第８回 ステアラボ人工知能セミナー

コンピュータビジョンによる動画認識

片岡 裕雄, *Ph.D.*

産業技術総合研究所

知能システム研究部門 コンピュータビジョン研究グループ (CVRG)

http://www.hirokatsukataoka.net/

動画認識の問題設定

動画には何が映っていますか？

画像識別

正解例：

```
predicted score/class1:        9.1733407974243 planetarium
predicted score/class2:        7.5232272148132 solar dish, solar collector, solar furnace
predicted score/class3:        7.3946485519409 racket, racquet
predicted score/class4:        7.3526573181152 tennis ball
predicted score/class5:        7.3053464889526 bullet train, bullet
```

Powered by S. Zagoruyko, NIN@Torch7 (https://github.com/torch/tutorials/tree/master/7_imagenet_classification)

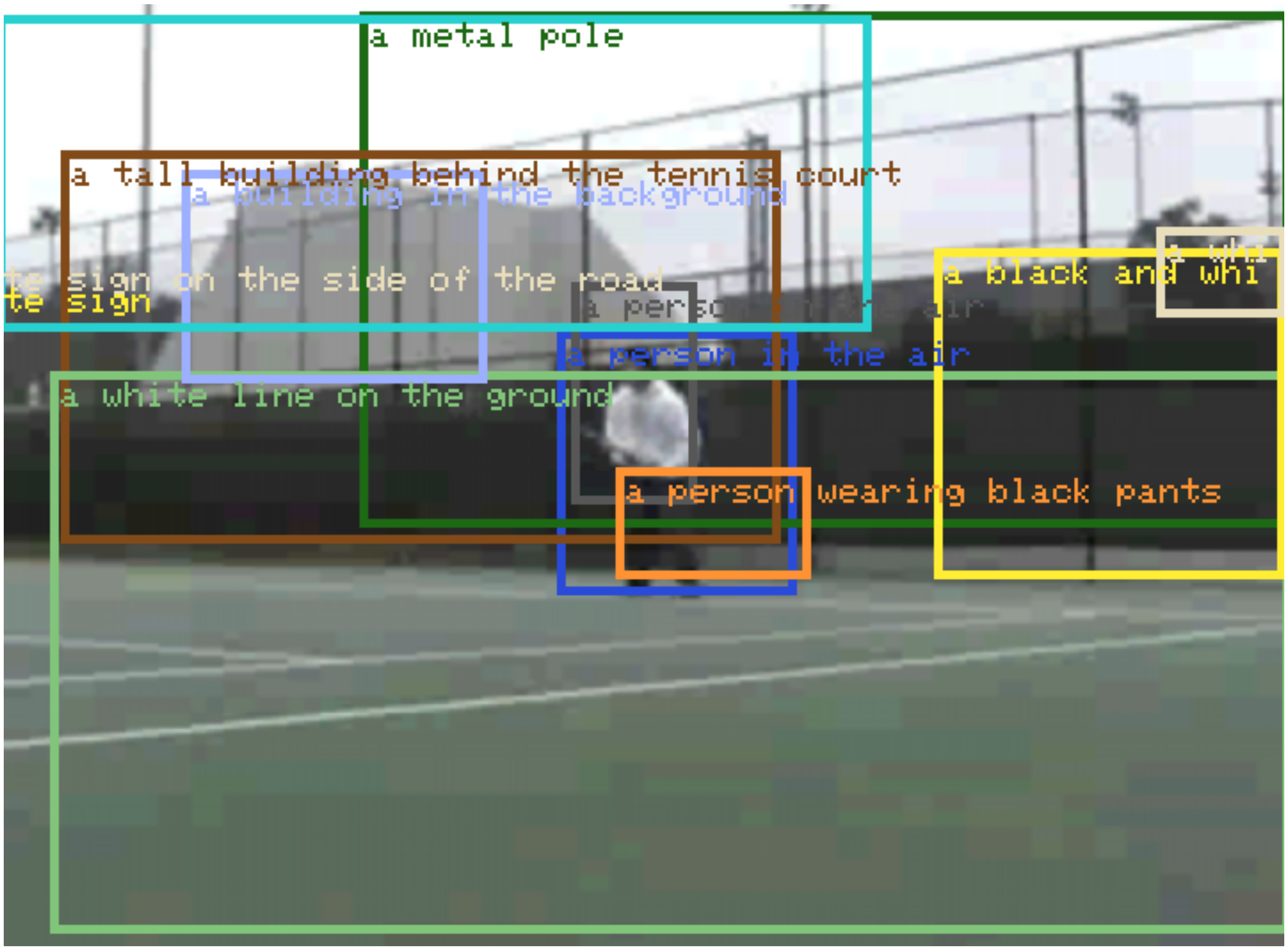
動画ラベル

正解例：TennisSwing

Powered by L. Wang, Very Deep Two-stream CNN (http://yjxiong.me/others/action_recog/)

画像説明文

正解：下図



Powered by J. Johnson, DenseCap (http://cs.stanford.edu/people/karpathy/densecap/)

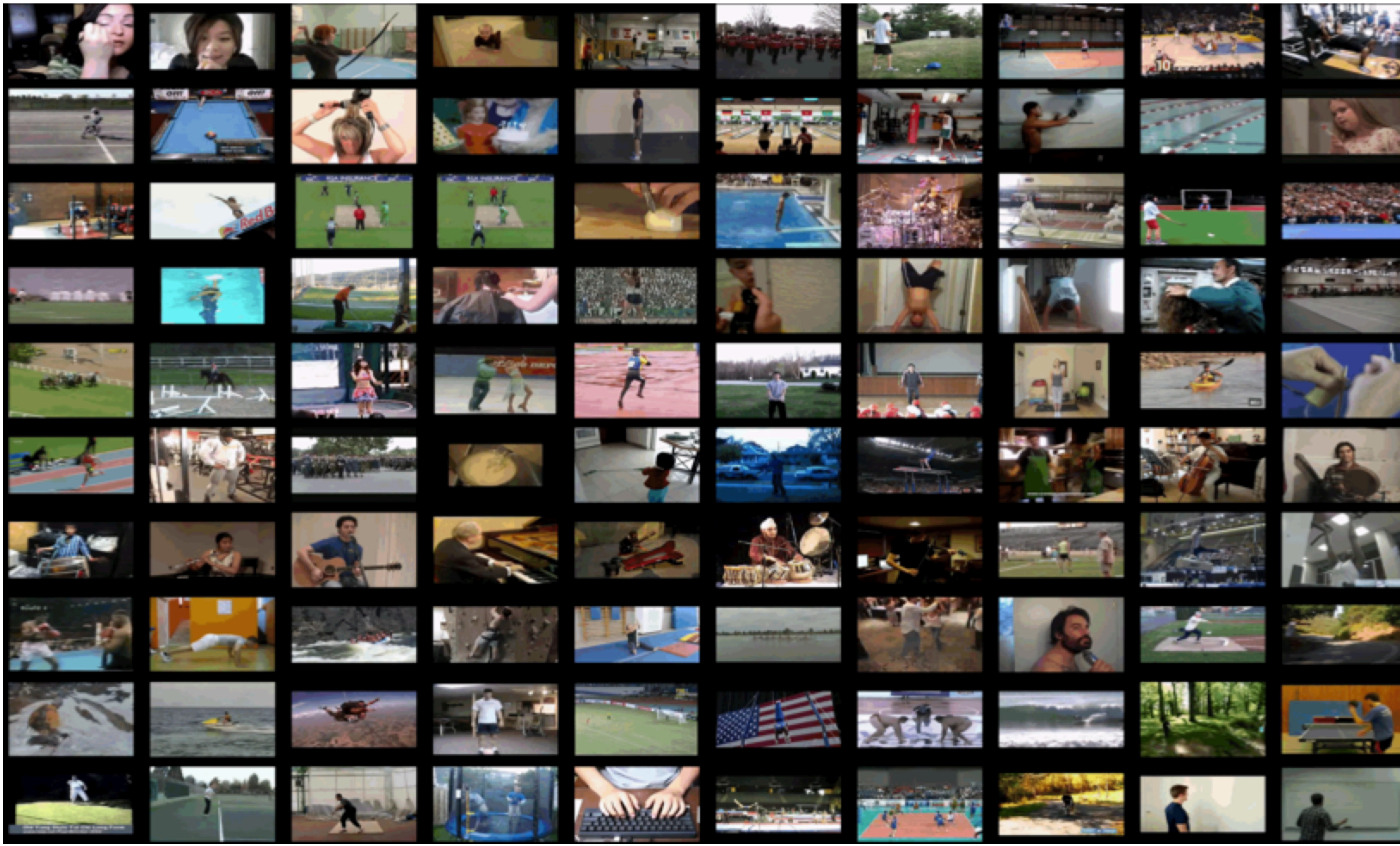
動画解析って難しい

何を認識する？

– 物体？行動？シーン？解説文？

頻出データセット：UCF101

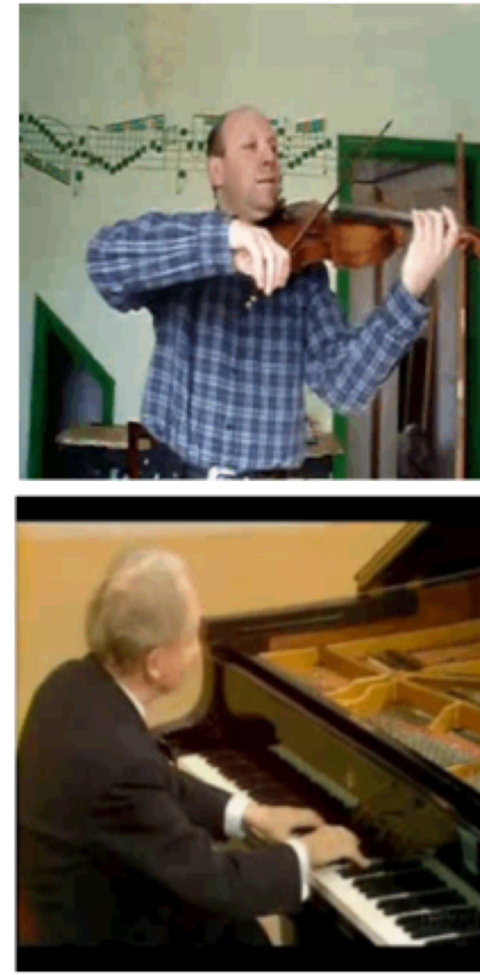
101行動を認識する



Soomro, K., Zamir, A.R. and Shah, M. “UCF101: A dataset of 101 human actions classes from videos in the wild”, arXiv pre-print, 2012.

