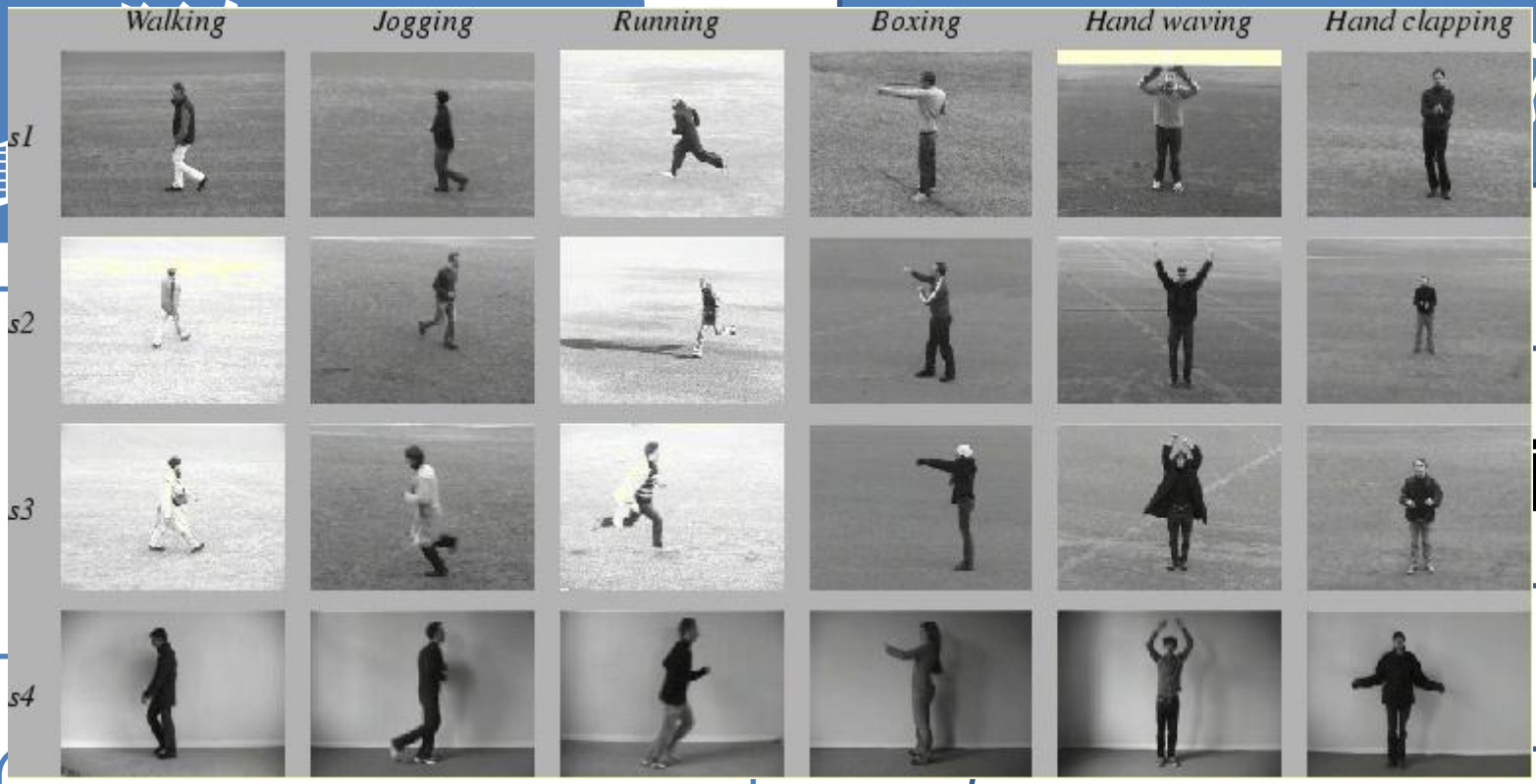
Web動画・画像を用いた
特定動作ショットの自動収集

DO HANG NGA　○樋爪 和也　　柳井 啓司
電気通信大学　　情報工学科

背景



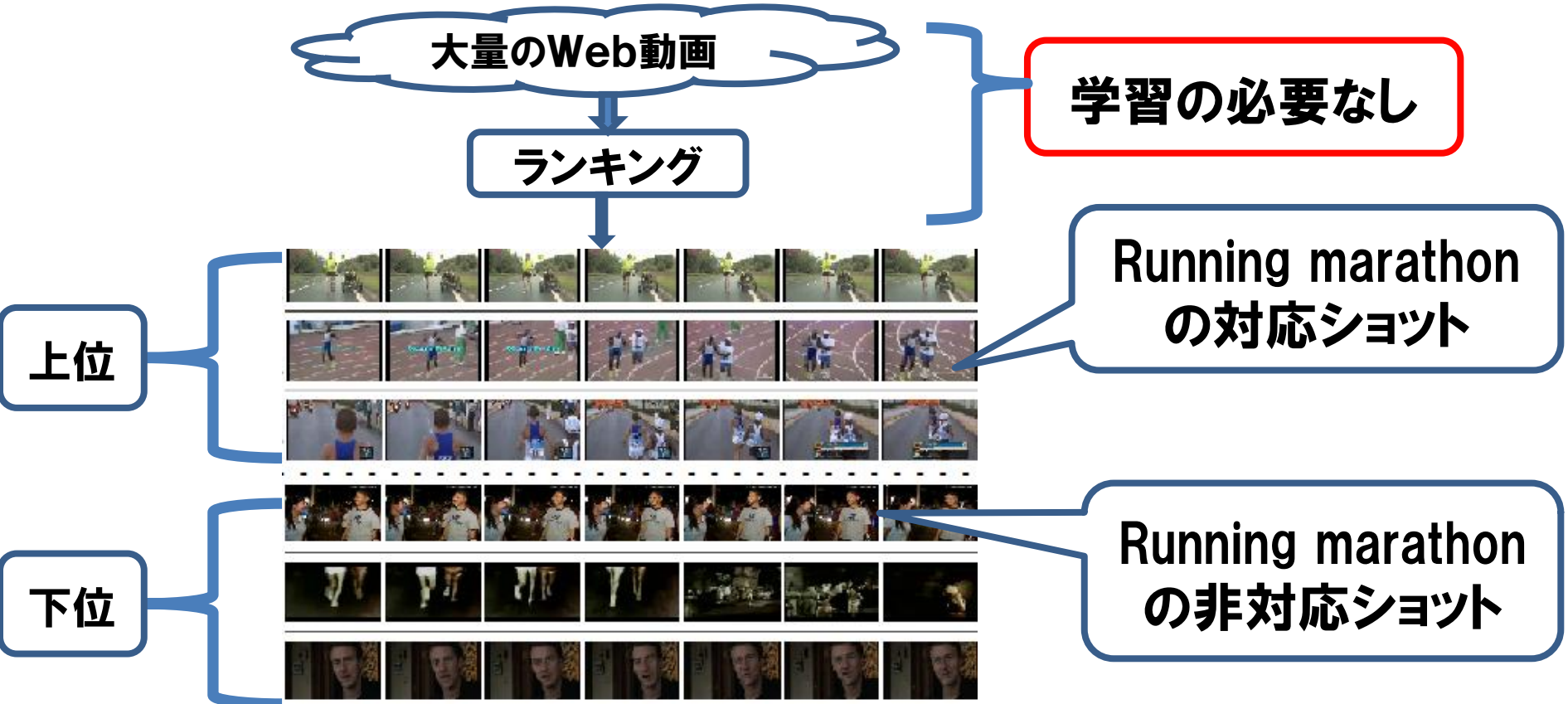
教師信号あり

動画量が少ない

研究の目的

特定動作についてのWebデータを使用して、

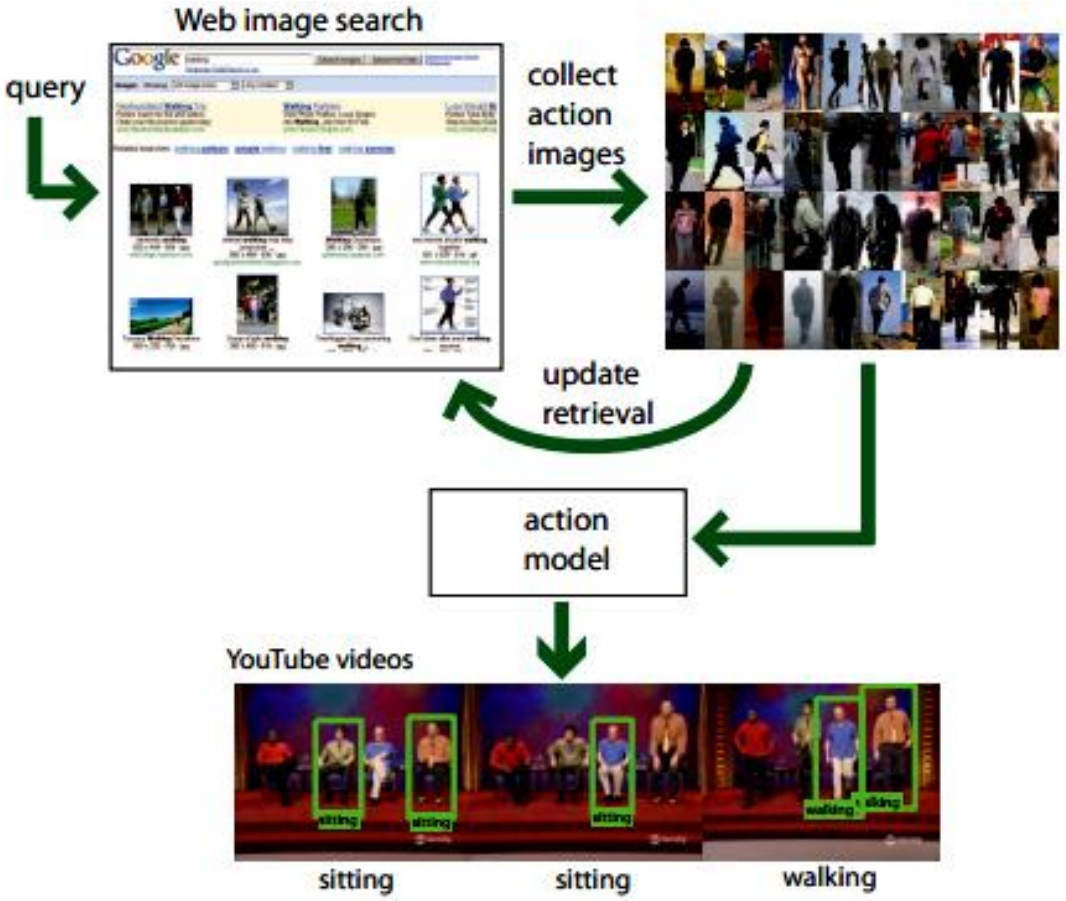
その動作の対応ショットを自動抽出



関連研究

