テレビ映像からの特定動作シーンの自動検出

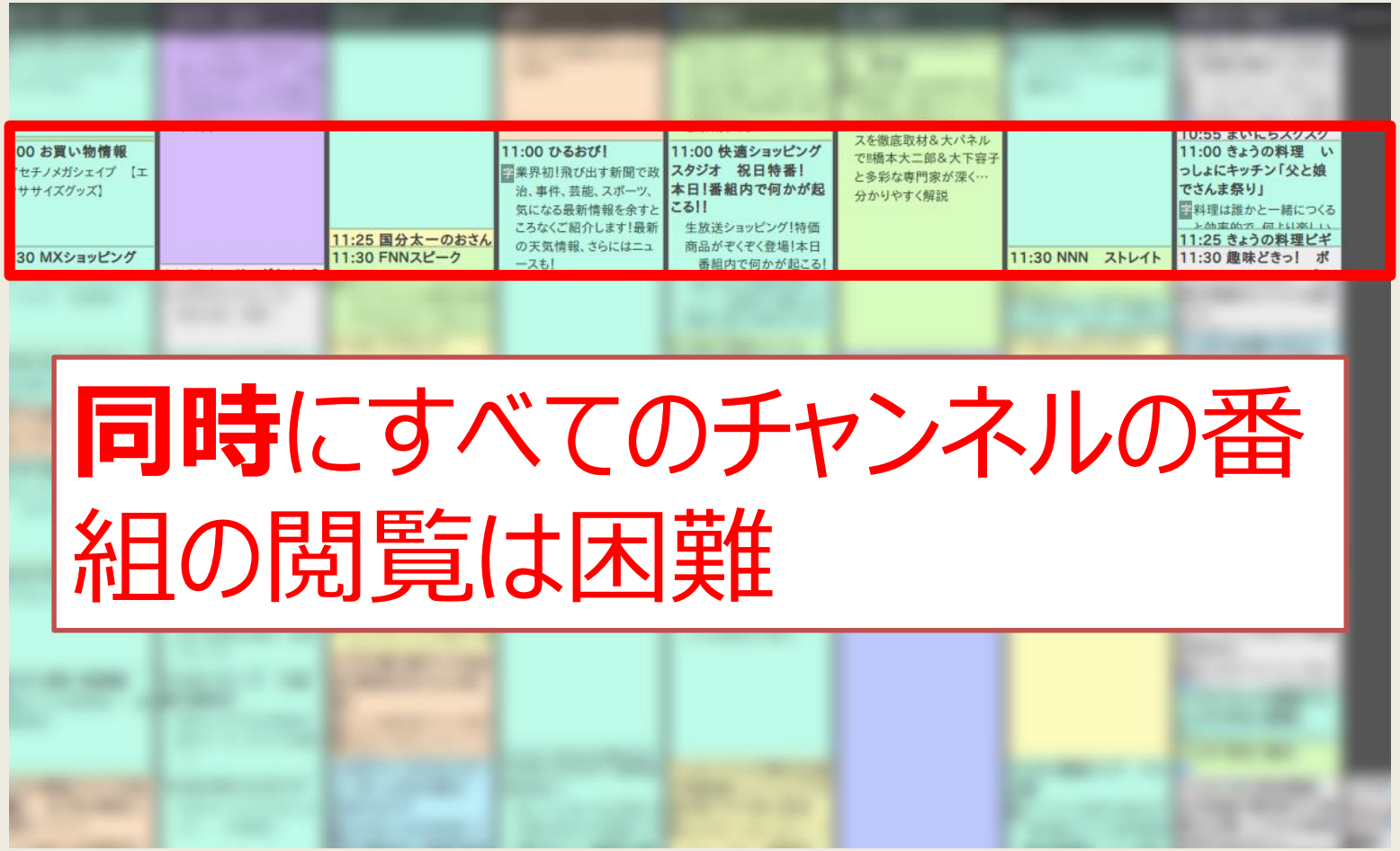
電気通信大学　情報理工学部
総合情報学科　柳井研究室
小林　隼人 柳井 啓司

はじめに

テレビのデジタル化による**多チャンネル化**

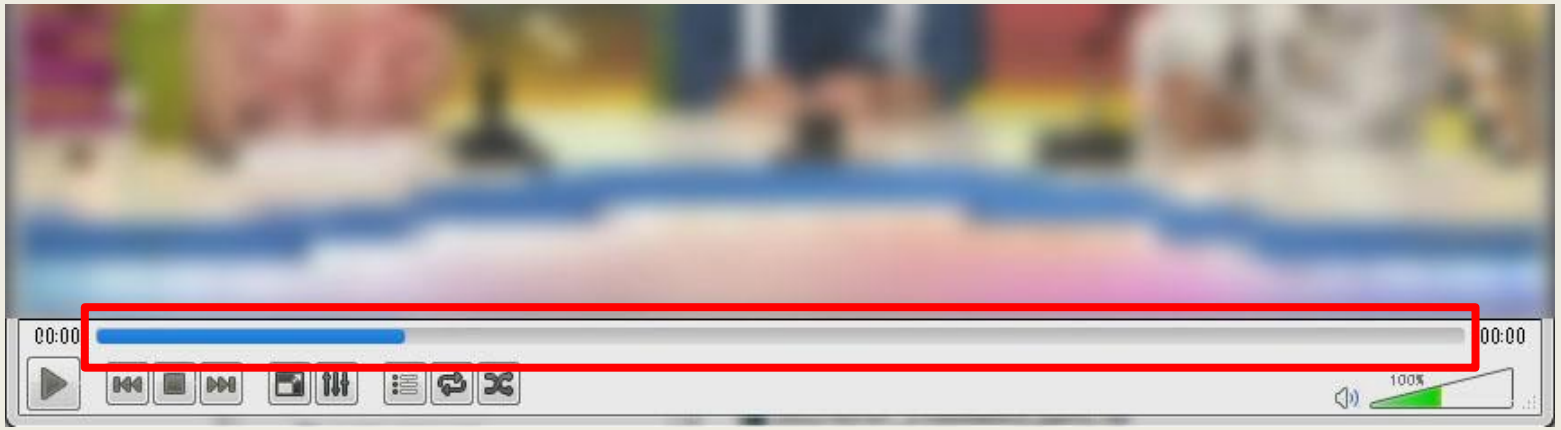
はじめに

テレビのデジタル化による**多チャンネル化**



はじめに

録画しても



見たいシーンのみを探すのも困難

研究目的

テレビ映像から
特定の動作を
自動的に検出

見たいシーンを
すぐに閲覧可能

例）食事シーン

動作認識の手法

1. **動き**を元に認識する方法

2. **物体やシーン認識**を使って認識する方法

動作認識の手法

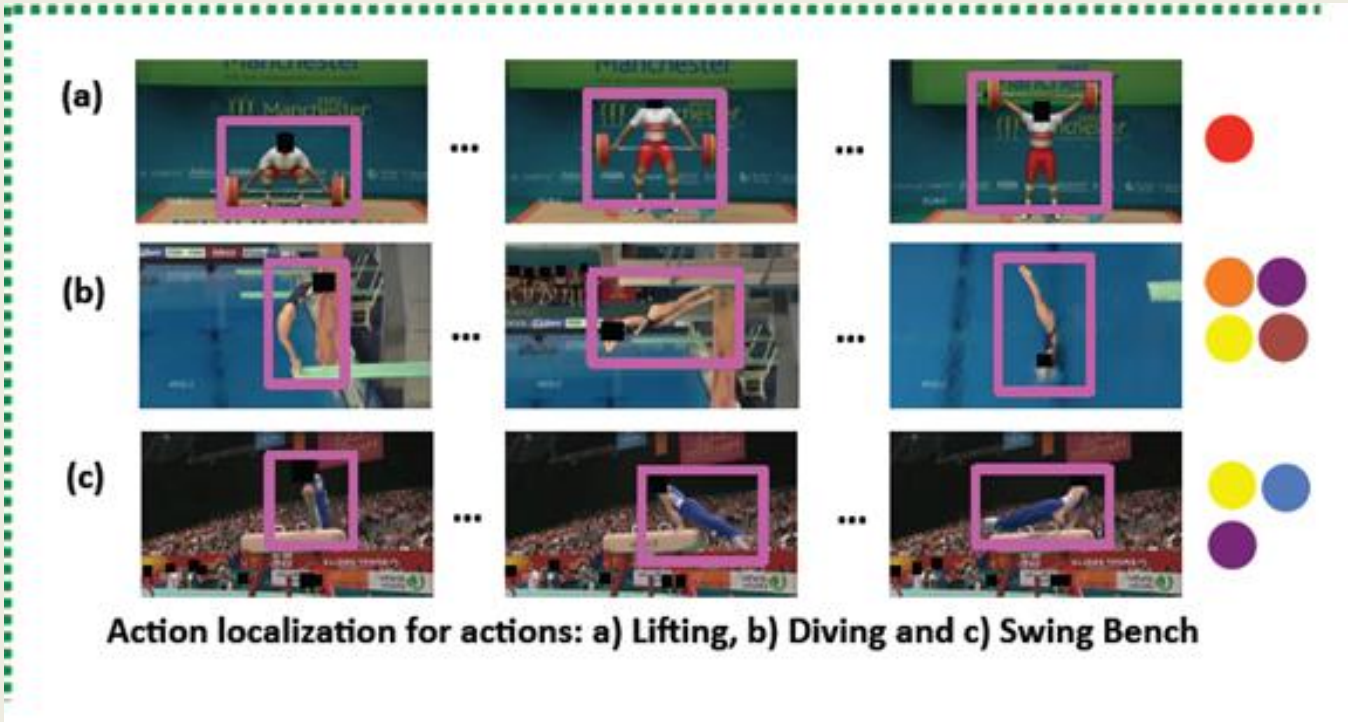
1. **動き**を元に認識する方法

Wangらの研究

Action recognition by dense trajectories, CVPR, 2013

動作認識の手法

2. **物体やシーン認識**を使って認識する方法



Action localization for actions: a) Lifting, b) Diving and c) Swing Bench

Jainらの研究
What do 15,000 object categories tell us about classifying and localizing actions? , CVPR, 2015