UCF101の詳細

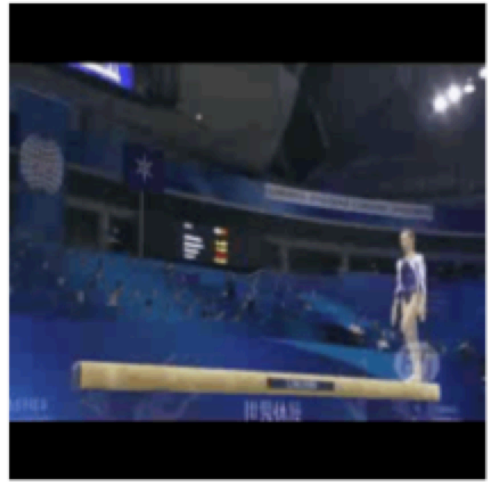
スポーツ



楽器演奏

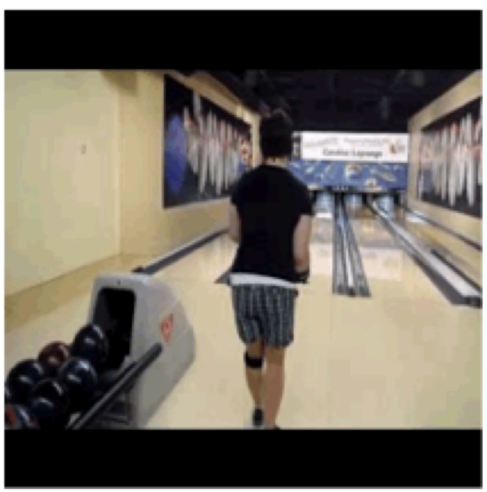
UCF101の詳細

Biking



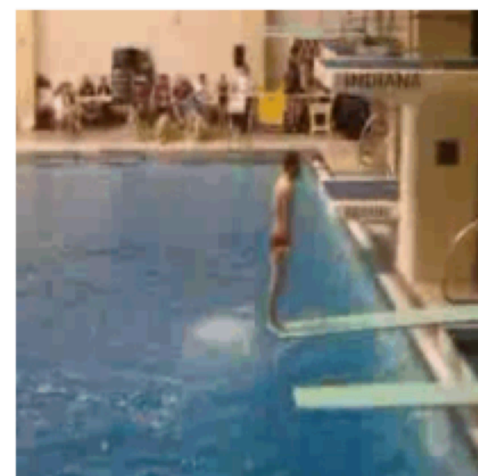
BalanceBeam

TennisSwing



Bowling

Fencing



Diving

頻出データセット：HMDB51

Kuehne, H. , Jhuang, H., Garrote, E. , Poggio, T. and Serre, T. “HMDB: A Large Video Database for Human Motion Recognition,” in ICCV, 2011.

HMDB51の詳細

顔表情

インタラクション

同データセットの問題：１動画１ラベル

問題を簡単にする

- 動画入力 => モーション表現 => 動画ラベル推定
- 本発表においては動画-ラベルが１対１

Motion Descriptor

Tennis Swing

本講演の目的

動画認識

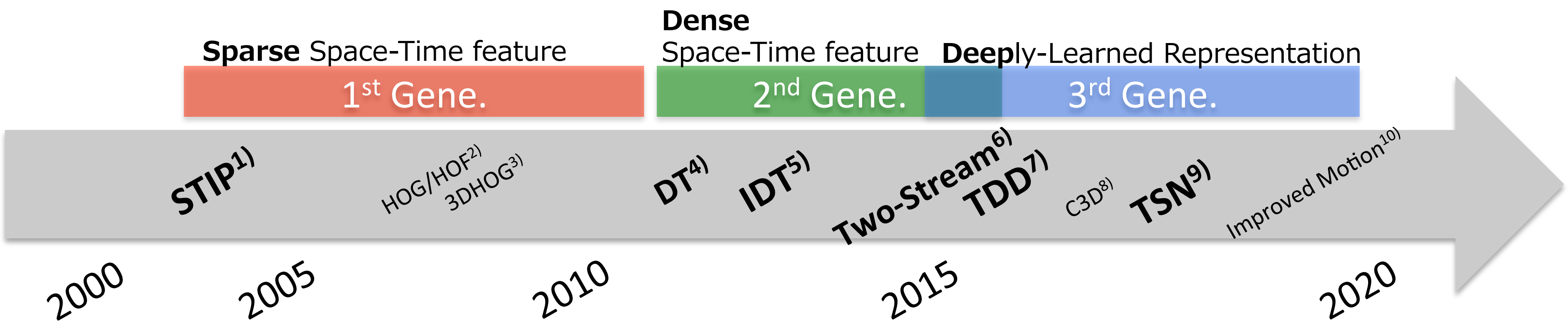
- （本講演においては）動画とラベルが１対１で対応
- 入力～モーション表現～ラベルを推定

講演で学習すること

- 動画認識の流れについて把握
- 重要技術の理解を深める
- 今後の展望「何をすれば良いか？」

動画解析の流れ

動画認識の流れ – Sparse, Dense and Deep



1) Laptev, I. and Lindeberg, T. “Space-Time Interest Points,” International Conference on Computer Vision (ICCV), pp.432–439, 2003.

2) Laptev, I., Marszalek, M., Schmid, C. and Rozenfeld, B. “Learning realistic human actions from movies,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1–8, 2008.

3) Klaser, A., Marszalek, M., and Schmid, C. “A Spatio-Temporal Descriptor Based on 3D-Gradients,” British Machine Vision Conference (BMVC), 2008.

4) Wang, H., Klaser, A., Schmid, C. and Liu, C.-L. “Action recognition by dense trajectories,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3169–3176, 2011.

5) Wang, H. and Schmid, C. “Action Recognition with Improved Trajectories,” International Conference on Computer Vision (ICCV), pp.3551–3558, 2013.

6) Simonyan, K. and Zisserman, A. “Two-Stream Convolutional Networks for Action Recognition in Videos,” Neural Information Processing Systems (NIPS), 2014.

7) Wang, L., Qiao, Y. and Tang, X. “Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

8) D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks“, ICCV 2015.

9) Wang, L., Xiong, Y., Wang, Z. Qiao, Y., Lin, D., Tang, X. and Gool, L. C. “Temporal Segment Networks: Towards Good Practices for Deep Action Recognition,“ in ECCV 2016.

10) He, Y., Shirakabe, S., Satoh, Y. and Kataoka, H. “Human Action Recognition without Human,“ in ECCV WS 2016.

Space-Time Interest Points (STIP)

STIP

エポックメイクなアイディア- 動画認識はここから始まった

– 時系列特徴点の提案

– 画像の2次元特徴点と3次元特徴点

  • （直感的には）フレーム間差分から特徴点検出

STIPの基本的な情報

考案者

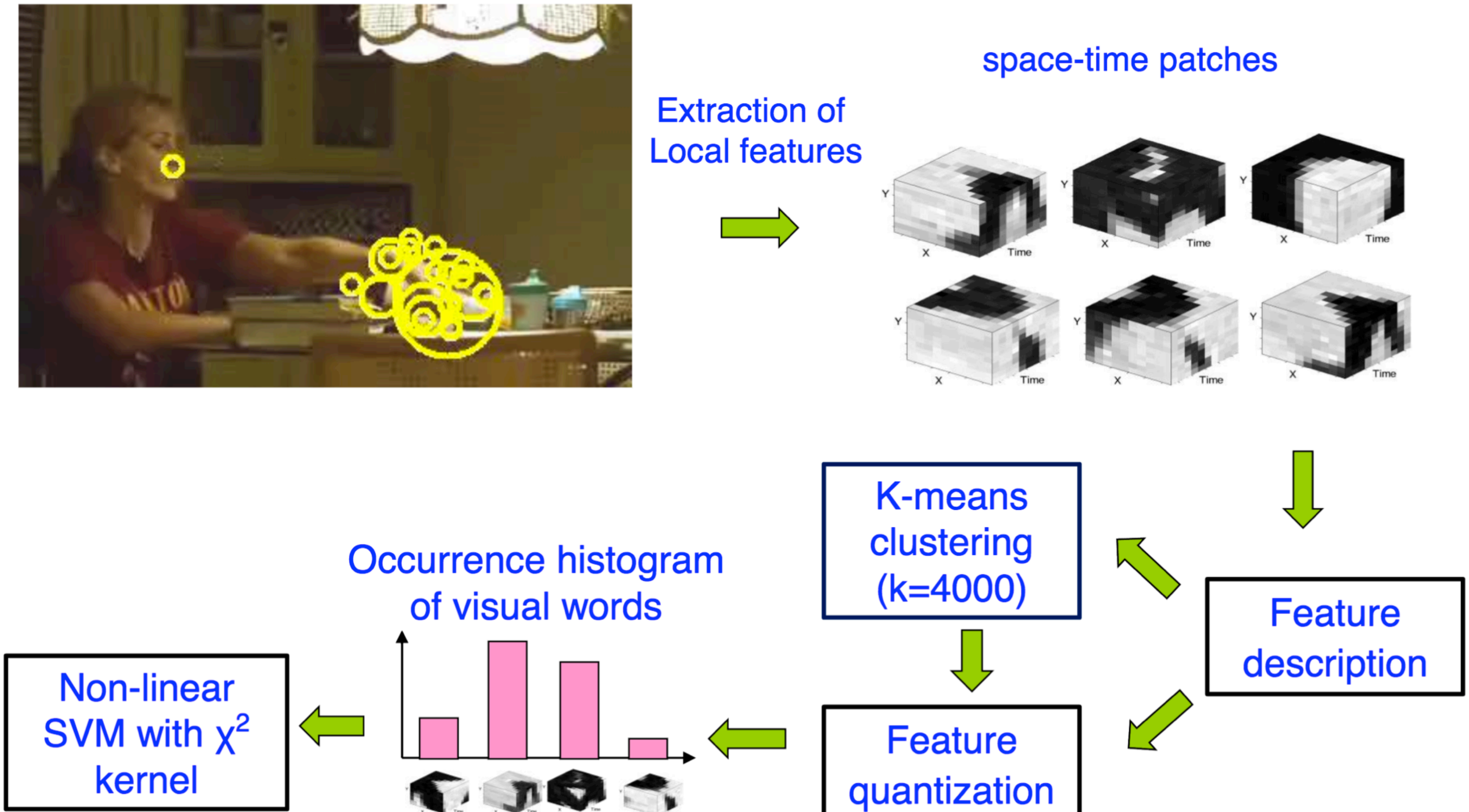
- Ivan Laptev (発表当時KTH所属、現INRIA)
- ICCV2003 (Journal: IJCV2005)

手法

- 時系列の特徴点検出というコンセプトが新しかった
- 特徴点検出・局所特徴抽出・コードワード化・識別の４ステップ

Laptev, I. and Lindeberg, T. “Space-Time Interest Points,” International Conference on Computer Vision (ICCV), pp.432–439, 2003.