N. I. Cinbis, R. G. Cinbis and S. Sclaroff:

“Learning actions from the web”, ICCV2009



Cinbisらの研究

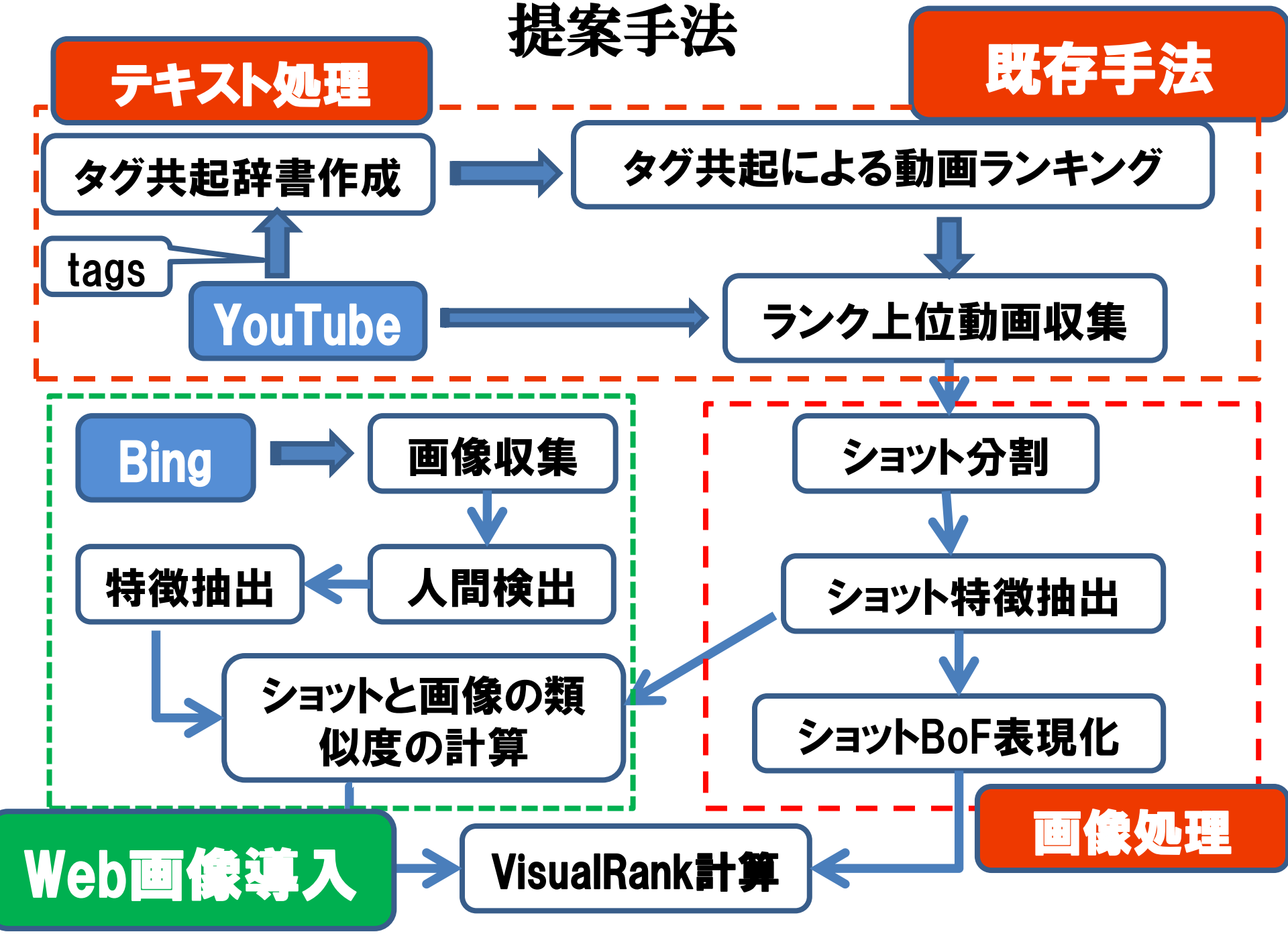
Web画像　静的特徴

我々の研究

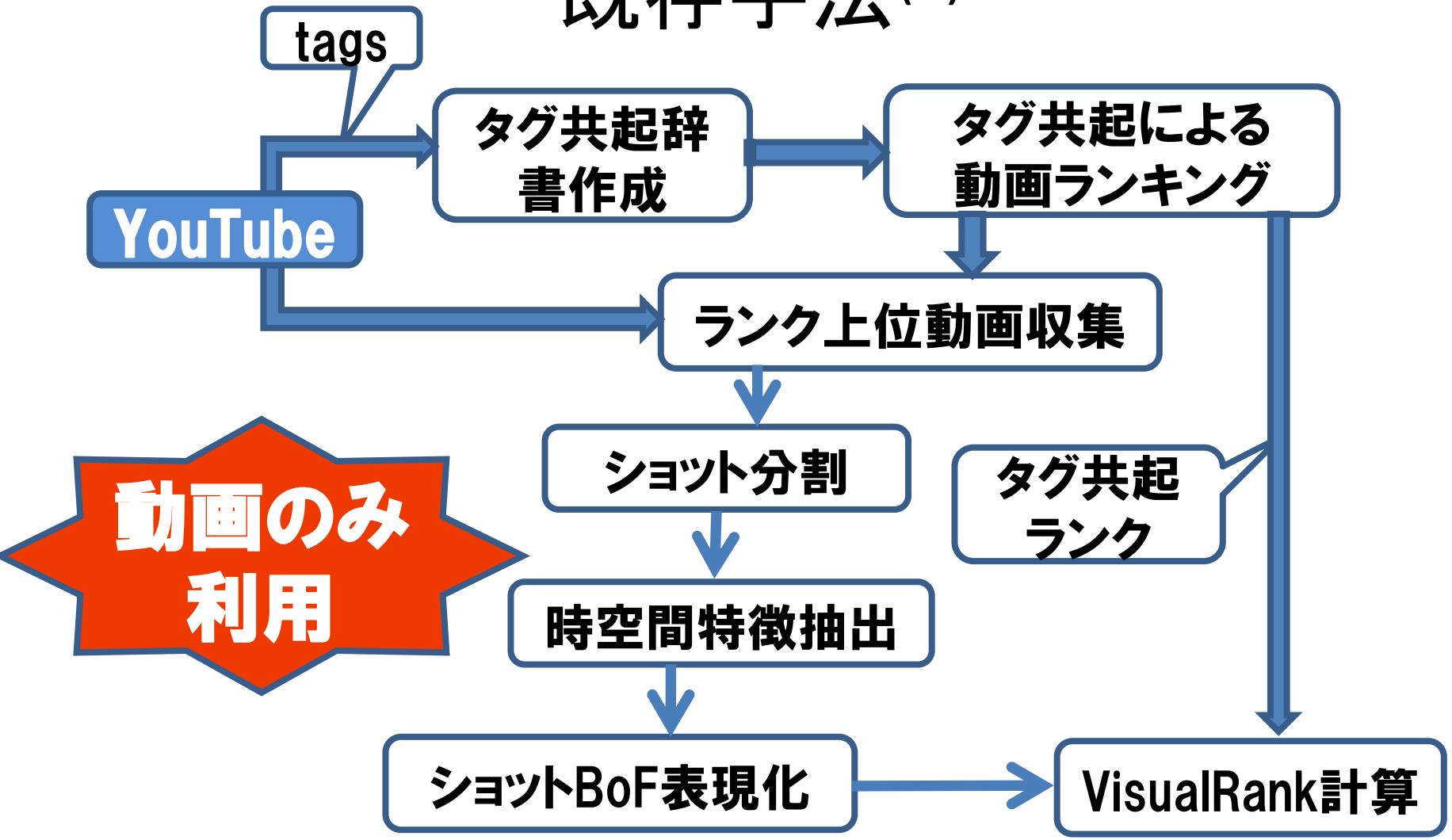
Web動画＋Web画像

時空間特徴

提案手法


テキスト処理
既存手法
タグ共起辞書作成
タグ共起による動画ランキング
tags
YouTube
ランク上位動画収集
Bing