提案手法

単体の認識手法

動作認識

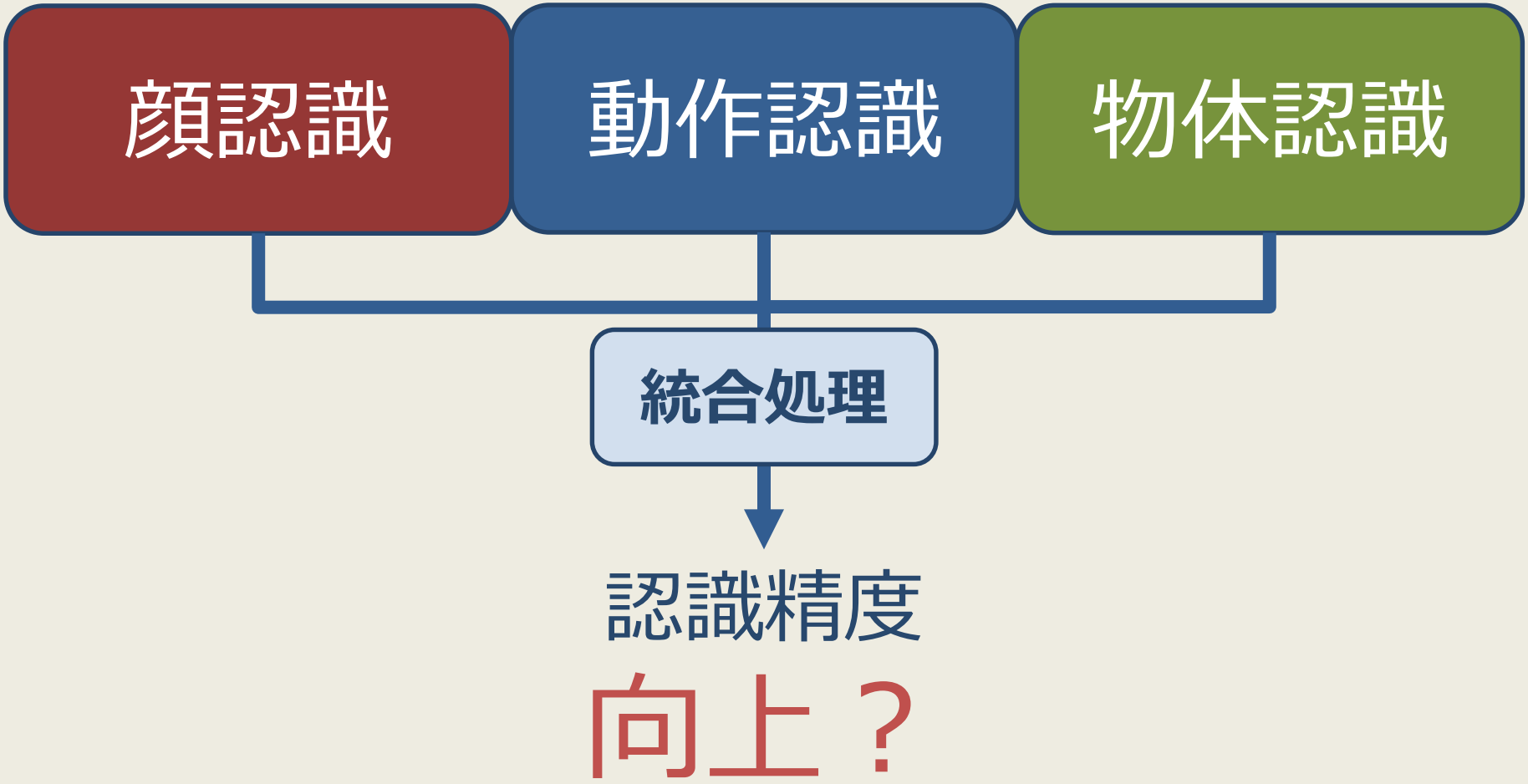
↓

認識精度

低い

提案手法

複数の認識手法



提案手法

テレビ映像

統合処理

顔認識 ＋ 動作認識 ＋ 物体認識

動作シーン

動作認識

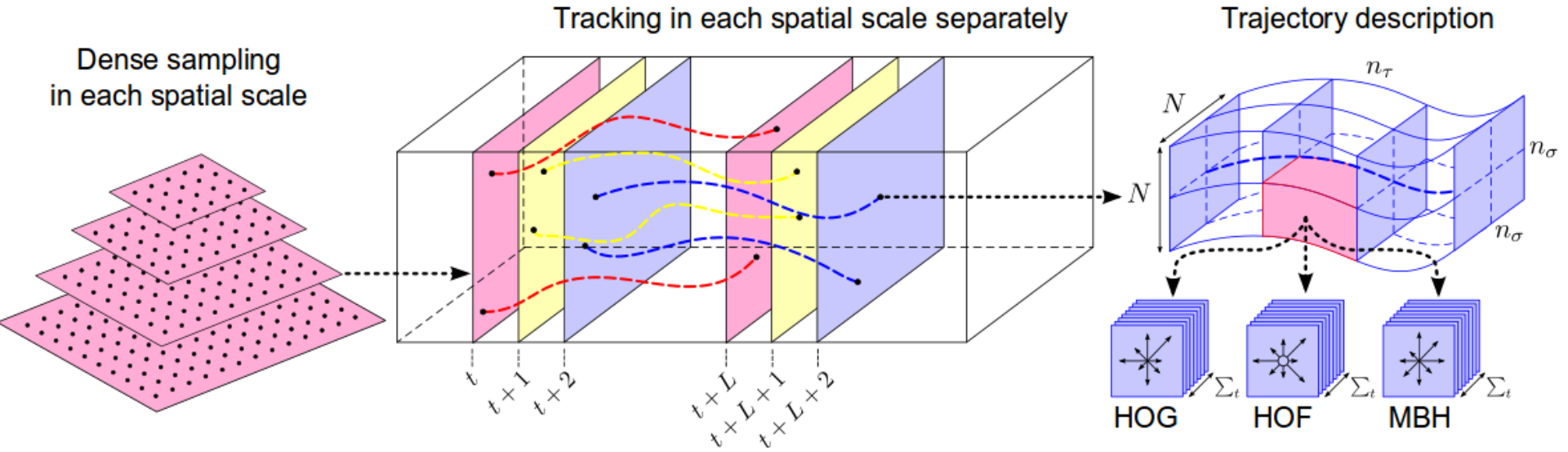
統合処理

顔認識 ＋ 動作認識 ＋ 物体認識

動作認識

時空間特徴抽出

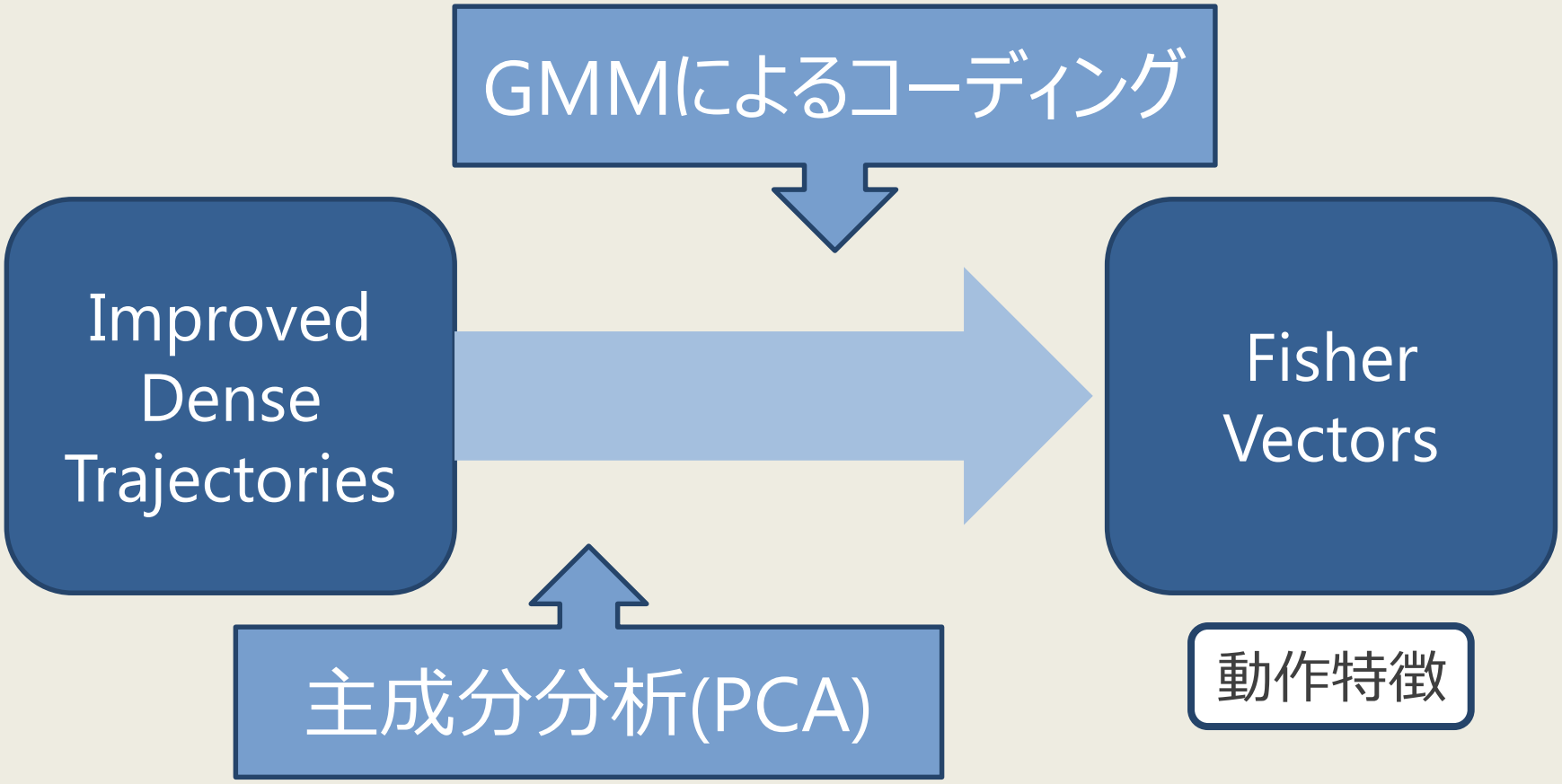
Improved Dense Trajectories



Wangらの研究
Action recognition by dense trajectories, CVPR, 2013

動作認識

時空間特徴抽出



動作認識

SVMによる分類器

学習データ

- 食事動作
  - Youtubeから収集した食事シーン100本
- 非食事動作
  - UCF-101から収集した動作320本

動作認識

予備実験

学習データにおける動作単体での
SVMの5-fold cross validationによる評価

分類率 = **92.5 %**

顔認識

統合処理

顔認識 ＋ 動作認識 ＋ 物体認識

顔認識

人がいないのに食事動作認識した例

誤認識を減らしたい!!