STIPの概要



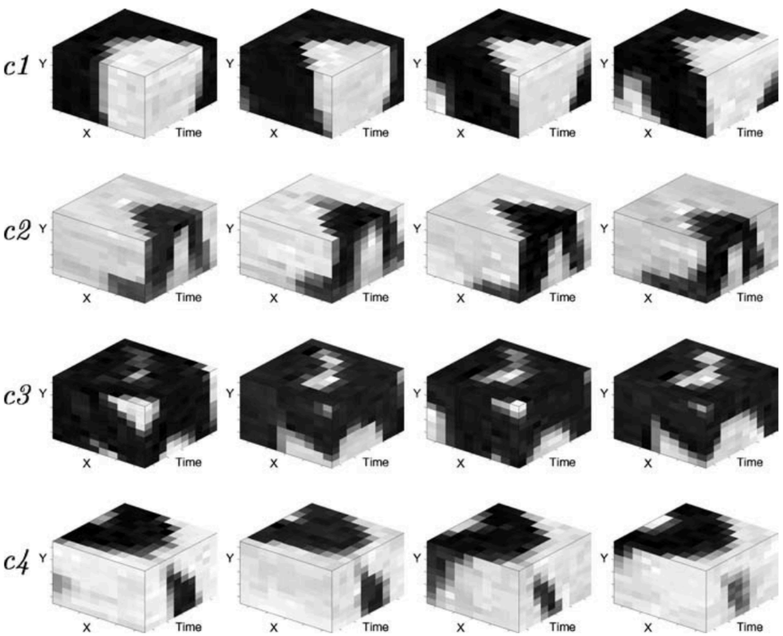
[Laptev, Marszałek, Schmid, Rozenfeld 2008]

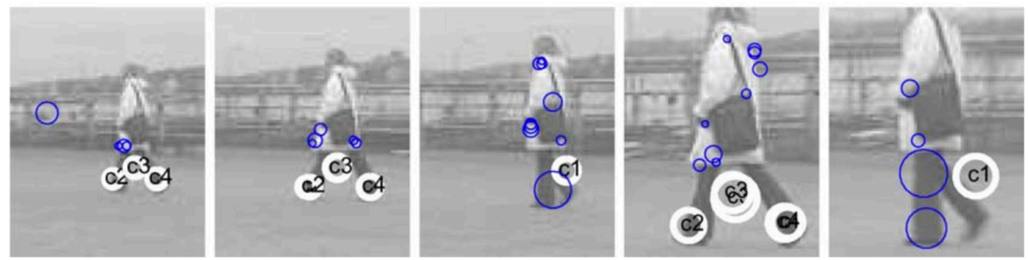
http://michaelryoo.com/cvpr2014tutorial/cvpr2014_tutorial_emerging_topics_laptev.pdf

特徴点検出

XYTの3次元空間に対して特徴点検出

– 2次元特徴点 (Harris operator)の拡張





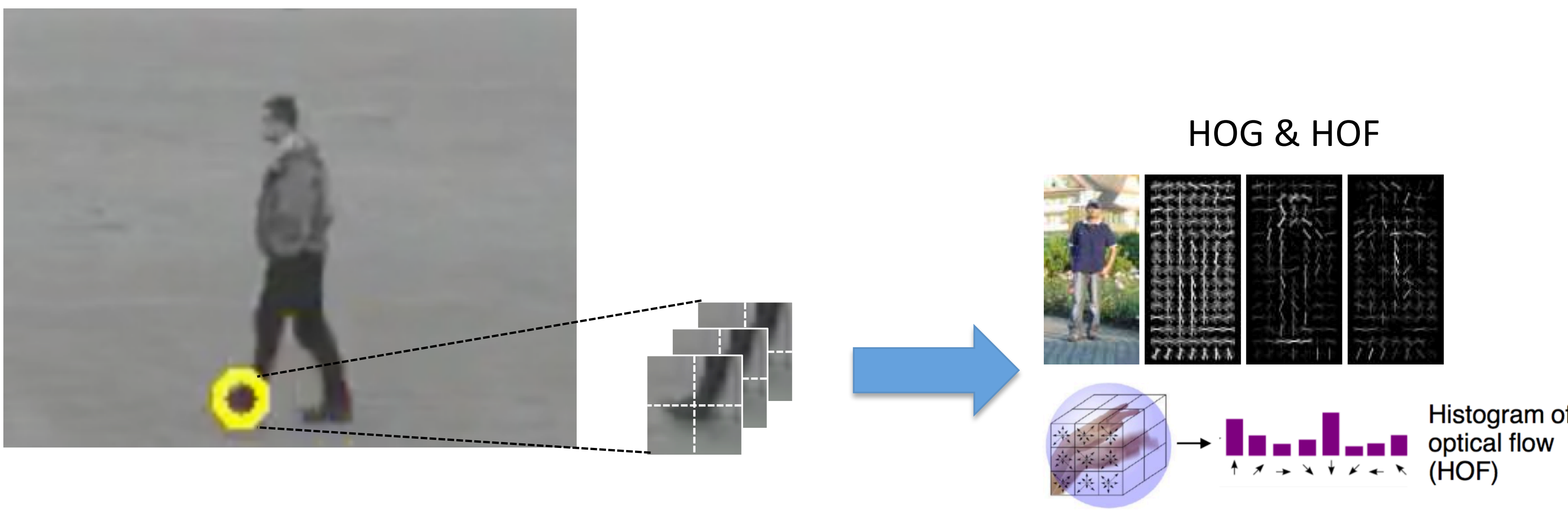
$$\mu = g(\cdot; \sigma_i^2, \tau_i^2) * \begin{pmatrix} L_x^2 & L_x L_y & L_x L_t \\ L_x L_y & L_y^2 & L_y L_t \\ L_x L_t & L_y L_t & L_t^2 \end{pmatrix}$$

$$\begin{aligned} H &= \det(\mu) - k\ \mathrm{trace}^3(\mu) \\ &= \lambda_1\lambda_2\lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3. \end{aligned}$$

局所特徴抽出

時系列の画像パッチから特徴抽出

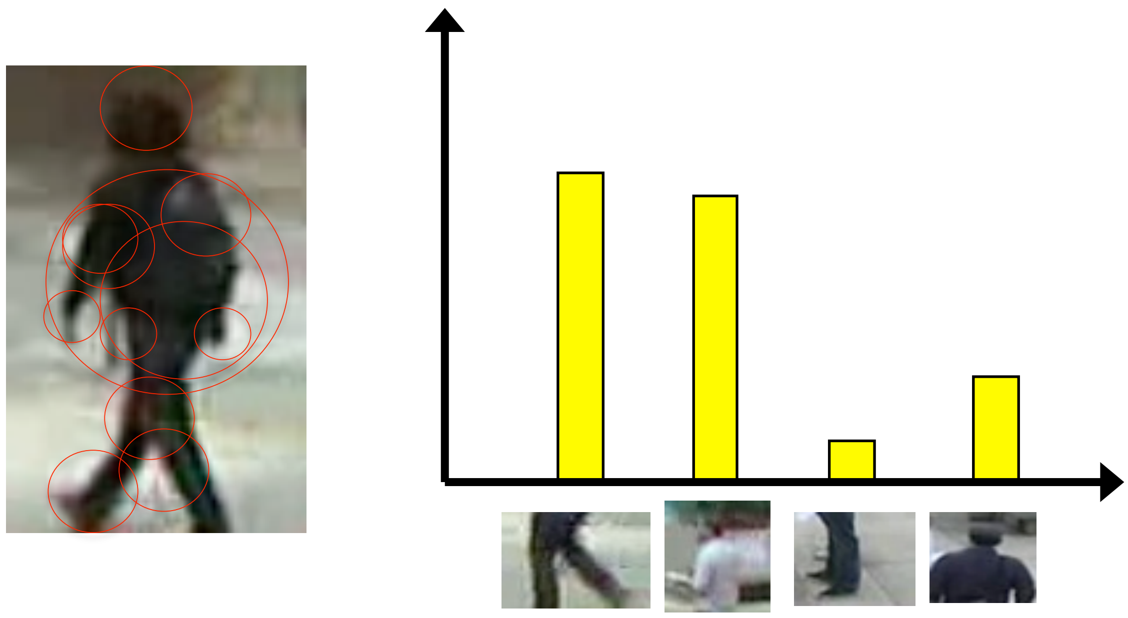
- XYT: 2x2x3のパッチから局所特徴抽出
- HOG & HOFがスタンダード
  - HOG: 勾配特徴のベクトル化
  - HOF: オプティカルフローのベクトル化



コードワード化：Bag-of-features (BoF)

「ベクトル量子化」と「最近傍割り当て」によりベクトル表現

- ベクトル量子化：予めk-meansによりベクトル(e.g. HOG/HOF)をクラスタリングして辞書を作成
- テスト時には辞書を参照してコードワードベクトルを生成、識別器によりラベルを推定



UCF101, HMDB51における精度

以降、比較のため代表的なデータであるUCF101やHMDB51における各手法のパフォーマンスを載せます

Approach	UCF101 (mAP)	HMDB51 (mAP)
STIP (HOG/HOF)	43.9	20.0

Dense Trajectories (DT)
Improved Dense Trajectories (IDT)