画像収集
特徴抽出
人間検出
ショット分割
ショット特徴抽出
ショットと画像の類似度の計算
ショットBoF表現化
Web画像導入
VisualRank計算
画像処理

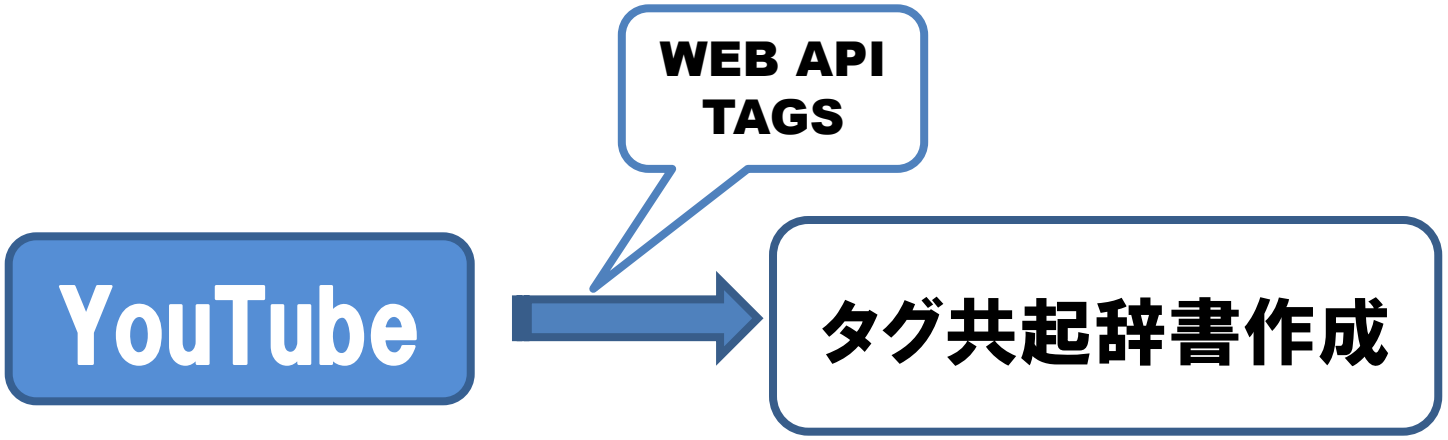
既存手法(*)



(*)・DoHang Nga, 柳井啓司: 大量のWeb動画からの教師なし特定動作ショット抽出, MIRU2011

・H.N.Do, K.Yanai: Automatic Construction of an Action Database using Web Videos, ICCV2011