顔認識

- 人が写っていない部分
  - 動作認識する必要はない

顔検出を使用

人のいる部分のみ動作認識が可能

物体認識

統合処理

顔認識 ＋ 動作認識 ＋ 物体認識

物体認識

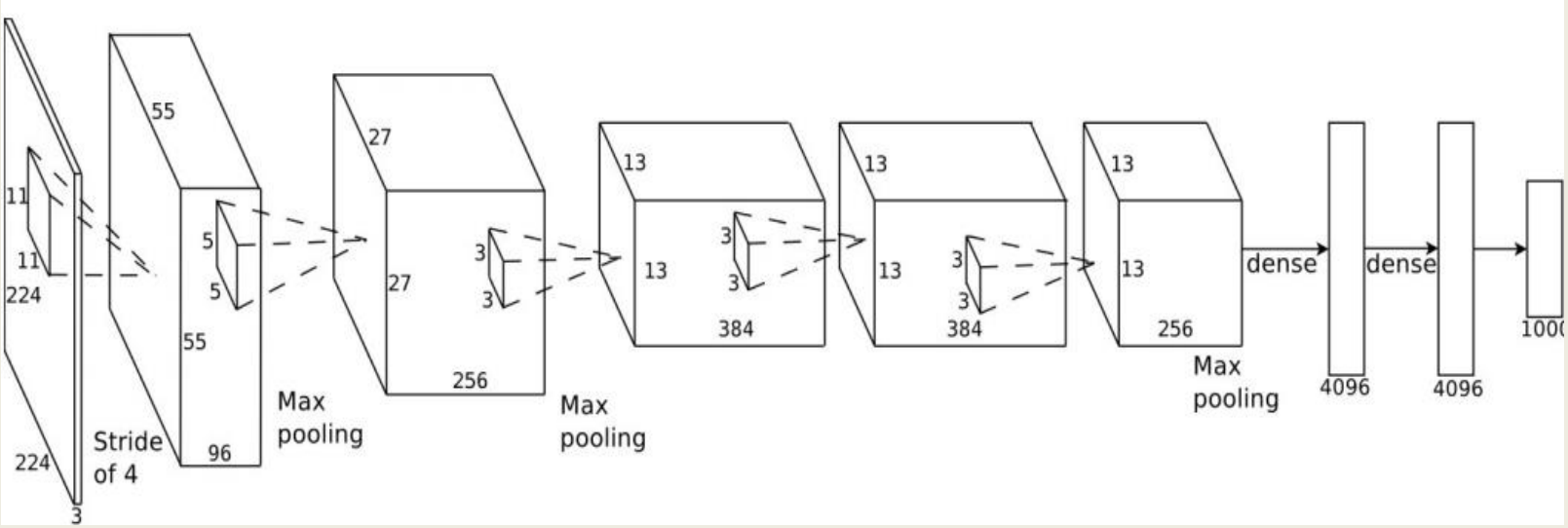
意味的に違っても
動作的に類似した動作の存在

writing

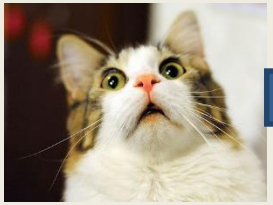
eating

物体認識

- 食事認識エンジンによる食事分類
- Deep learningを使用



食事



非食事

物体認識

食事・非食事分類

意味的に関連性のある動作のみを検出可能

物体認識

予備実験

– テレビ映像から食事が分類できるか実験

- 分類率：使用した画像の中で正しく認識された画像数の割合

$$\text{分類率} = \frac{true\ positive + true\ negative}{all\ images}$$

true positive = 「食事画像」として認識された食事画像数
true negative = 「非食事画像」として認識された非食事画像数
all images = 使用したすべての画像数