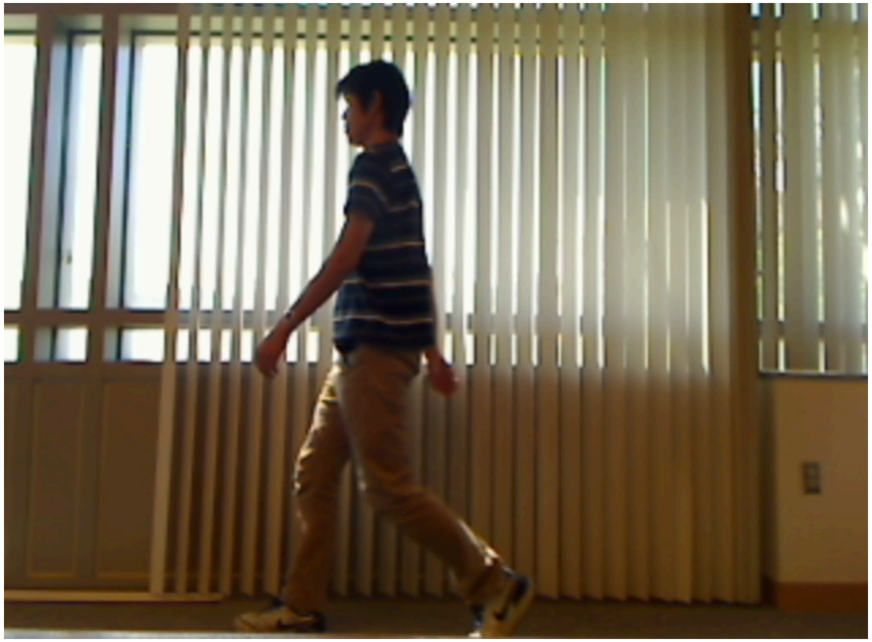
密(Dense)な動線(Trajectories)抽出

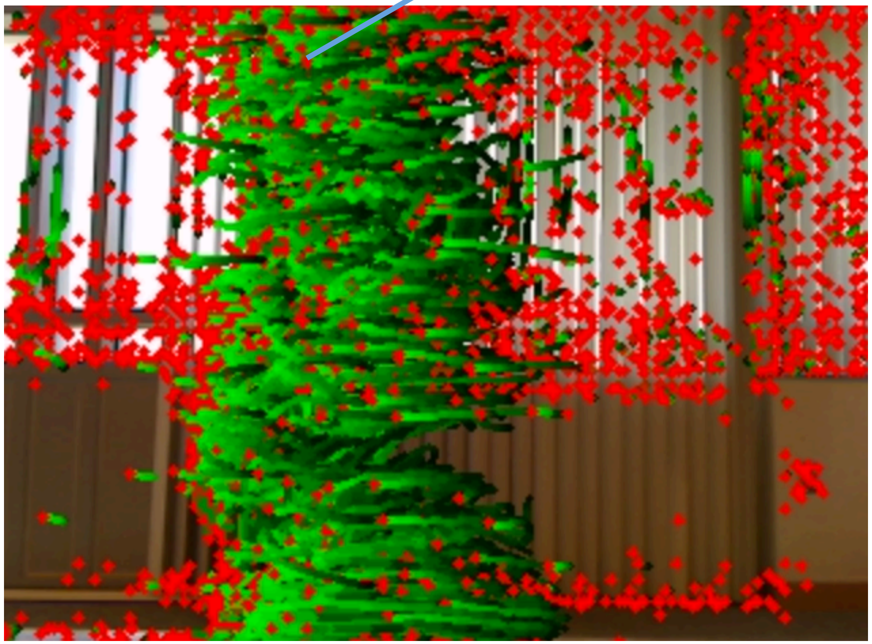
サンプリングを密にするのが精度が良いと判明

– STIPと比較すると、より「デンス」に

– 特徴点検出のみでなく、連結して「トラジェクトリ」に

圧倒的な特徴記述量！



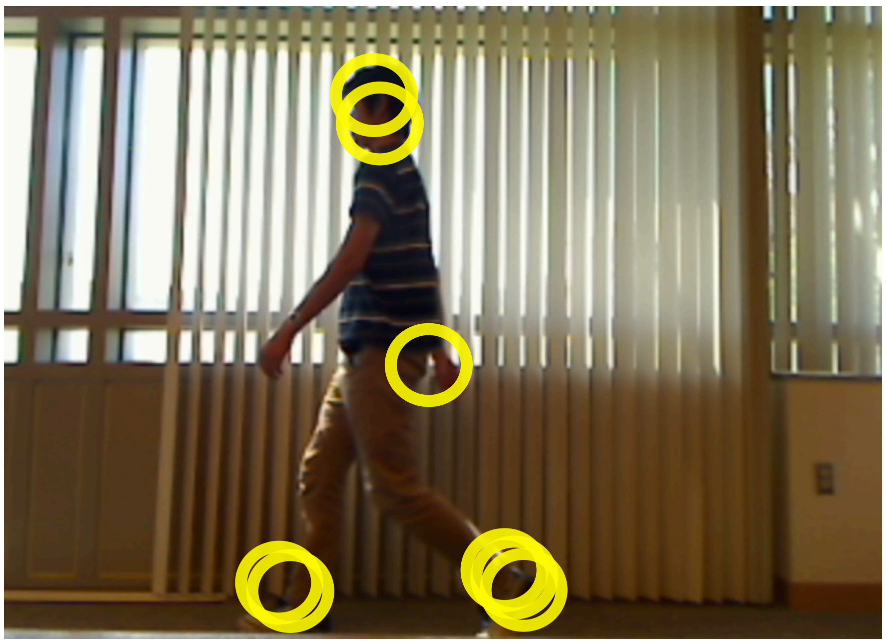


STIP vs DT

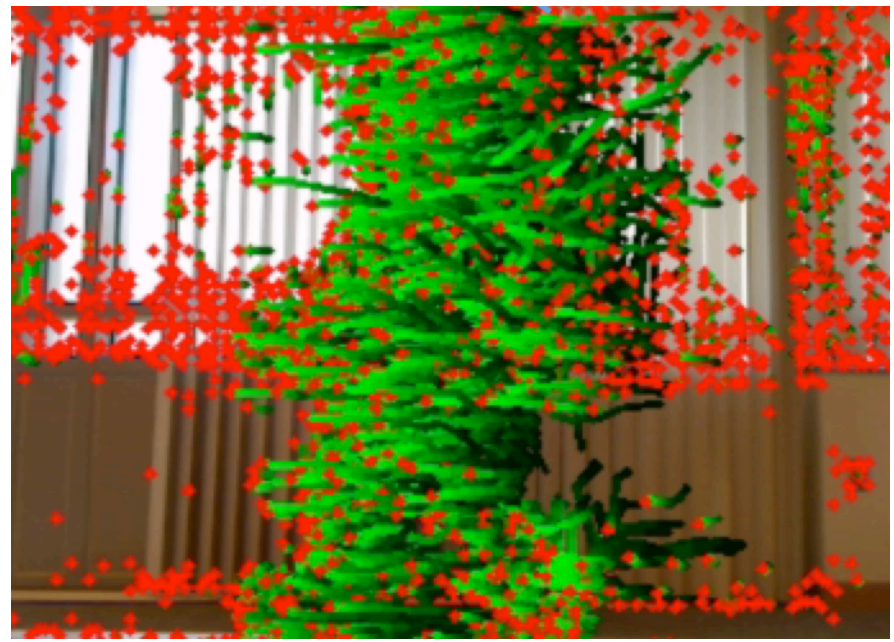
これだけ特徴記述量が違う！

– STIP: 数十サンプリング　　DT: 数百～数千サンプリング

STIP



DT



Dense Trajectoriesの基本的な情報

考案者

– Heng Wang (発表当時INRIA所属、現Amazon)
– CVPR2011 (Journal: IJCV2013)

手法

– STIPと比較して時間的にも空間的にも密な特徴点検出&追跡
– 特徴点検出・局所特徴抽出・コードワード化・識別の４ステップ

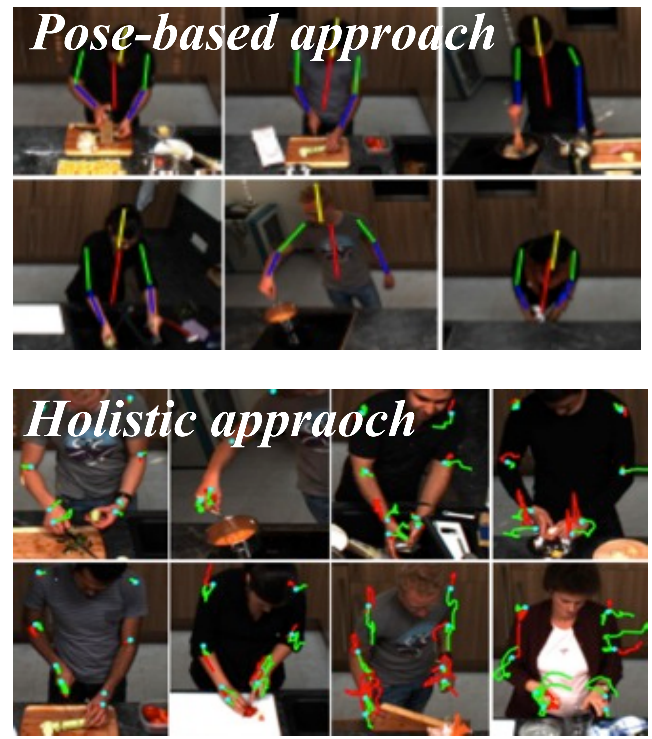
Wang, H., Klaser, A., Schmid, C. and Liu, C.-L. “Action recognition by dense trajectories,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3169–3176, 2011.

3D姿勢ベースの認識よりもロバスト

詳細行動認識 (Fine-grained recognition)にて

– 詳細行動認識では微小な違いによるクラスの違いを見分ける課題

– 姿勢特徴 34.6%に対してアピアランス特徴(=Dense Traj.)は59.2%を記録



M. Rohrbach, et al., “A Database for Fine Grained Activity Detection of Cooking Activities”, in CVPR2012.

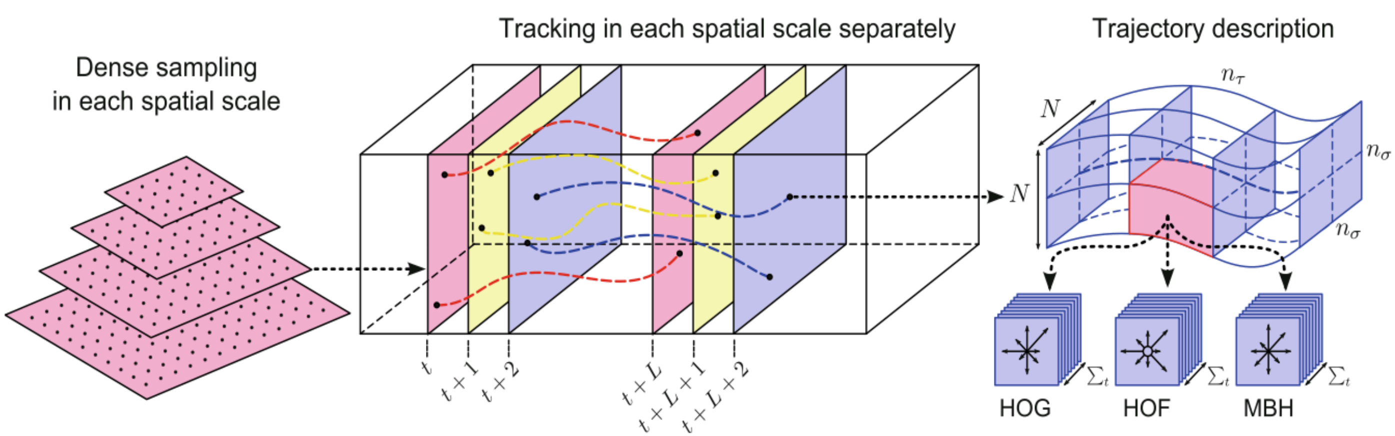
アルゴリズム

(i)特徴点追跡

- ピラミッド画像と特徴点サンプリング
- Farneback Optical Flowによるフロー抽出
- Median Filteringによるフローの対応付け

(ii)特徴記述

- HOG, HOF, MBH, Trajectory特徴の記述
- Bag-of-wordsによる特徴ベクトル化

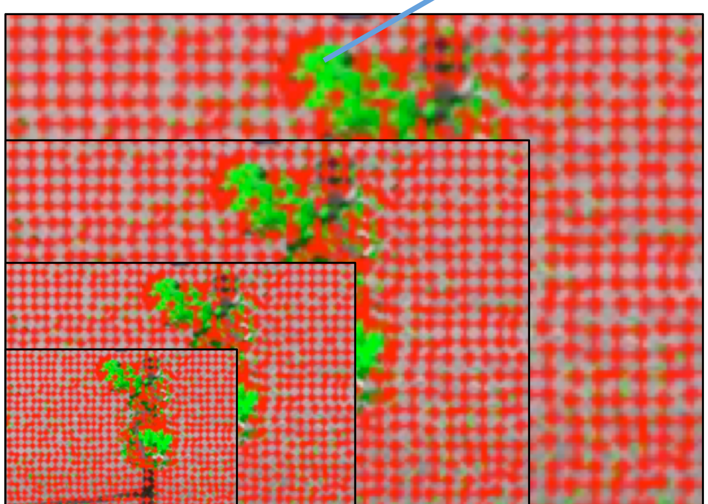


ピラミッド画像と特徴点サンプリング

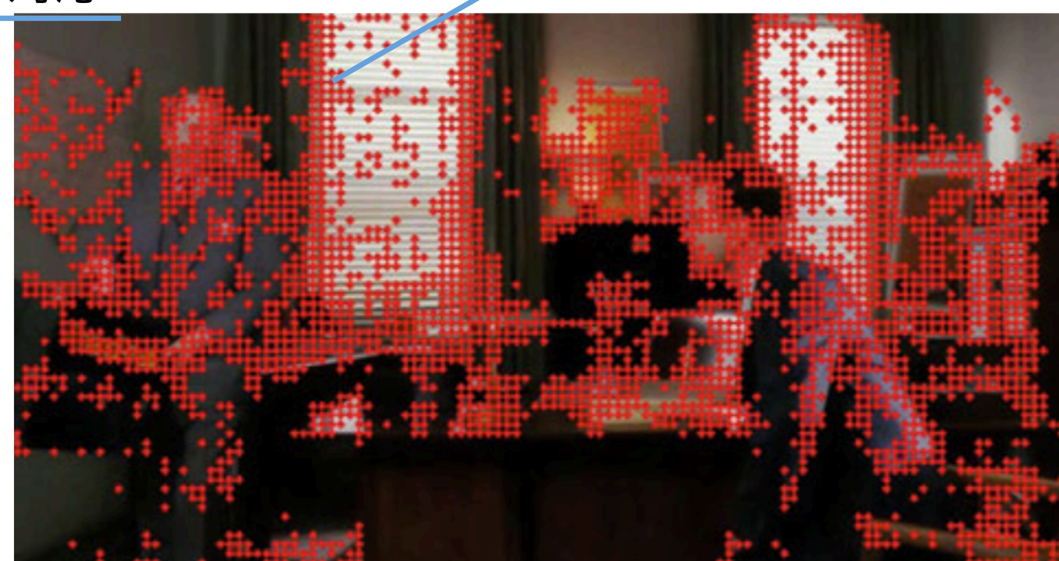
スケール対応(ピラミッド構造)と密な記述(サンプリング)