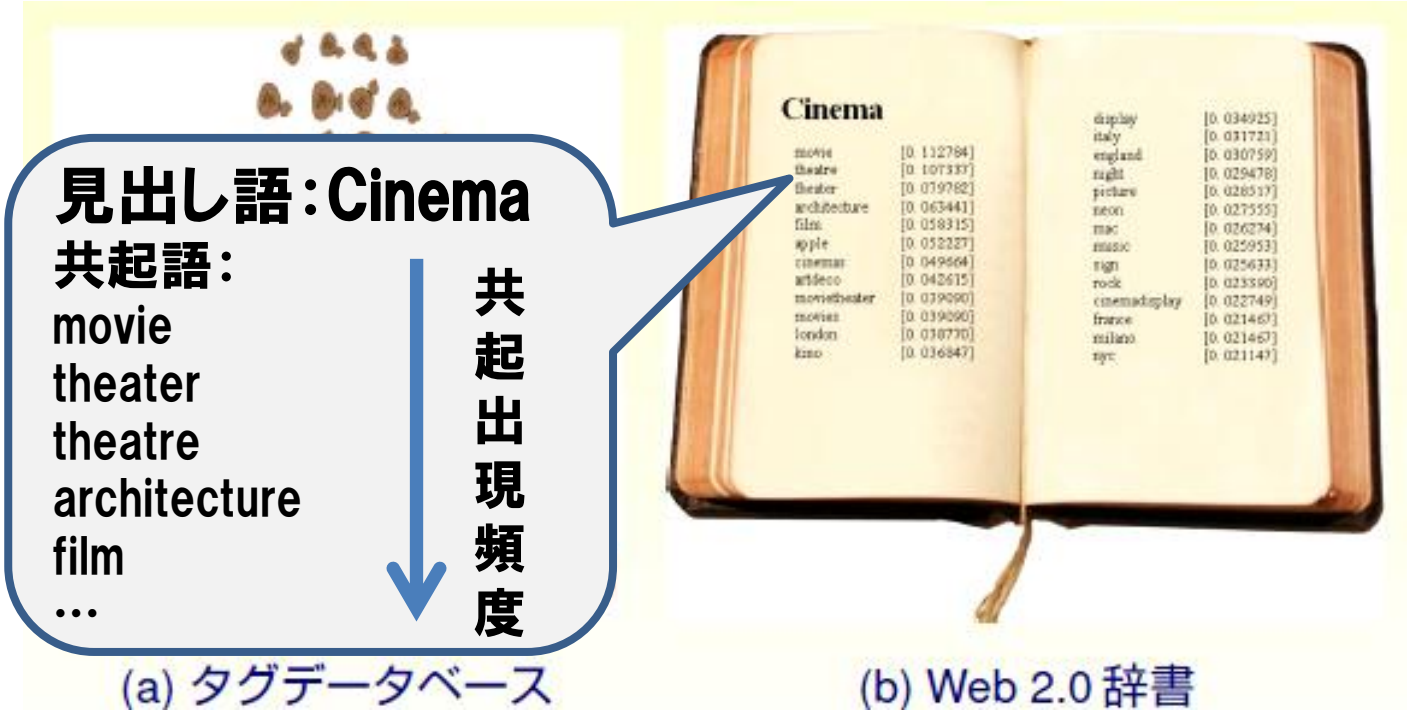
既存手法:タグ共起辞書作成ステップ



タグ共起辞書

タグ共起辞書： Web2.0辞書(*)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



(*)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

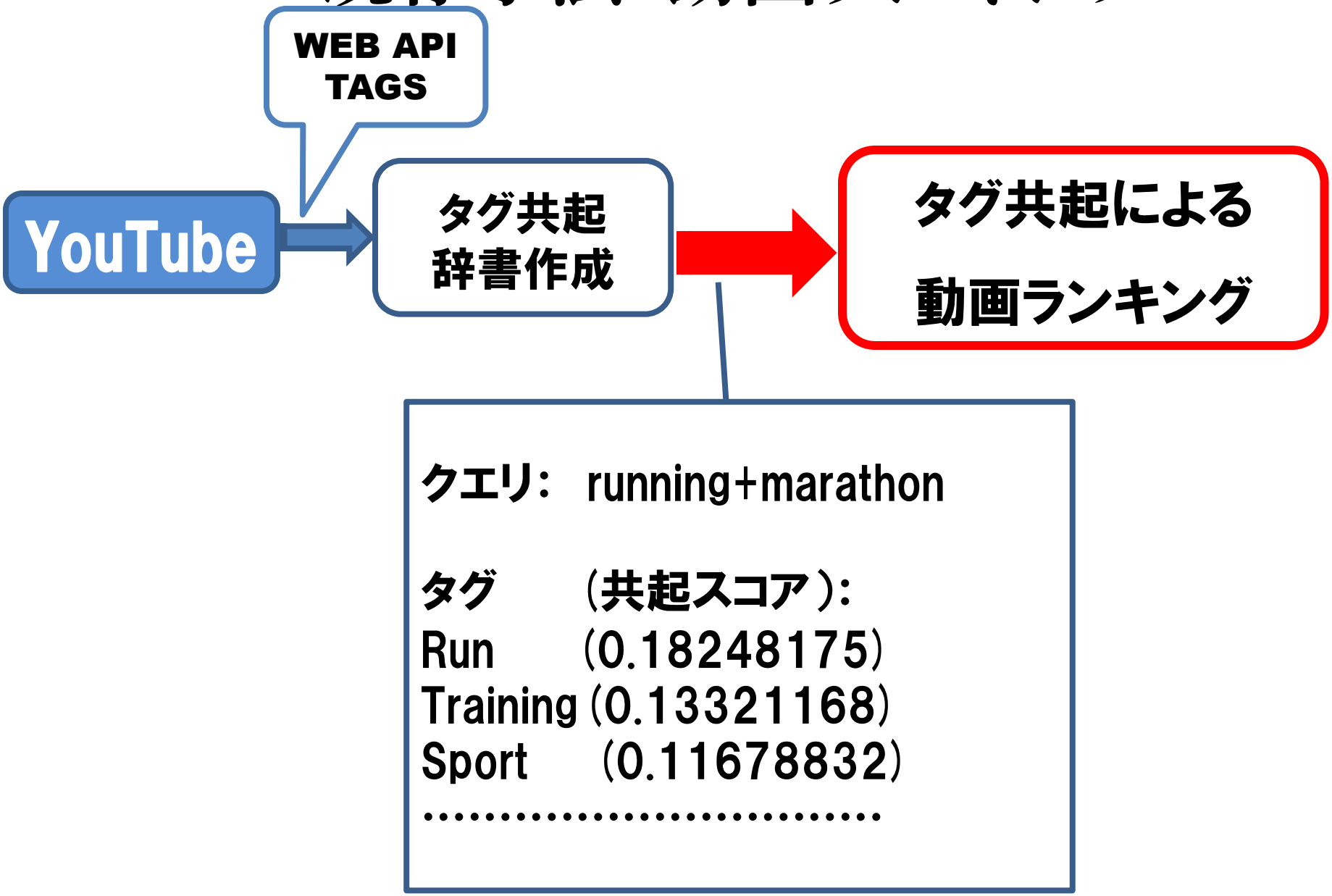
ステップ2 タグを集計し、出現頻度上位2000タグについて、それぞれ1000動画のタグを収集

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

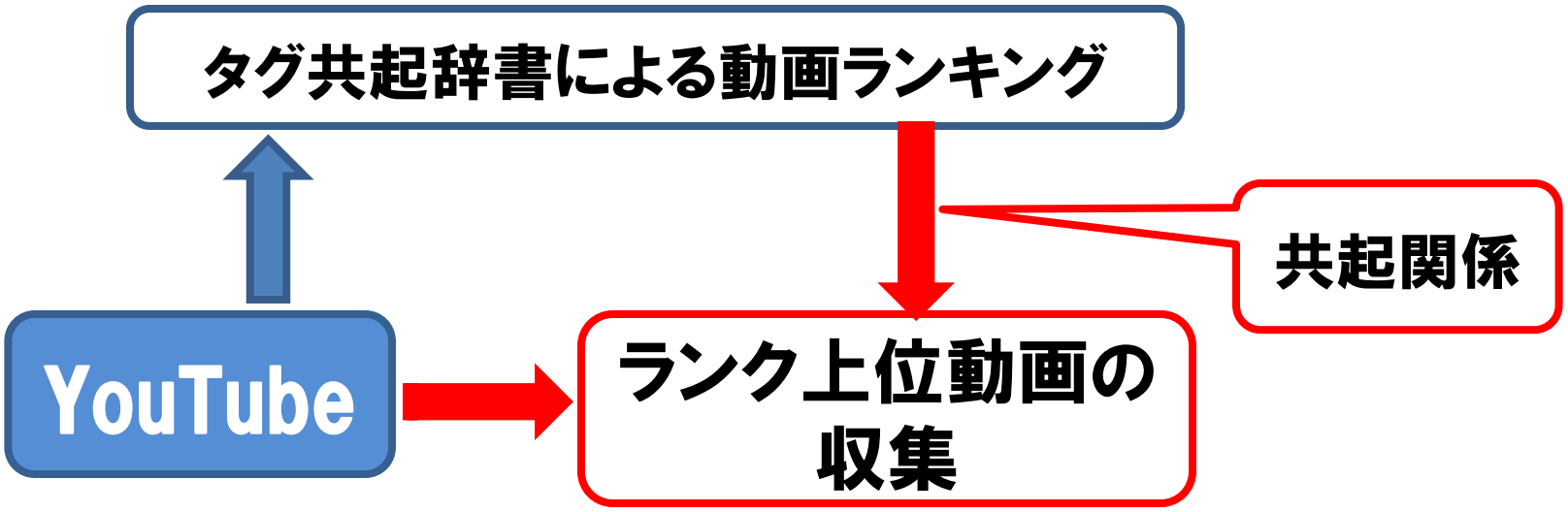
$$P(b \mid a) = \frac{count(a,b)}{count(a)}$$