物体認識

- テレビ映像の食事分類率

分類率 = **90.0 %**

統合処理

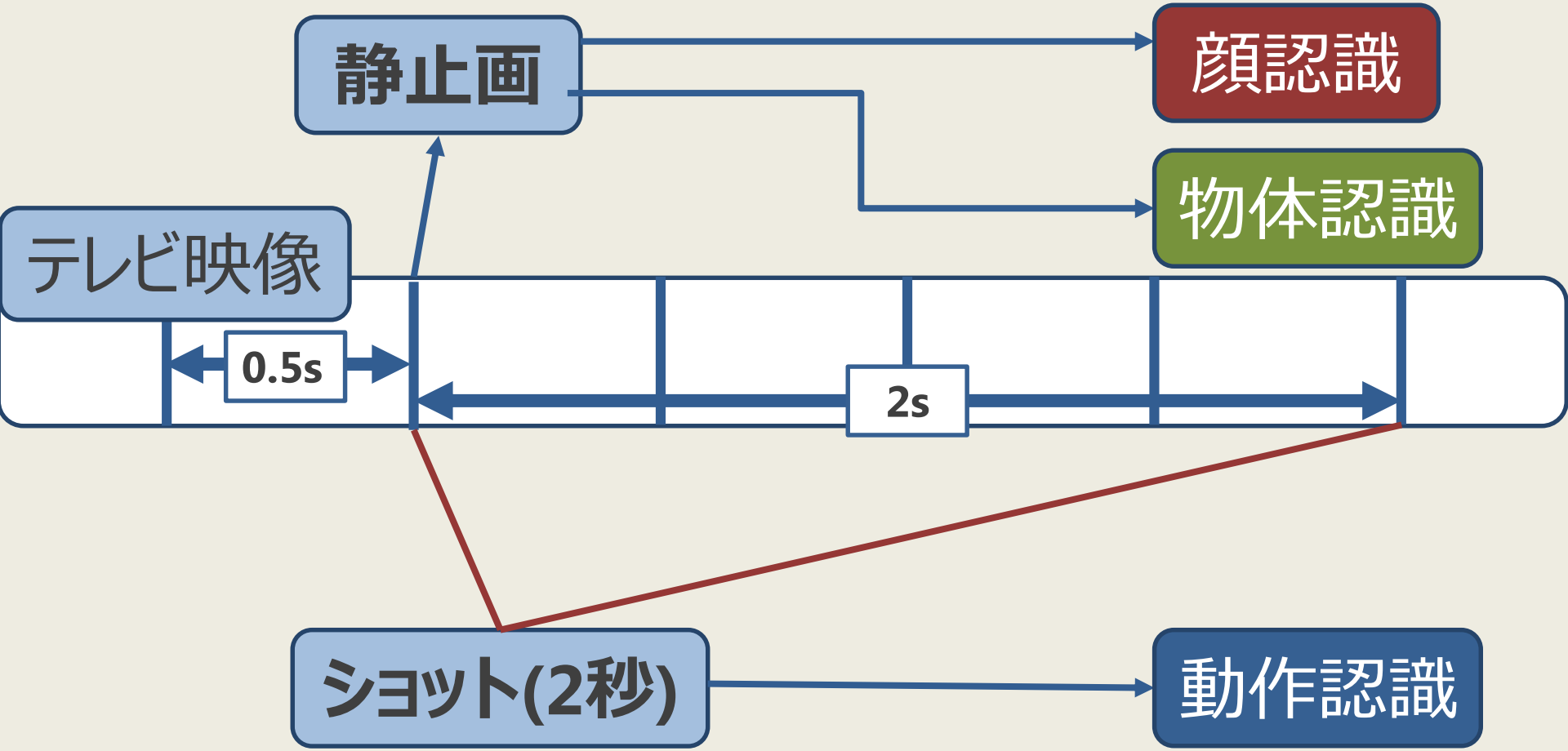
統合処理

顔認識 ＋ 動作認識 ＋ 物体認識

認識の統合処理

テレビ映像の処理

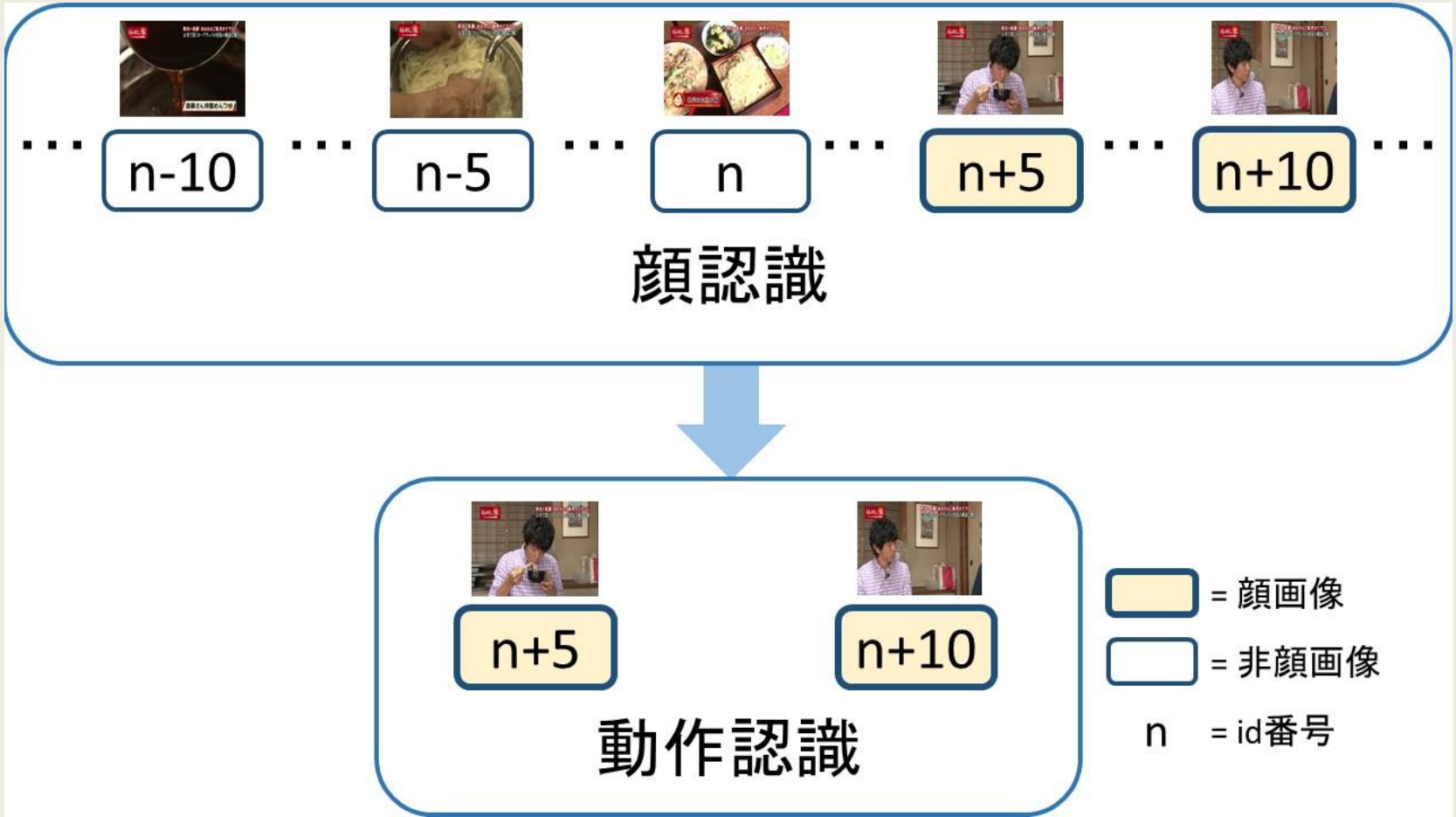
– 0,5秒おきに静止画，ショット(2秒)を抽出



認識の統合処理

顔認識 + 動作認識

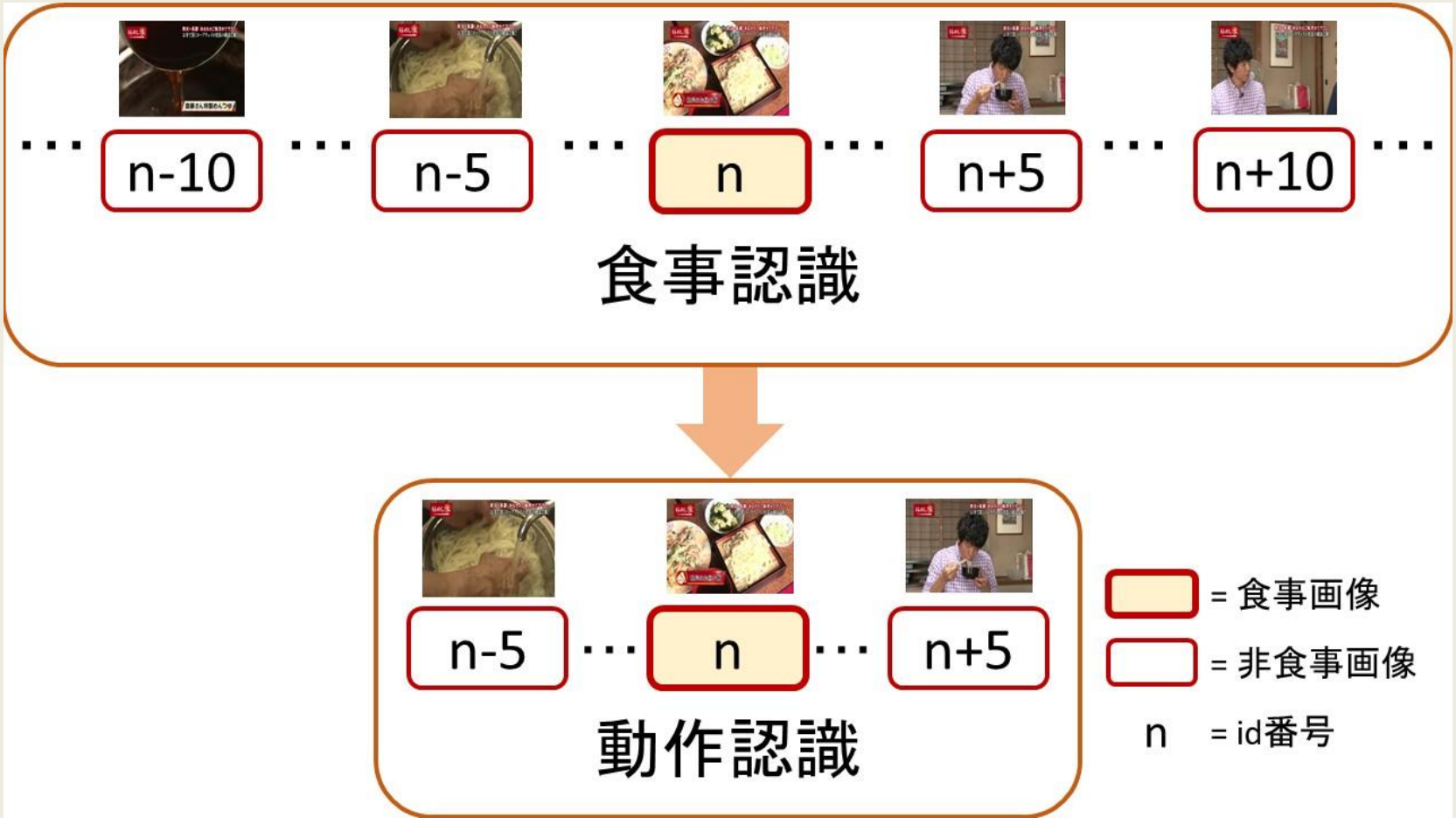
- 顔認識**できたもの**に対して動作認識



認識の統合処理

物体認識 + 動作認識