- ピラミッド構造(多段階スケーリング)によりスケールの違いに対応
  - スケールは1/√2 ずつ小さくする
- 特徴点は各スケールにおいてサンプリング
  - 格子幅は実験的に5x5[pxls]と設定
  - コーナー $T = 0.001 \times \max_{i \in I} \min(\lambda_i^1, \lambda_i^2), \qquad (1)$

拡大・縮小に対応



詳細なサンプリング

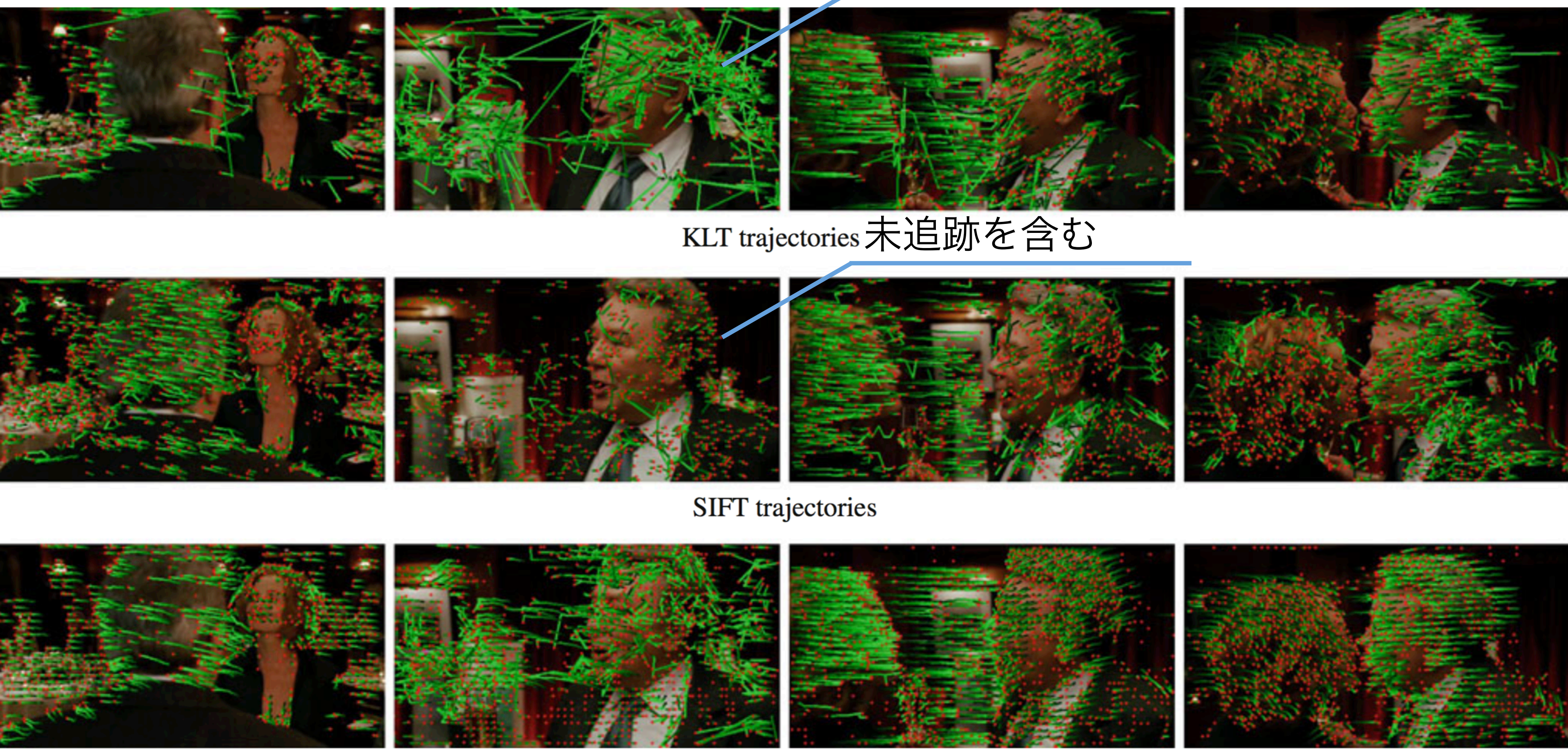


Farneback Optical Flowによるフロー抽出

密なオプティカルフロー

- Farneback Optical FlowはOpenCVにも実装済
- KLT法(上), SIFT(中)と比較

ノイズを多数含む



KLT trajectories

未追跡を含む

SIFT trajectories

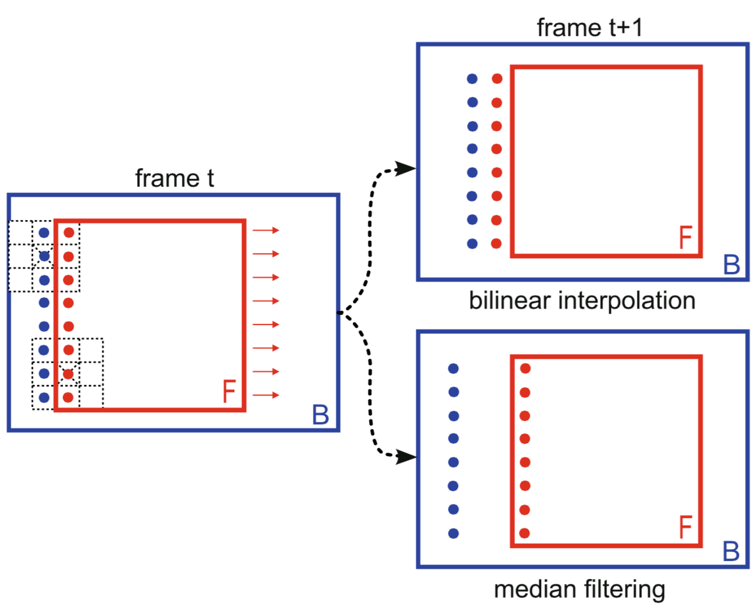
Dense trajectories

Median Filterによるフロー対応付け

フレーム t と t+1 のフロー対応付け

- Bilinear補間では物体と背景の境界にブラーが発生するため，フローの対応付けに失敗する場面も多く存在する
- Median FilterではBilinear補間よりも境界が鮮明になり，失敗回数が少ないことが明らかにされた
- (x,y): 画像上の座標 M: medianのカーネル ω: オプティカルフロー

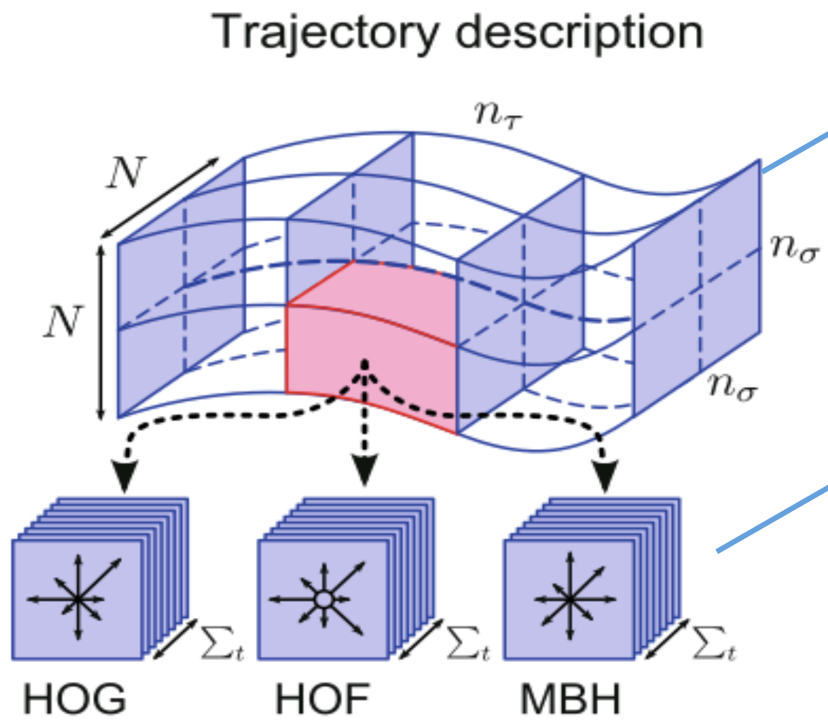
$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega_t)|_{(x_t, y_t)}, \quad (2)$$



特徴記述

動線を中心とした局所領域から特徴記述

- 追跡された動線 (Tフレーム連結)の周辺領域を抽出
- 周辺領域から局所特徴量HOG, HOF, MBHと，フロー形状を取得
- 下図のように2(x方向)x2(y方向)x3(t方向)の領域から取得した特徴を連結
- Bag-of-words(BoW)により特徴ベクトルを算出する



フローの中心から局所領域を取得

複数の局所特徴量を抽出

HOG特徴量

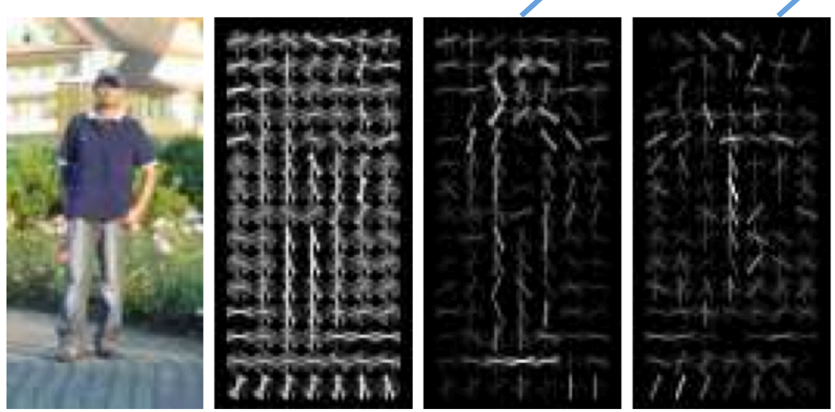
Histograms of Oriented Gradients (HOG)

- 物体のおおまかな形状を表現可能
- 局所領域をブロック分割して特徴取得
- エッジ勾配(下式g(x,y))により量子化ヒストグラム作成
- 勾配毎のエッジ強度(下式m(x,y))を累積

$$
\begin{aligned}
m(x,y) &= \sqrt{f_x(x,y)^2 + f_y(x,y)^2} \\
g(x,y) &= tan^{-1}\frac{f_y(x,y)}{f_x(x,y)} \\
f_x(x,y) &= I(x+1,y) - I(x-1,y) \\
f_y(x,y) &= I(x,y+1) - I(x,y-1)
\end{aligned}
$$