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

既存手法：動画ランキング

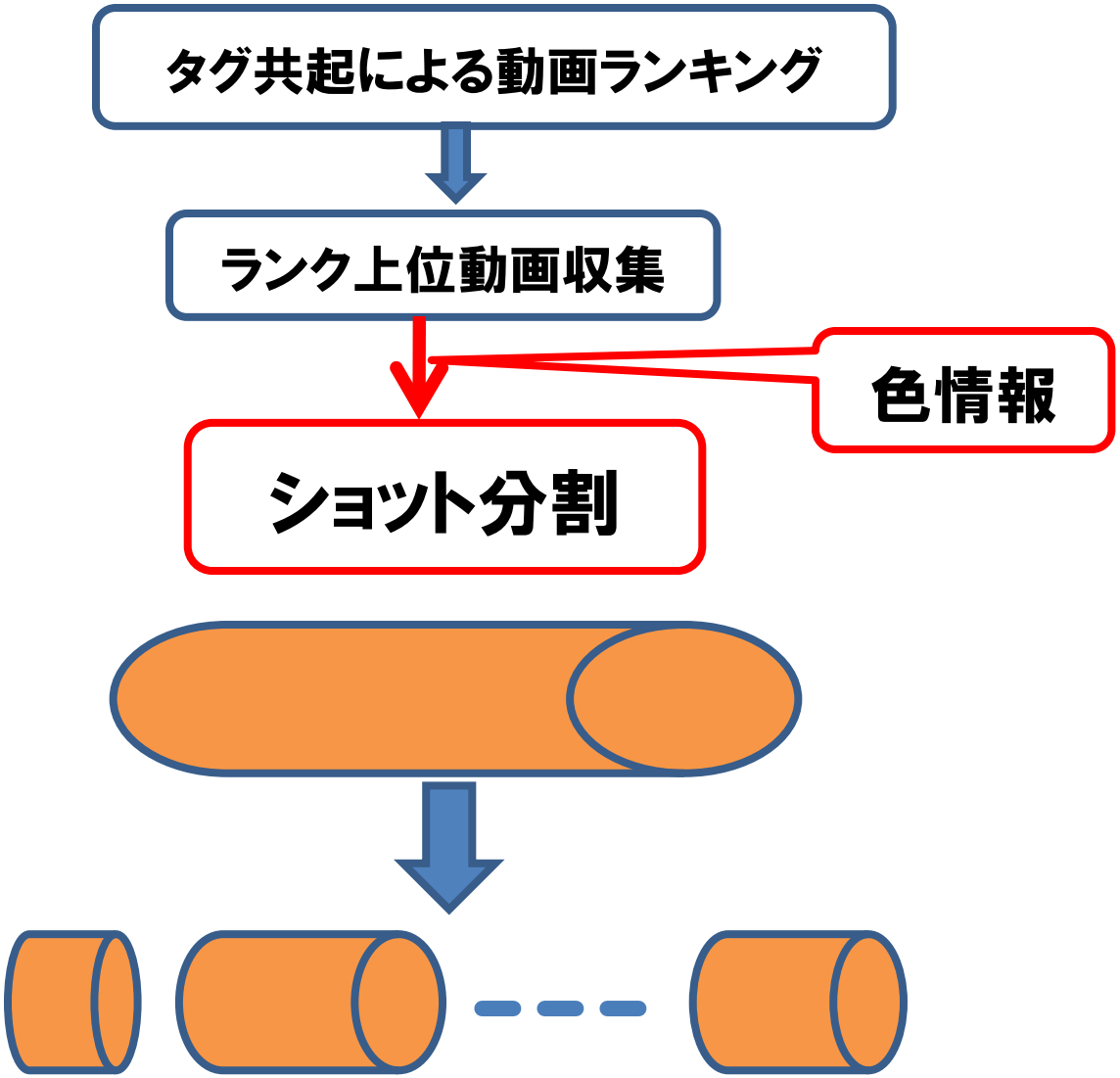


既存手法：動画収集ステップ

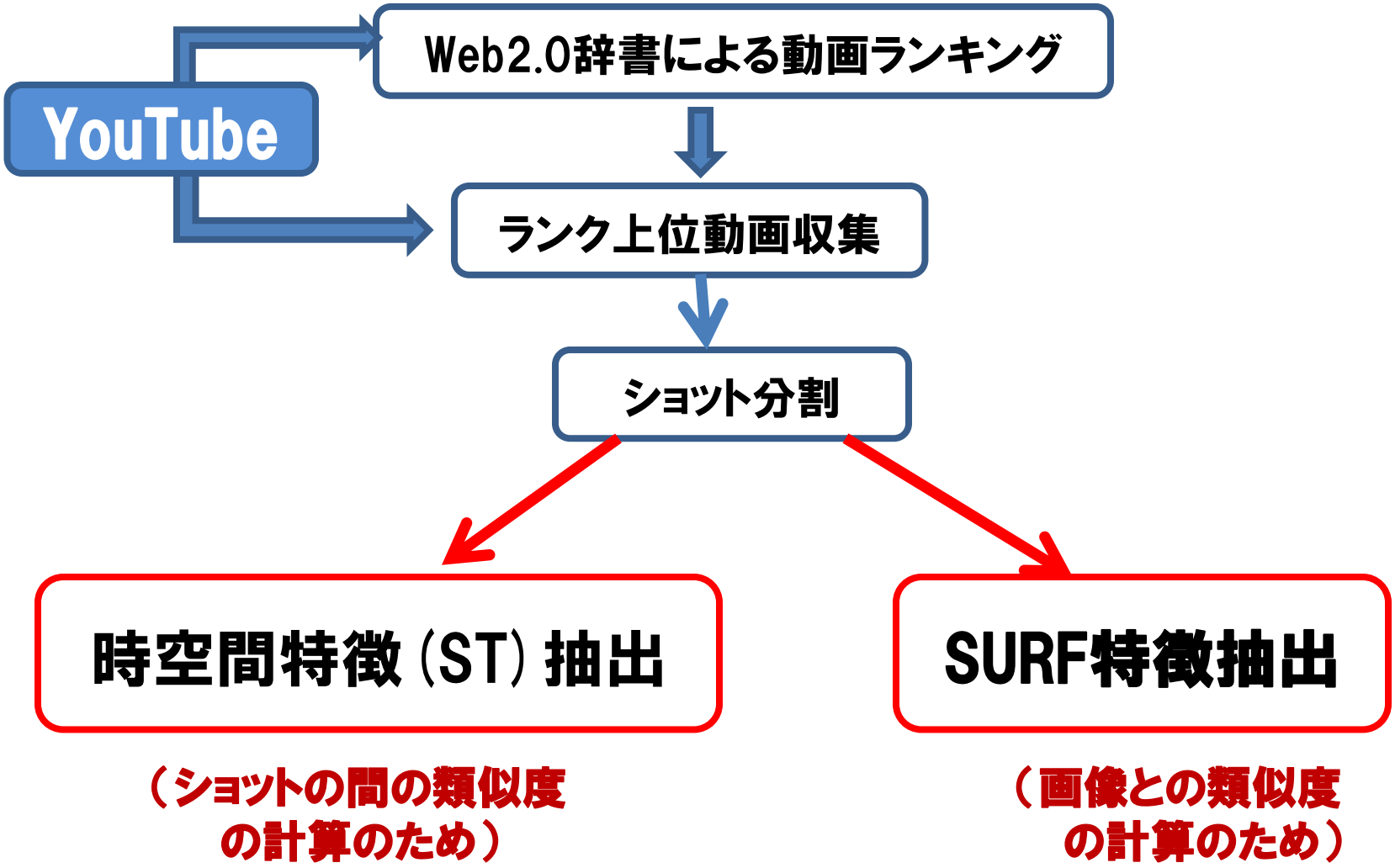


※上位200動画のみ

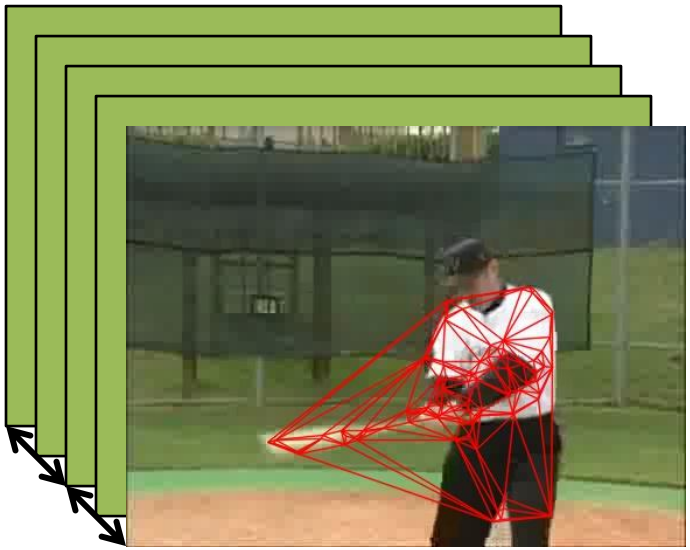
既存手法：ショット分割ステップ



特徴抽出ステップ


YouTube
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
SURF特徴抽出
（ショットの間の類似度の計算のため）
（画像との類似度の計算のため）