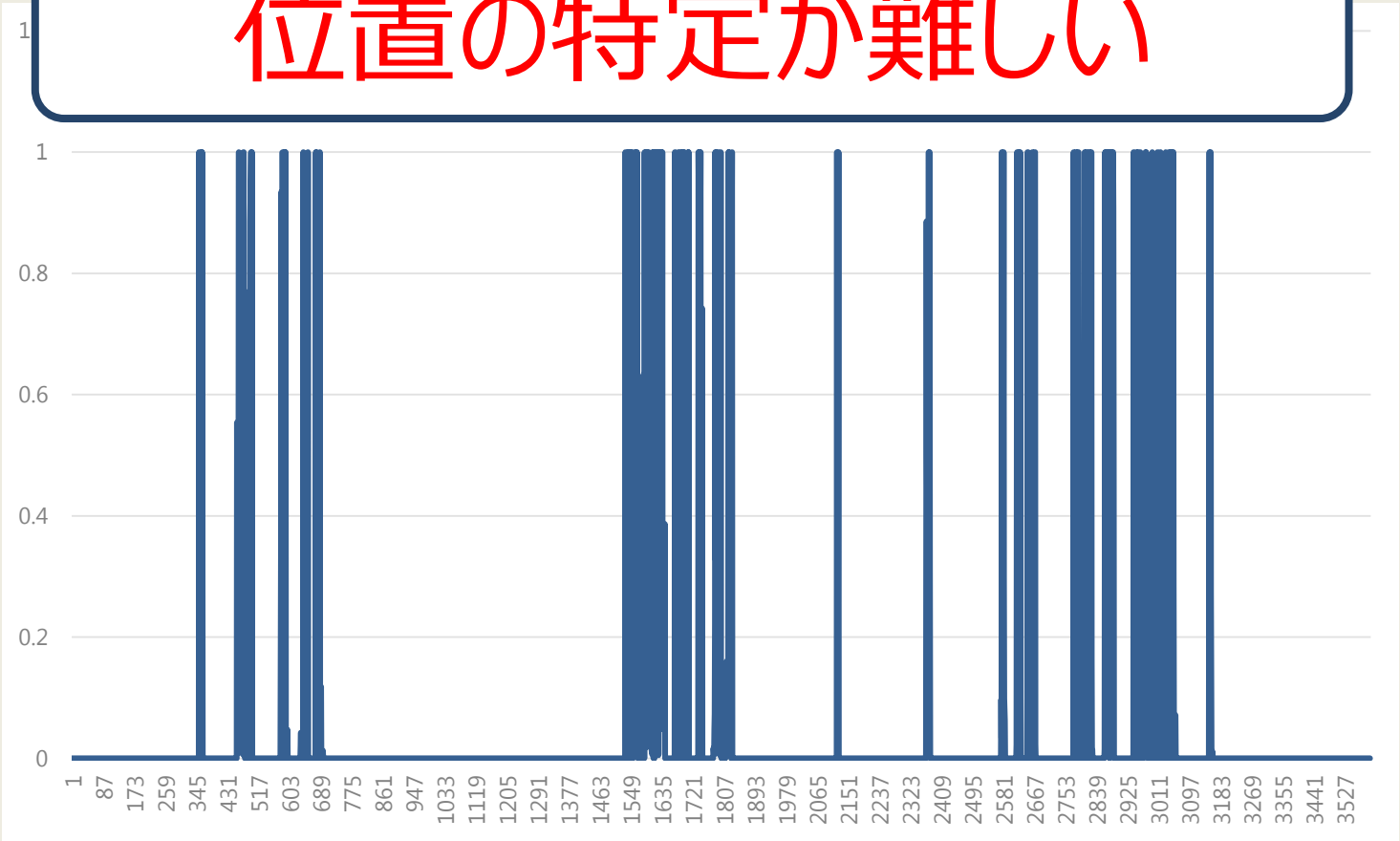
- 物体認識**できたもの周辺**に対して動作認識



シーンの検出

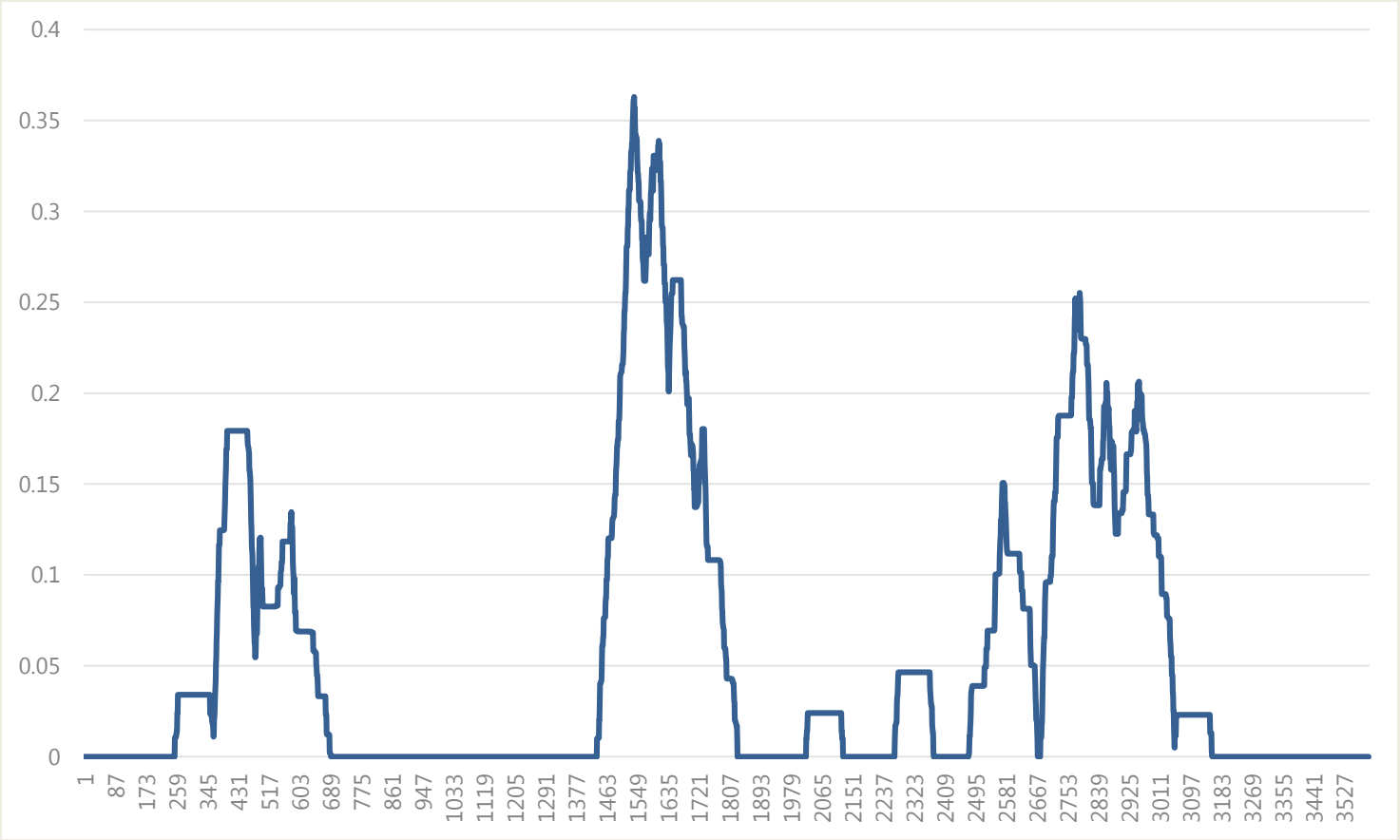
確率値をそのままプロットしたグラフ例

位置の特定が難しい



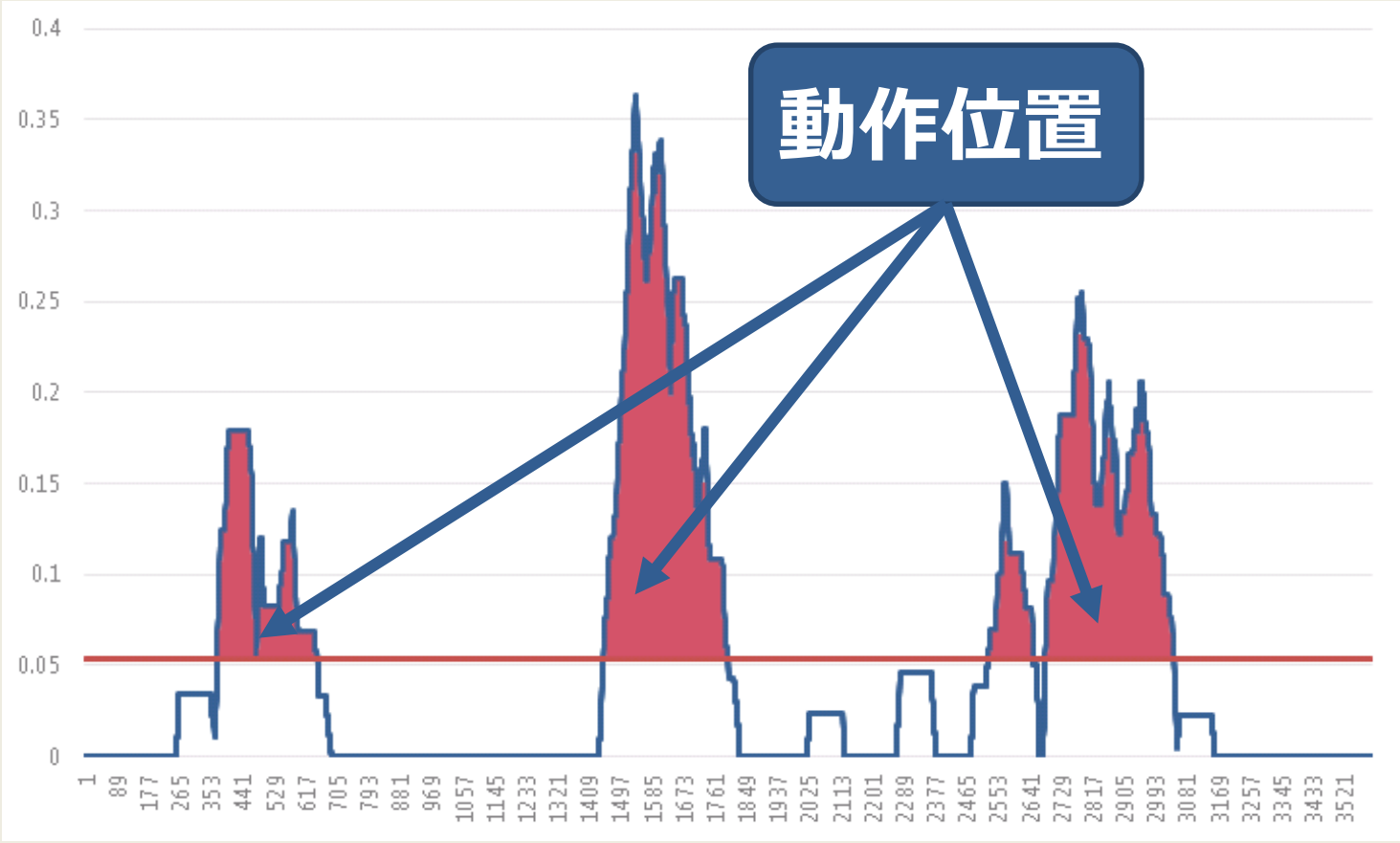
シーンの検出

類似度の前後50ショットの値の平均値を分布



シーンの検出

グラフの平均以上の値の集合を動作位置とする



実験：各組み合わせでの比較

- 動画に食事シーンが含まれているか実験

- 以下のパターンで実験
  - 動作認識
  - 動作認識+顔認識