歩行者画像から取得した形状

背景から取得した形状

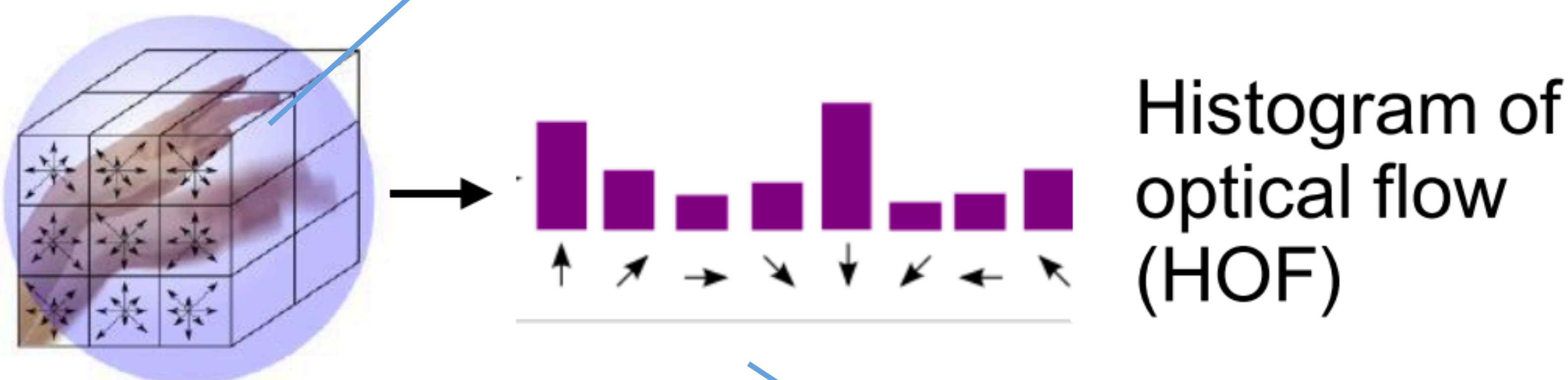


HOF特徴量

Histograms of Optical Flow (HOF)

- 局所領域をブロック毎に分割
- 前後フレーム(tとt+1)のフローをブロックごとに記述
- フロー方向と強度(長さ)

前後2フレームからフローを算出



動作ベースの特徴ベクトルを取得

MBH特徴量

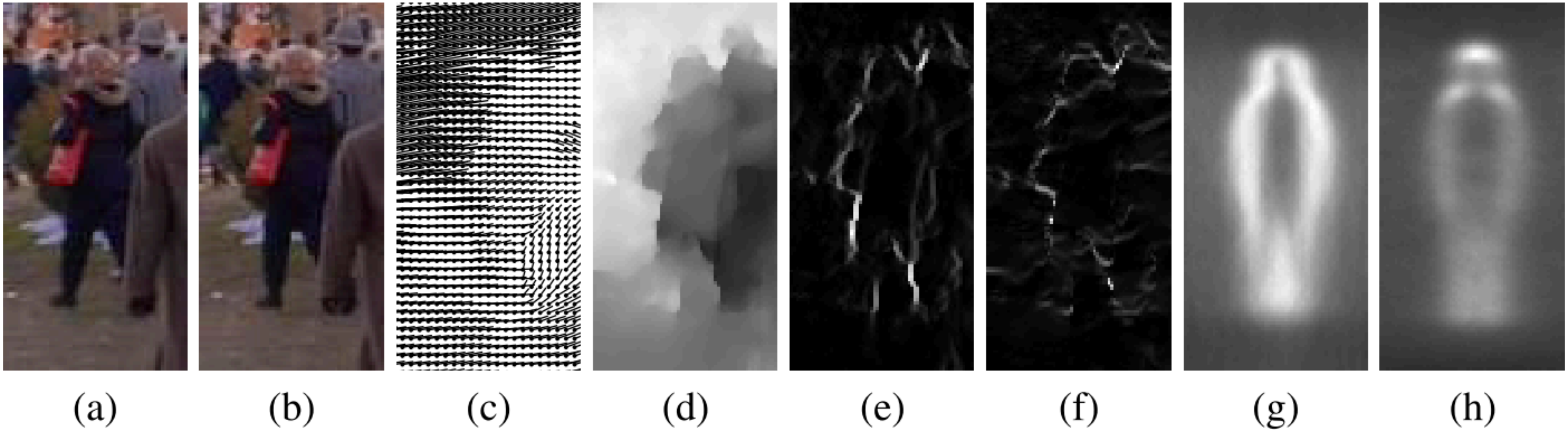
Motion Boundary Histograms (MBH)

– フレームt, t+1から取得したフローをエッジとみなして特徴記述

– 基本的に記述方法はHOGと同じ (Dense Trajectoriesでも同じ次元数を取得)

– MBHはフロー空間をx,y方向に微分した画像から特徴取得

=>x方向微分のMBH_x, y方向微分のMBH_yが取得される



(a)(b) t, t+1の元画像 (c) フロー空間 (d) フローの強度 (e)(f) x,y方向の勾配
(g)(h) x,y方向のMBH特徴

Trajectory特徴量

動線の形状を特徴量として扱う

- フレーム間のフローを算出
- 全体のフローの大きさで正規化

$$T = \frac{(\Delta P_t, \ldots, \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} ||\Delta P_j||}. \tag{3}$$

$$P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$$

Improved Dense Trajectories

考案者

– Heng Wang (発表当時INRIA所属、現Amazon)

– ICCV2013 (Journal: IJCV2015)

手法

– Dense Trajectoriesの改良

– 背景フローの除去、コードワード改良、人物検出を追加

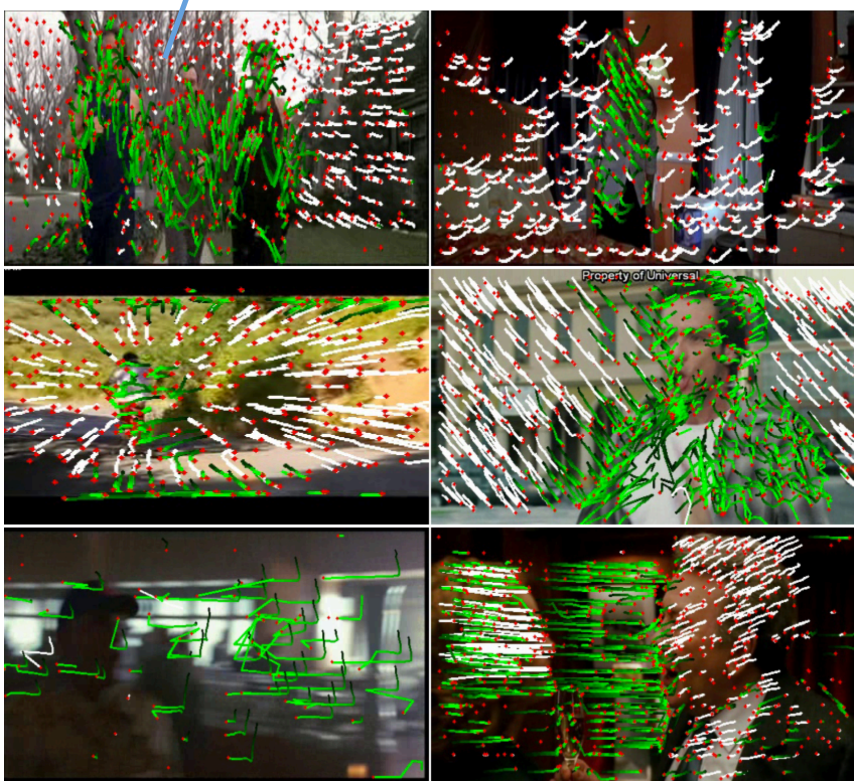
Wang, H. and Schmid, C. “Action Recognition with Improved Trajectories,” International Conference on Computer Vision (ICCV), pp.3551–3558, 2013.

Improved Dense Trajectories [H. Wang, ICCV2013]

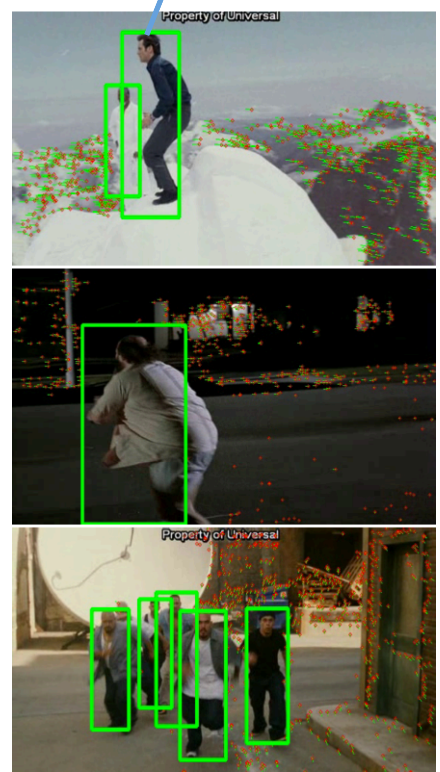
Dense Trajectoriesから3点の追加

- SURF + RANSACを用いたカメラ動作推定
- 人物検出枠によるROI(Region of interest)指定
- Fisher Vector (FV)の導入

カメラ動作以外が前景動作



行動認識では人物領域に限定



THUMOS@ICCV’13

Improved DTはワークショップTHUMOSで優勝

- THUMOS: The First International Workshop on Action Recognition with a Large Number of Classes, in conjunction with ICCV '13
- UCF50をさらに拡張したUCF101(101クラス認識)にて認識率を評価
- INRIAの研究グループはImproved Dense Trajectoriesを用いて85.9%の認識率を達成

UCF101, HMDB51における精度

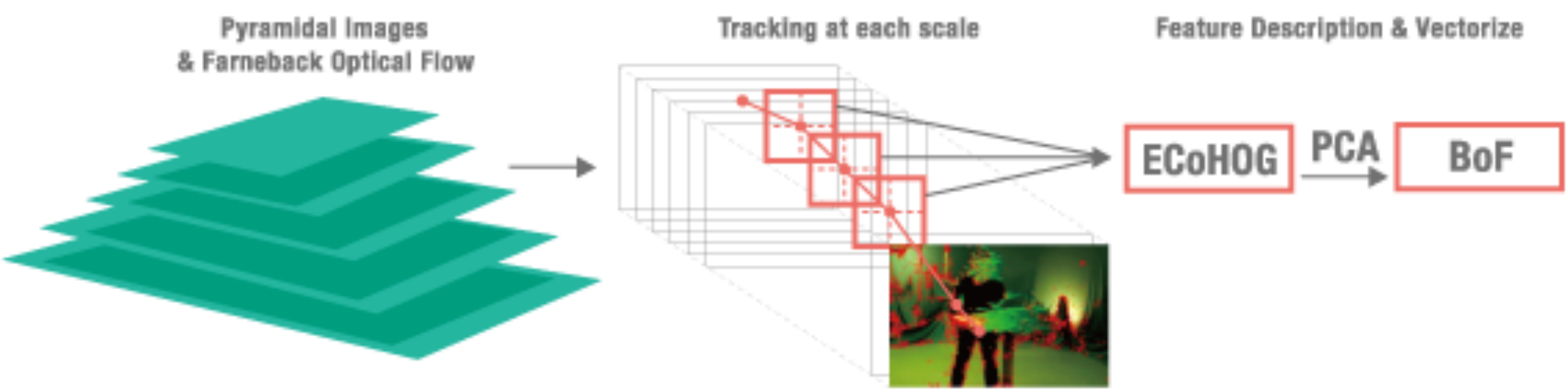
Approach	UCF101 (mAP)	HMDB51 (mAP)
STIP	43.9	20.0
DT	-	46.6
IDT	85.9	57.2

共起特徴量の導入

[H. Kataoka, ACCV2014]

– 共起特徴量をDense Trajectoriesに適用

– 「エッジペアカウント」や「エッジ強度ペア累積」により特徴表現



Kataoka, H., et al., "Extended Co-occurrence HOG with Dense Trajectories for Fine-grained Activity Recognition", in ACCV2014.