時空間特徴(*)



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

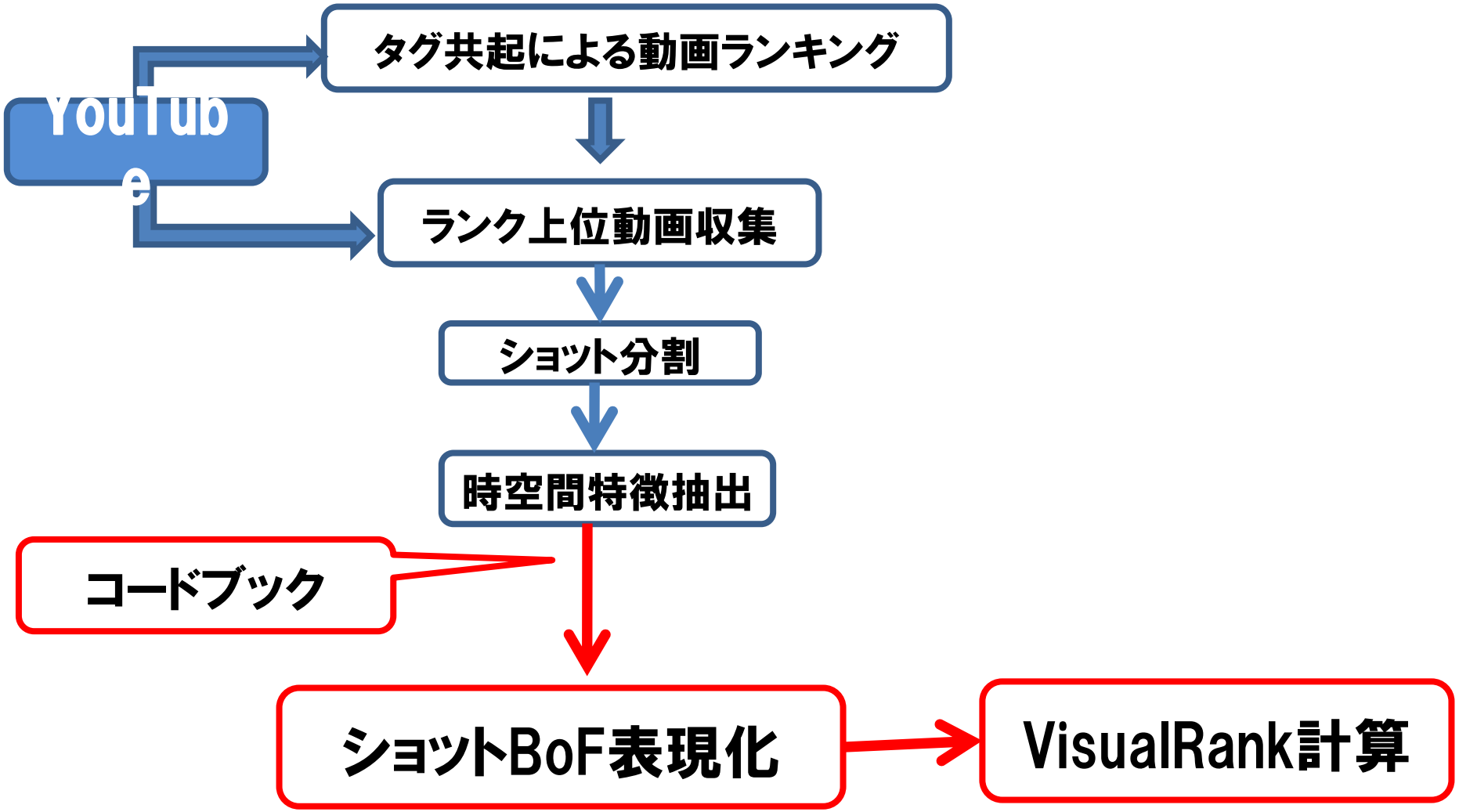
④. ドロネー三角形を作成
以降三点で一組の特徴と考える

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

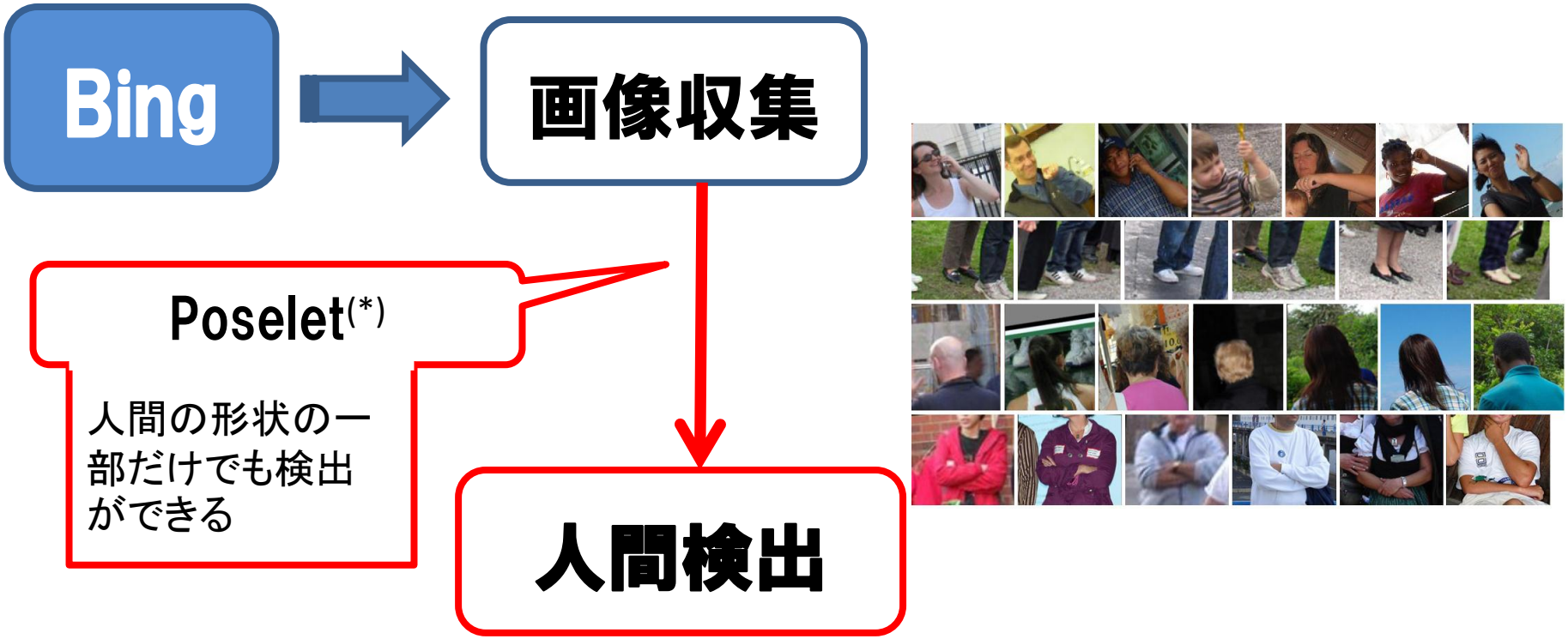
⑥ 視覚特徴と動き特徴を統合し、
特徴をヒストグラム化する

(*) A.Noguchi and K.Yanai: A SURF-based Spatio-Temporal Feature for feature-fusion-based action recognition, ECCV WS on Human Motion: Understanding, Modeling, Capture and Animation