  - 動作認識+食事認識
  - 動作認識+顔認識+食事認識

実験：各組み合わせでの実験

テストデータ

- 旅番組とバラエティ番組を一つずつ用意

1. 昼飯旅　～あなたのご飯見せてください！～
   - 放送時間：45分
2. 雨上がり食楽部
   - 放送時間：30分

実験：適合率, 再現率

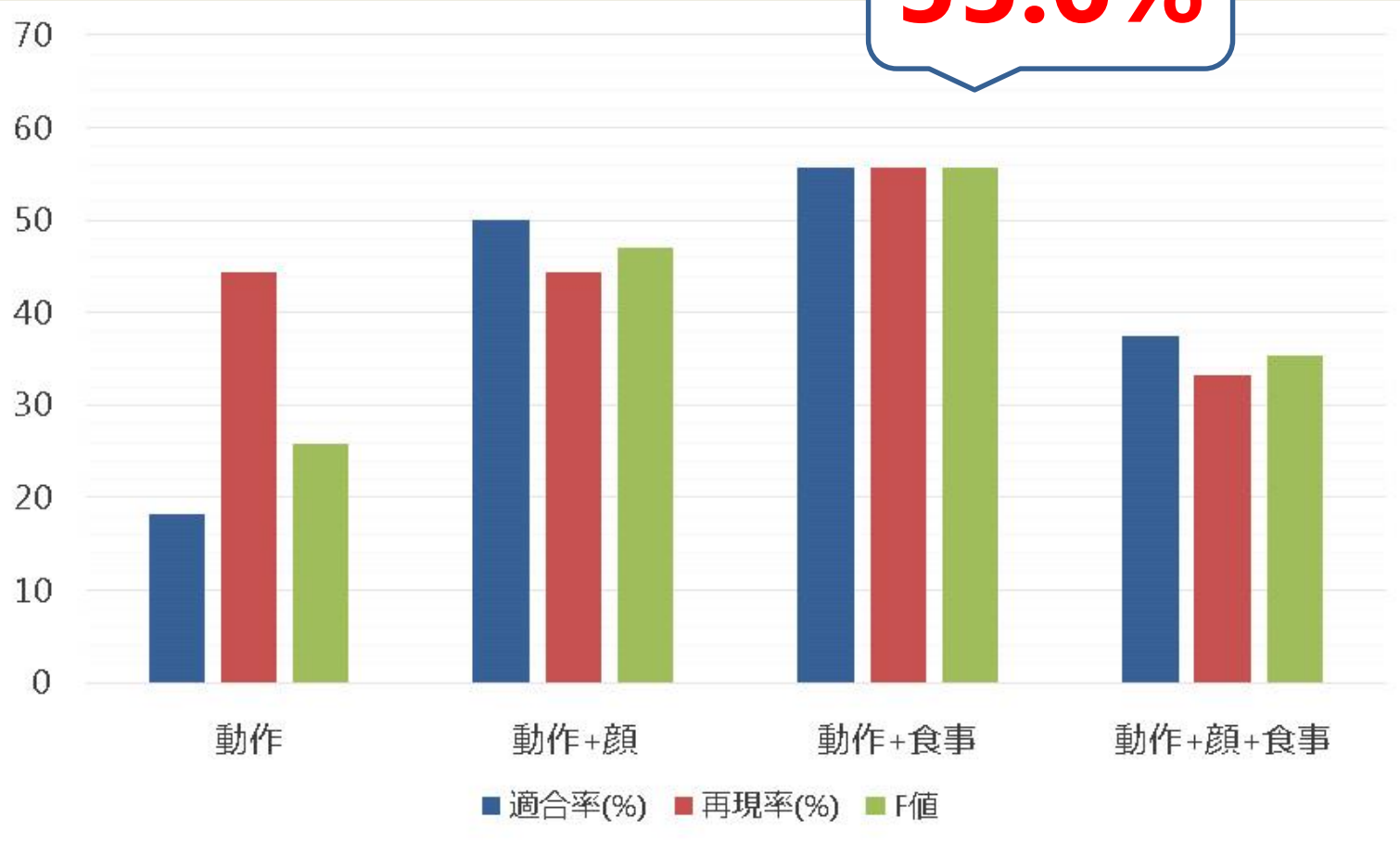
$$\text{適合率} = \frac{true\ positive}{true\ positive + false\ positive}$$

$$\text{再現率} = \frac{true\ positive}{true\ positive + false\ negative}$$