共起特徴量：CoHOGとECoHOG

「エッジペアのカウント」と「エッジ強度ペアの累積」

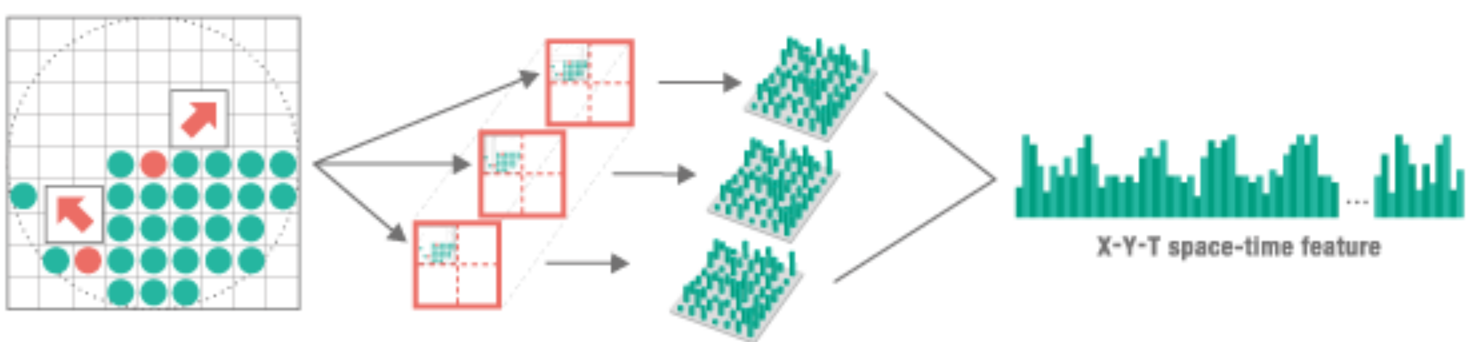
CoHOG

$$C_{x,y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m} \begin{cases} 1, & \text{if } d(p,q) = i \text{ and } d(x+p, y+q) = j \\ 0 & \text{otherwise} \end{cases}$$

Extended CoHOG

$$C_{x,y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m} \begin{cases} \|g_1(p,q)\| + \|g_2(p+x, q+y)\| \\ \text{if } d(p,q) = i \text{ and } d(p+x, q+y) = j \\ 0 \quad \text{otherwise} \end{cases}$$

– PCA による次元圧縮: 数千次元から数十次元オーダへ



パラメータ設定

PCAの次元数と共起特徴量取得のウインドウサイズ

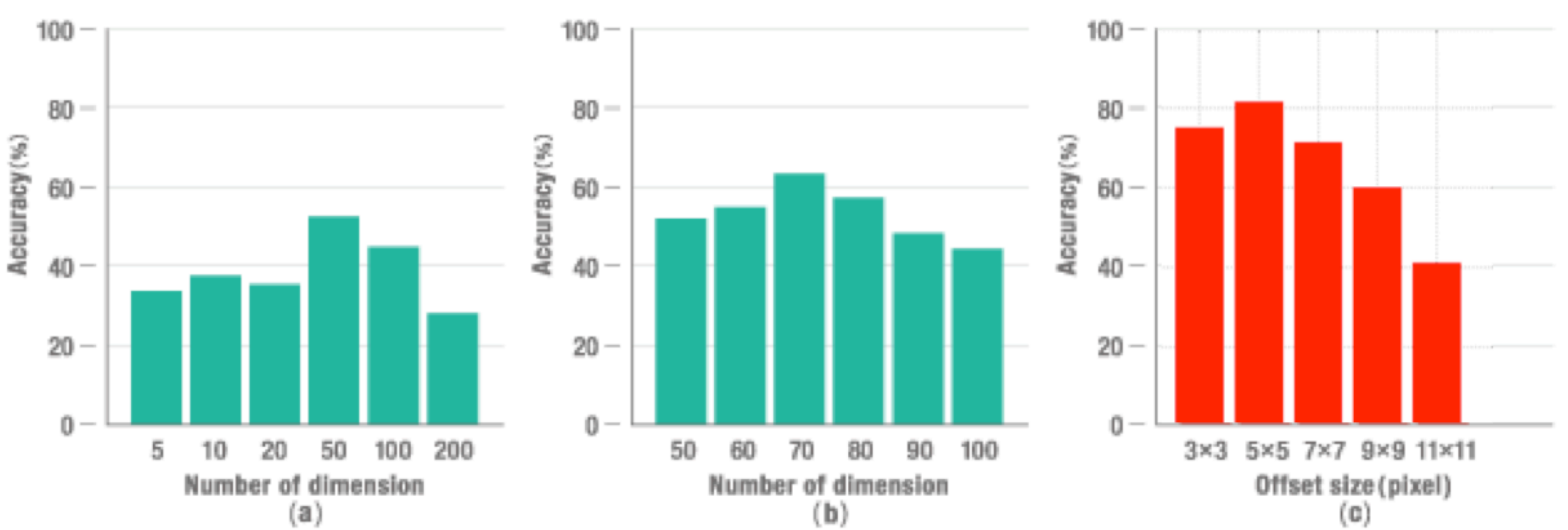
(a) PCA [dimensions] – 5, 10, 20, 50, 100, 200

(b) PCA [dimensions] – 50, 60, 70, 80, 90, 100

(c) Size of edge extraction window [pixels] – 3x3, 5x5, 7x7, 9x9, 11x11

– PCA: 70次元 => 寄与率と特徴空間サイズのバランス

– ウインドウサイズ: 5x5 pixels => 隣接画素の類似性を考慮(次ページ)

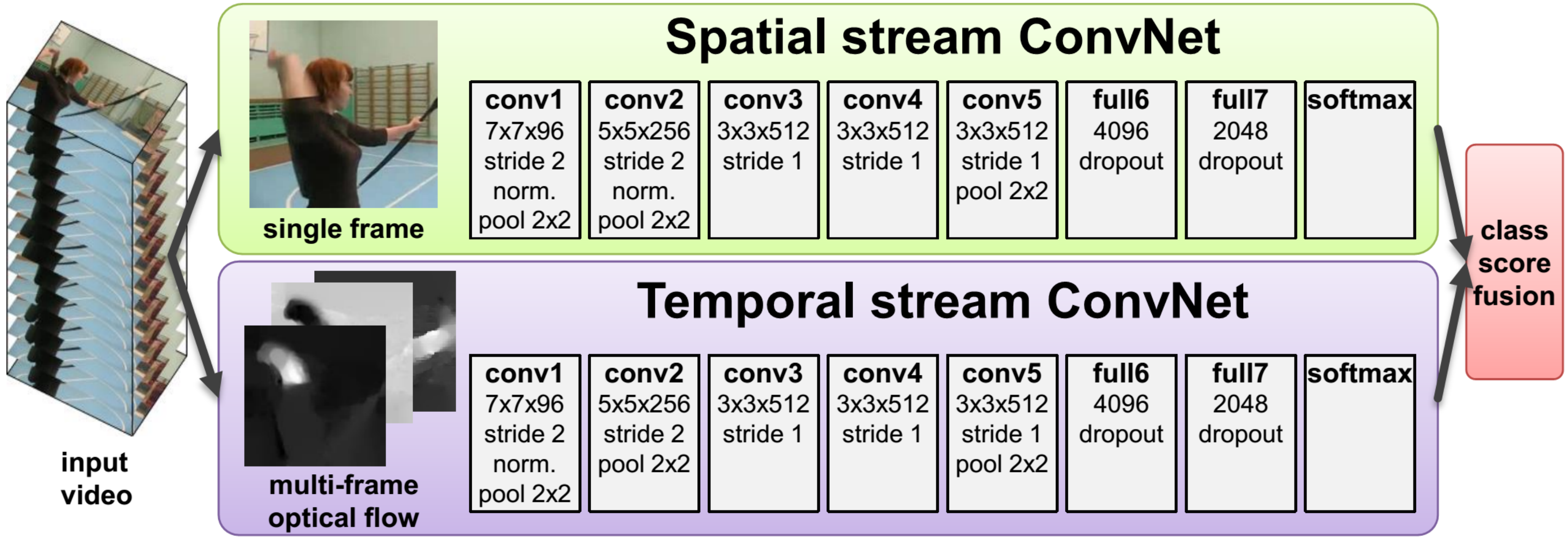


Two-stream CNN

Two-stream CNN

深層学習(Deep Learning)時代の動画認識手法

- 空間情報と時間情報の畳み込みにより精度を向上



Two-stream CNNの基本的な情報

考案者

- Karen Simonyan (発表当時Oxford所属、現Deep Mind)
- NIPS2014

手法

- RGB画像のみでなく、時間情報を画像に投影したフロー画像に対してCNN!

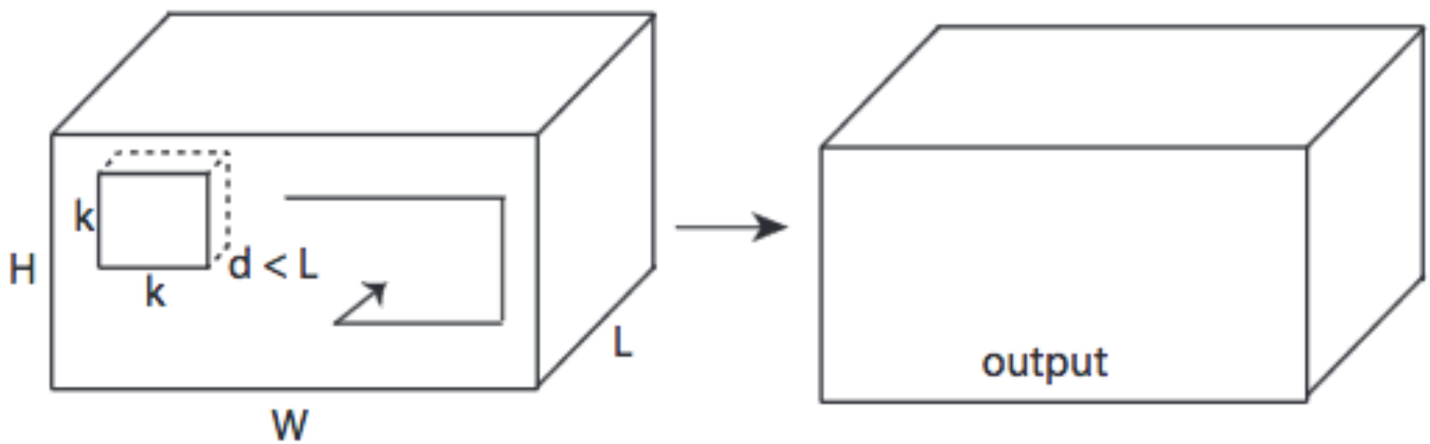
（XYTの3次元畳み込みで苦労していた（している））

Simonyan, K. and Zisserman, A. “Two-Stream Convolutional Networks for Action Recognition in Videos,” Neural Information Processing Systems (NIPS), 2014.

3次元の畳み込みについて

実際には (Two-Stream CNN) > (C3D: Spatiotemporal 3DCNN)

- 学習画像が足りない：2次元画像は1M@ImageNet、時系列画像は学習画像数の桁を上げないといけない？
- XYとTは性質が異なる：単純なXYTのカーネルではダメ？



(c) 3D convolution

Tran, D., Bourdev, L. , Fergus, R., Torresani, L. and M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks“, ICCV 2015.