既存手法：ショットBoF 化ステップ
VisualRank計算ステップ



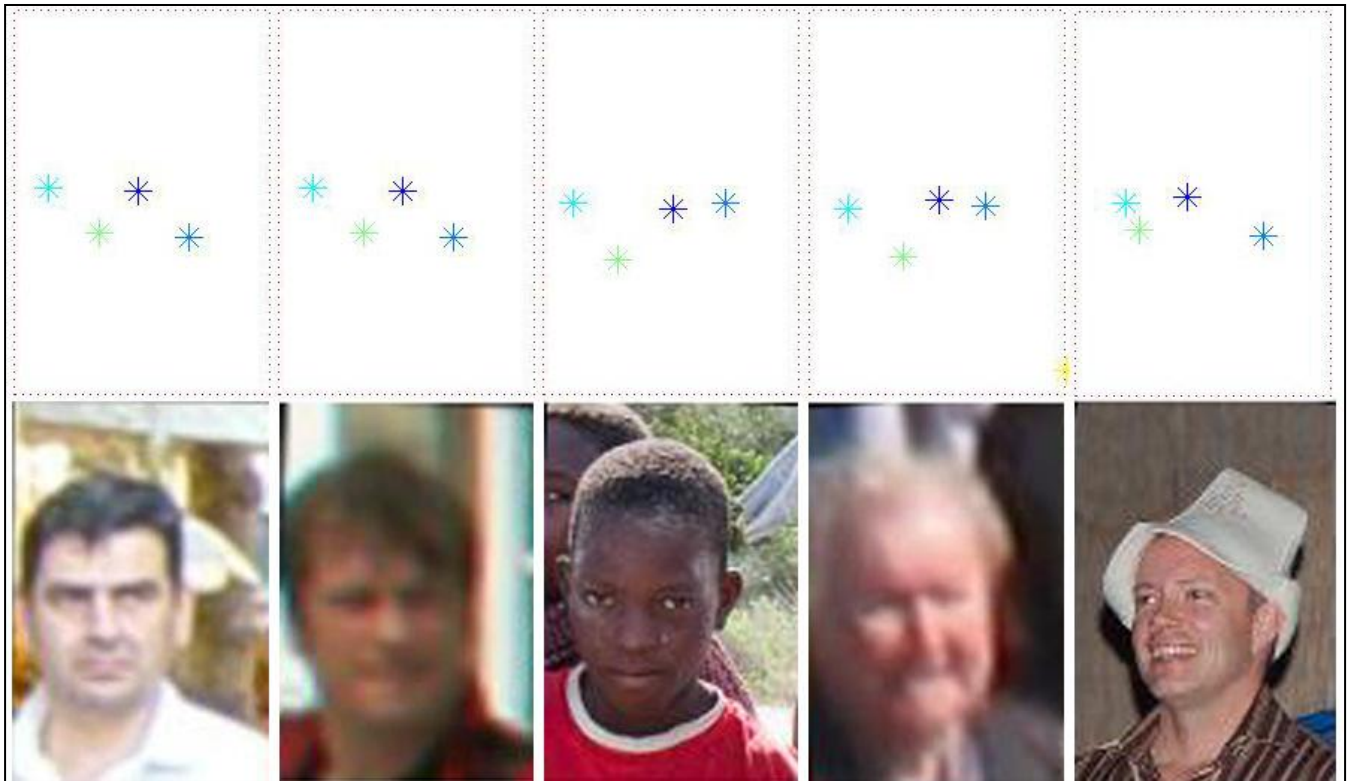
提案手法：Web画像の導入
Web画像収集ステップ



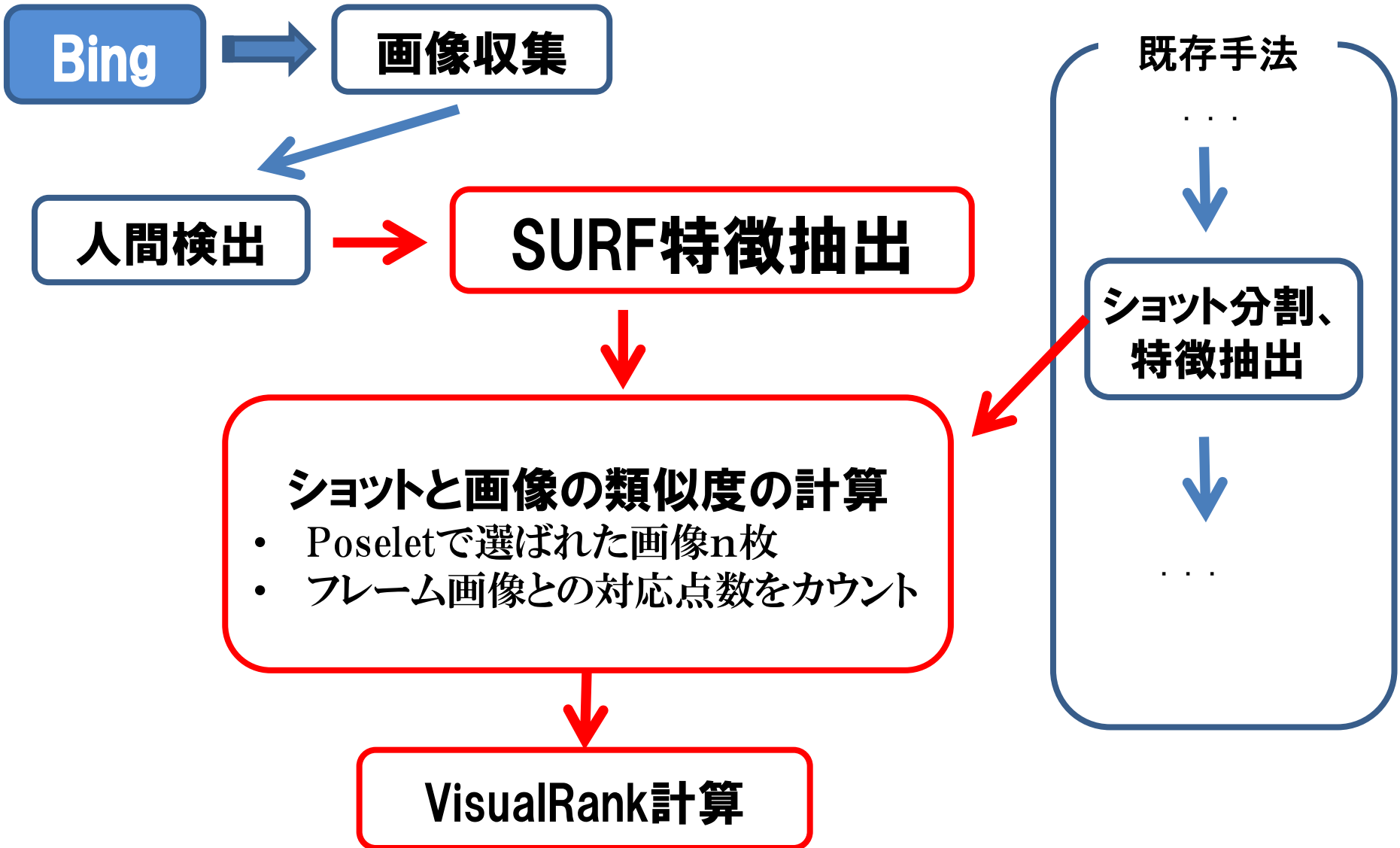
(*) Lubomir Bourdev, Jitendra Malik, Poselets: Body Parts Detectors Trained using 3D Human Pose Annotations, ICCV 2009