$$\text{F値} = \frac{2 * precision * recall}{precision + recall}$$

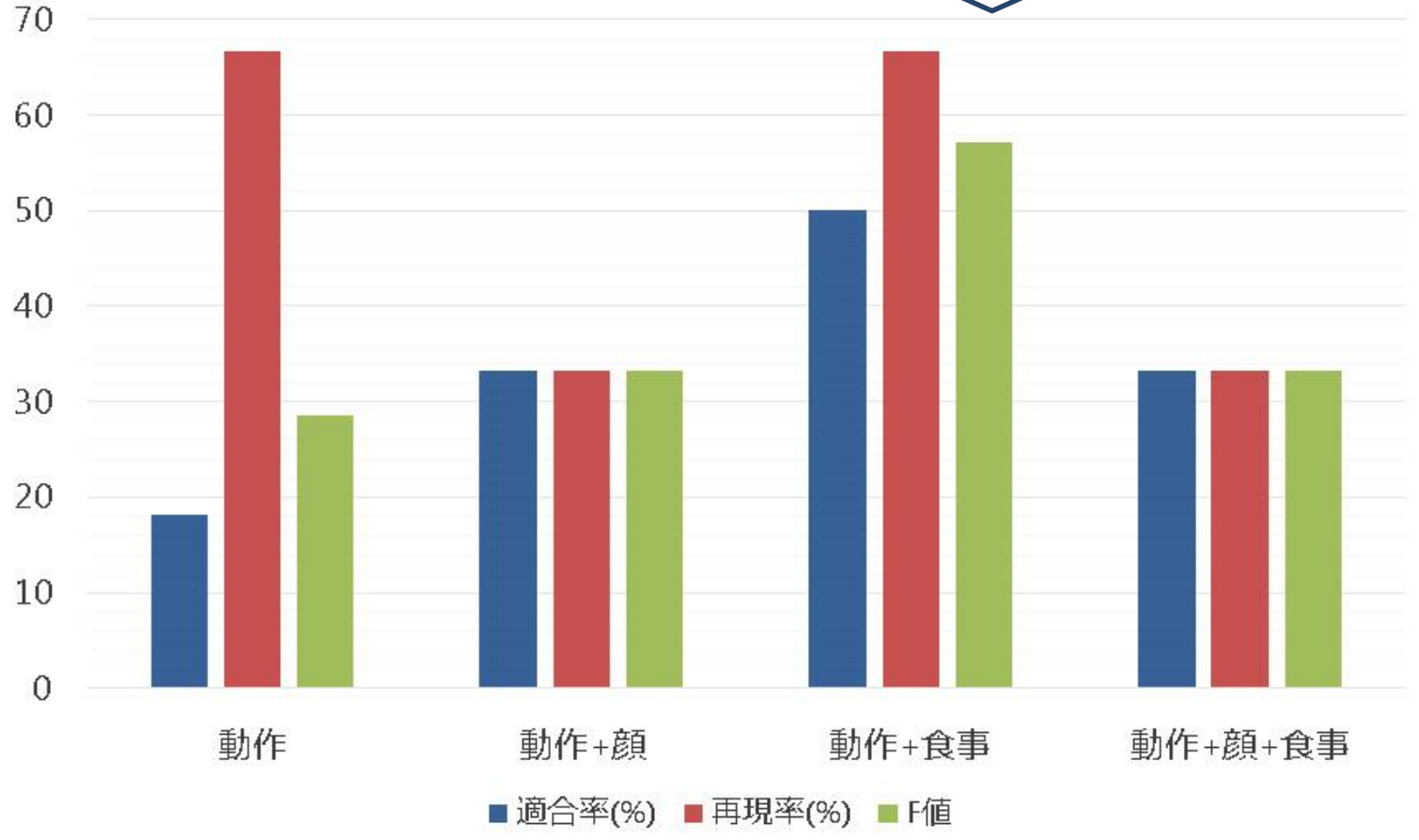
検出結果(昼飯旅)

55.6%



検出結果(雨上がり食楽部)

57.2%



検出の正解例(昼飯旅)

検出の正解例(雨上がり食楽部)

誤検出動作の例(昼飯旅)

未検出動作の例(昼飯旅)

まとめ

- テレビ映像から「食べる」動作を自動で検出, 認識精度の評価をした
- 複数の認識手法を組み合わせを実験
  - **食事認識＋動作認識**を組み合わせたものが最もよい結果となった

今後の課題

- 他の手法との組み合わせ
  - 例：姿勢推定, 字幕データetc
- 他の動作の推定

資料 1

	検出数	正解数	適合率	再現率	F値
動作	22	4	18.2	44.4	25.8
動作+顔	8	4	50	44.4	47.0
動作+食事	9	5	**55.6**	**55.6**	**55.6**
動作+顔+食事	8	3	37.5	33.3	35.3

	検出数	正解数	適合率	再現率	F値
動作	11	2	18.2	66.7	28.6
動作+顔	3	1	33.3	33.3	33.3
動作+食事	4	2	**50**	**66.7**	**57.2**
動作+顔+食事	3	1	33.3	33.3	33.3