Poselet

- 3D特徴点を使用し、パーツまたはポーズ毎に人間検出を行う
- PASCAL VOC 2007 challenge・personカテゴリでAP=0.365



提案手法:特徴抽出ステップ
類似度計算ステップ



ショットのVisualRankの計算

- VisualRank(*)計算：

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

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

- 補正ベクトル（バイアスなし）：

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

補正ベクトルの設定

- 既存手法：共起スコアの高いショットにバイアス

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

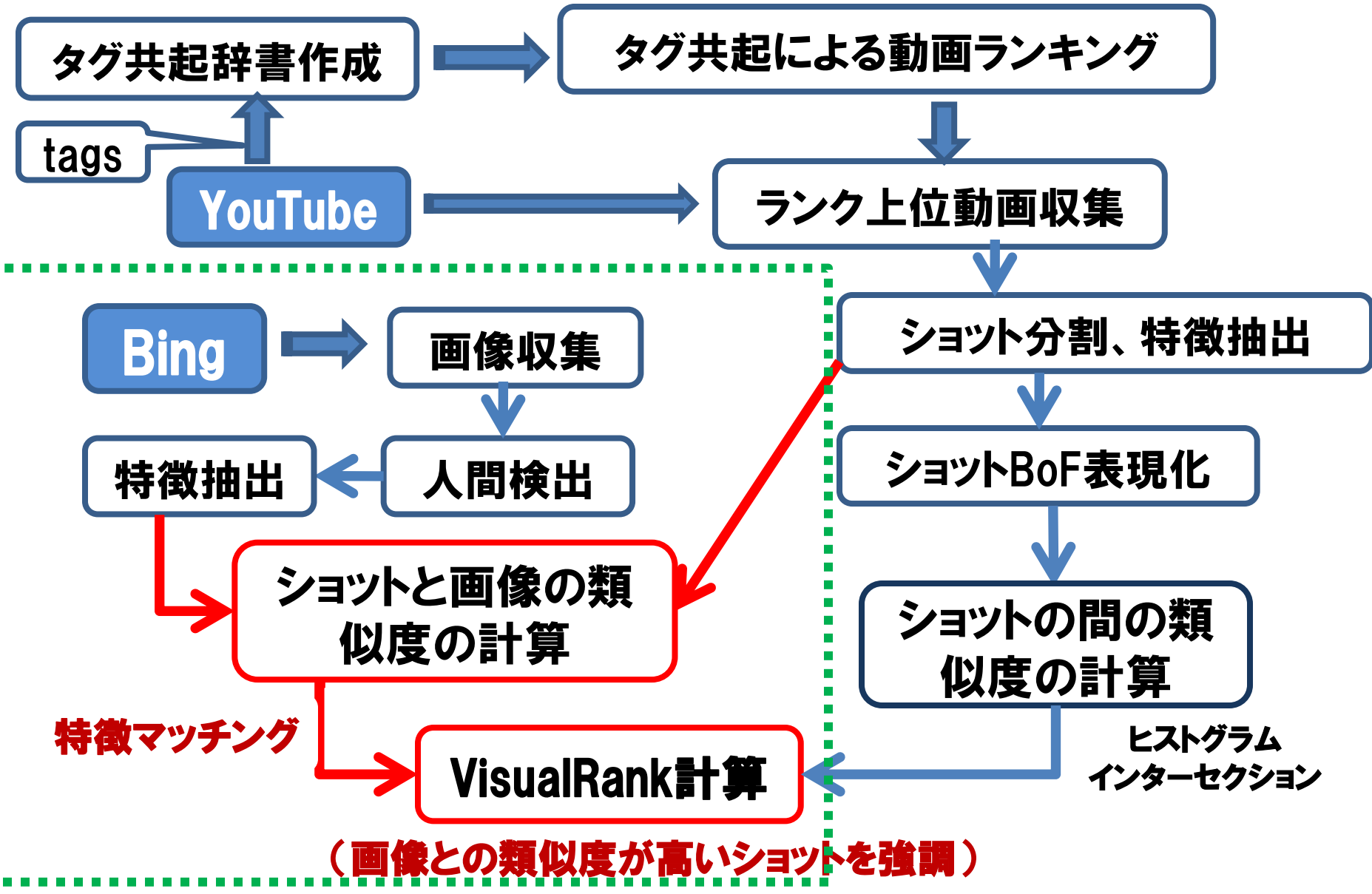
実験設定：
n ≈ 2000, m = 1000

- **提案手法：Poseletで選択された画像との類似度が高いショットにバイアス**

$$p_i = \frac{\exp(\gamma\, \mathrm{S(i)})}{\sum_{j=1}^{n} \exp(\gamma\, \mathrm{S}(j))} \quad \gamma = log2 : (\text{定数値})$$

S(i)：ショットの類似度

提案手法


タグ共起辞書作成
タグ共起による動画ランキング
tags
YouTube
ランク上位動画収集
Bing
画像収集
ショット分割、特徴抽出
特徴抽出
人間検出
ショットBoF表現化
ショットと画像の類似度の計算
ショットの間の類似度の計算
特徴マッチング
VisualRank計算
ヒストグラムインターセクション
（画像との類似度が高いショットを強調）

実験

- 目的: Web画像導入の有効性の検討

- 既存手法のデータおよび結果評価法を利用
  ーランキング後の上位1～100ショットについての
  適合率で結果を評価

実験

- 実験1：既存手法で適合率が50%以下の6種類の動作を選び、実験を行う

- 実験2：精度が良い種類に対する提案手法の有効性の検討
  ― 既存手法で適合率が50%以上の4種類

- 実験3：Poseletで選択する画像数の影響の検討
  ― 既存手法で適合率が10%以下の6種類

実験データ1

テーブル1：適合率が50%以下の6種類

動作	動画数	利用ショット数
bake+bread	198	2000
brush+teeth	173	1652
iron+clothes	181	1944
jog	169	2000
jump+rope	162	1675
wash+face	173	1277
平均	176	1758

Web画像適用の有効性の検討(その1)

動作	既存手法	手動で選択 20画像	Poselet適用あり TOP20画像	Poselet適用なし TOP20画像
bake+bread	6	16	19	12
brush+teeth	28	38	33	27
iron+clothes	47	48	47	49
jog	5	21	19	14
jump+rope	26	24	34	30
wash+face	29	30	29	24
平均	23.5	29.5	30.2	26.0

実験データ２

適合率が50%以上の4種類

動作	動画数	利用ショット数
curl+bicep	165	832
do+yoga	151	1641
ride+bicycle	197	2000
laugh	196	2000
平均	120	1412

実験2の結果
Web画像適用の有効性の検討(その2)

動作	既存手法	提案手法
curl+bicep	58	42
do+yoga	77	40
ride+bicycle	62	55
laugh	50	15
平均	61.8	38.3

23.5%↓

実験3のデータ

適合率が10%以下の6種類

動作	動画数	利用ショット数
boil+egg	187	2000
head+ball	183	1973
cook+rice	190	2000
grill+fish	191	2000
swim+butterfly	193	2000
swim+backstroke	177	1777
平均	187	1958

実験3
Poseletで選択する画像数の影響の検討

動作	既存手法	10画像	20画像	30画像	50画像
boil+egg	9	10	**13**	7	6
head+ball	9	7	**10**	6	6
cook+rice	6	15	**16**	15	13
grill+fish	5	21	23	**27**	17
swim+butterfly	7	29	33	30	**37**
swim+backstroke	9	10	11	**13**	12
平均	7.5	15.3	17.7	16.3	15.2

結論

- Web動画からの自動ショット抽出において，Web画像を導入した。
  - 低い精度の動作に関して，精度が向上。
  - ただし，元の精度が高い場合，精度低下。

今後の課題

- Web画像の選択の仕方の改良
  - Poselet以外の人物検出手法の利用
  - 動作対象物体の認識

- 画像とショットの類似度の計算法の改良
  - 多数画像(Web画像) 対 多数画像(フレーム) の新しい類似度計算手法の考案
  - BoFや色などの特徴の利用

データセット公開

http://mm.cs.uec.ac.jp/webvideo/video.html

Exp No.	Tag-based Ranking	Biased damp. vec.	Visual Feature	Mean prec@100
RND	Randomly-selected 100 shots			14.2%
TAG	✔	—	—	23.5%
1	—	—	ST	33.7%
2	✔	—	ST	41.0%
3(1)	✔	✔(1)	ST	**47.3%**
3(2)	✔	✔(2)	ST	44.8%
5	✔	✔(1)	Motion	31.8%
6	✔	✔(1)	Appear.	39.7%
7	✔	✔(1)	Fusion	**49.5%**

(1) $p_i = \begin{cases} \frac{1}{m} & 1 \le i \le m \\ 0 & m < i \le n \end{cases}$, $n \approx 2000, m = 1000$

(2) $p_i = \frac{S_c(j)}{C}$, $C = \sum_{j=1}^{n} S_c(j)$

Sc(j)：ショットjのビデオのタグ共起スコア