Two-stream CNN [Simonyan+, NIPS14]

空間的・時間的な畳み込みの結果を統合

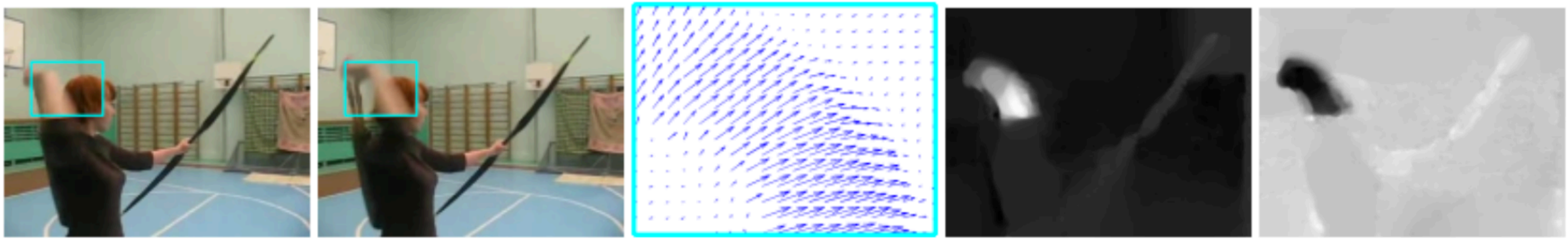
- 空間 (Spatial-stream)：RGBの入力から識別結果を出力
- 時間 (Temporal-stream)：Flow蓄積画像の入力から識別結果を出力、時間情報を予め２次元画像に投影
- 統合：クラスごとの確率分布

RGBの入力


Spatial stream ConvNet
single frame
conv1
7x7x96
stride 2
norm.
pool 2x2
conv2
5x5x256
stride 2
norm.
pool 2x2
conv3
3x3x512
stride 1
conv4
3x3x512
stride 1
conv5
3x3x512
stride 1
pool 2x2
full6
4096
dropout
full7
2048
dropout
softmax

Stacked Flowの入力

(x, y)方向に分割、フローを画像に投影



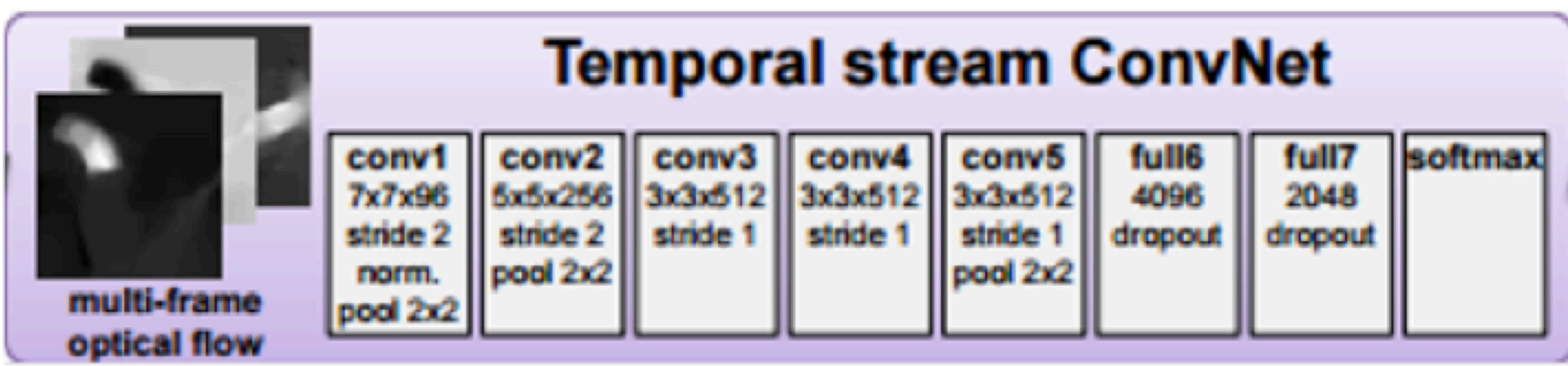
(a) (b) (c) (d) (e)

$$I_\tau(u, v, 2k-1) = d^x_{\tau+k-1}(u, v), \tag{1}$$
$$I_\tau(u, v, 2k) = d^y_{\tau+k-1}(u, v), \quad u = [1; w], v = [1; h], k = [1; L].$$

Stacked Flowの入力

X方向

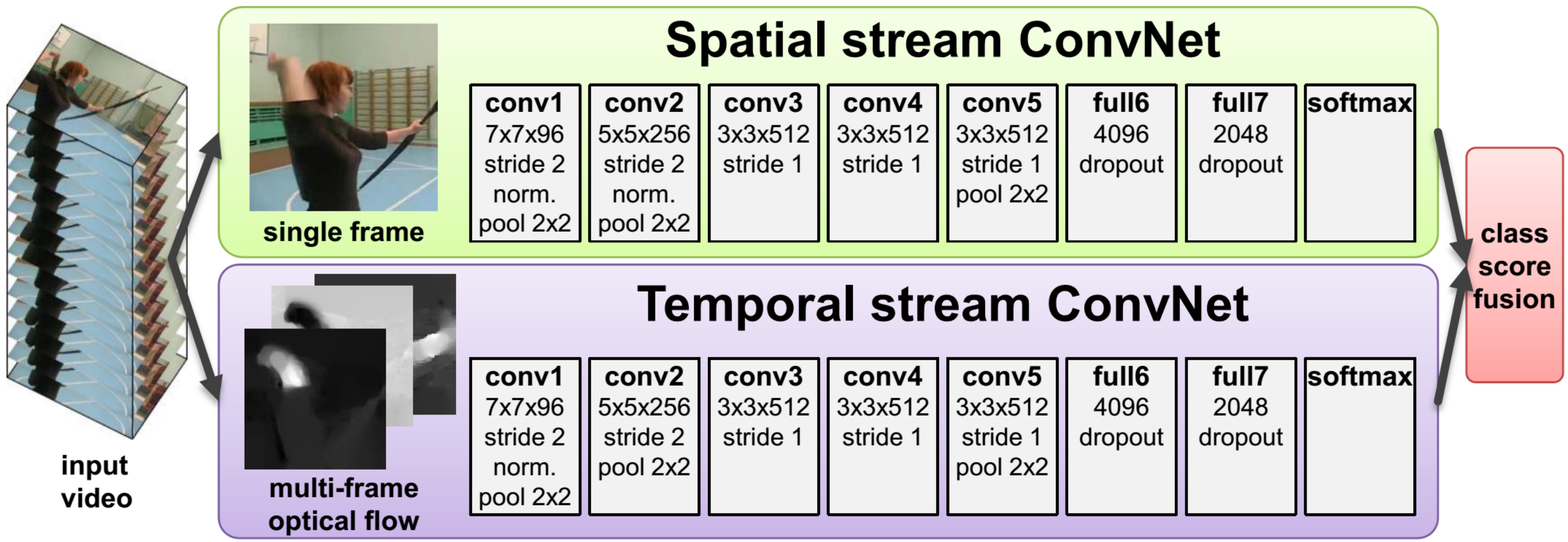
Y方向



スコアの統合

RGB + Stacked Flow

– スコアを (i) 単純平均 (ii) SVMの出力値 により統合



UCF101, HMDB51における精度

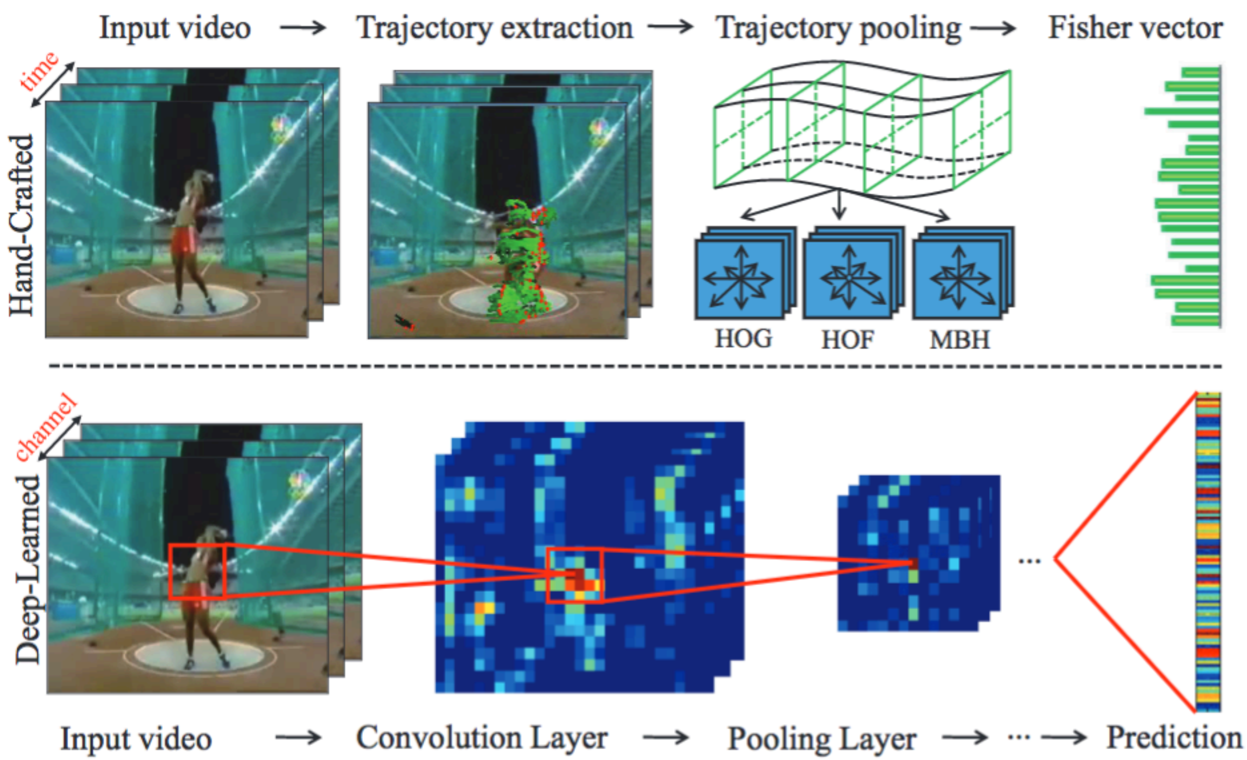
Approach	UCF101 (mAP)	HMDB51 (mAP)
STIP	43.9	20.0
DT	-	46.6
IDT	85.9	57.2
Two-stream	88.0	59.4

Traj.-pooled Deep-conv. Desc. (TDD)

TDD

ハンドクラフトと深層学習の統合

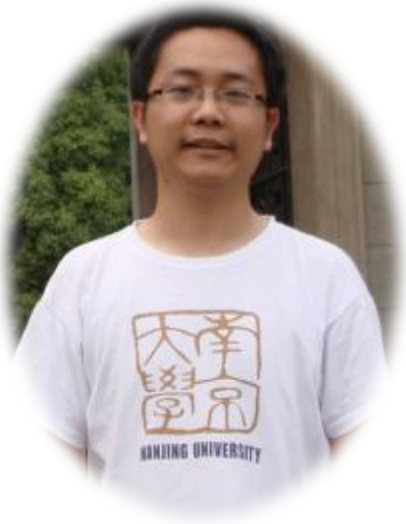
– IDTを用いるが、特徴を畳み込みマップから抽出



TDDの基本的な情報

考案者

- Limin Wang (発表当時CUHK所属、現ETHZ)
- CVPR2015



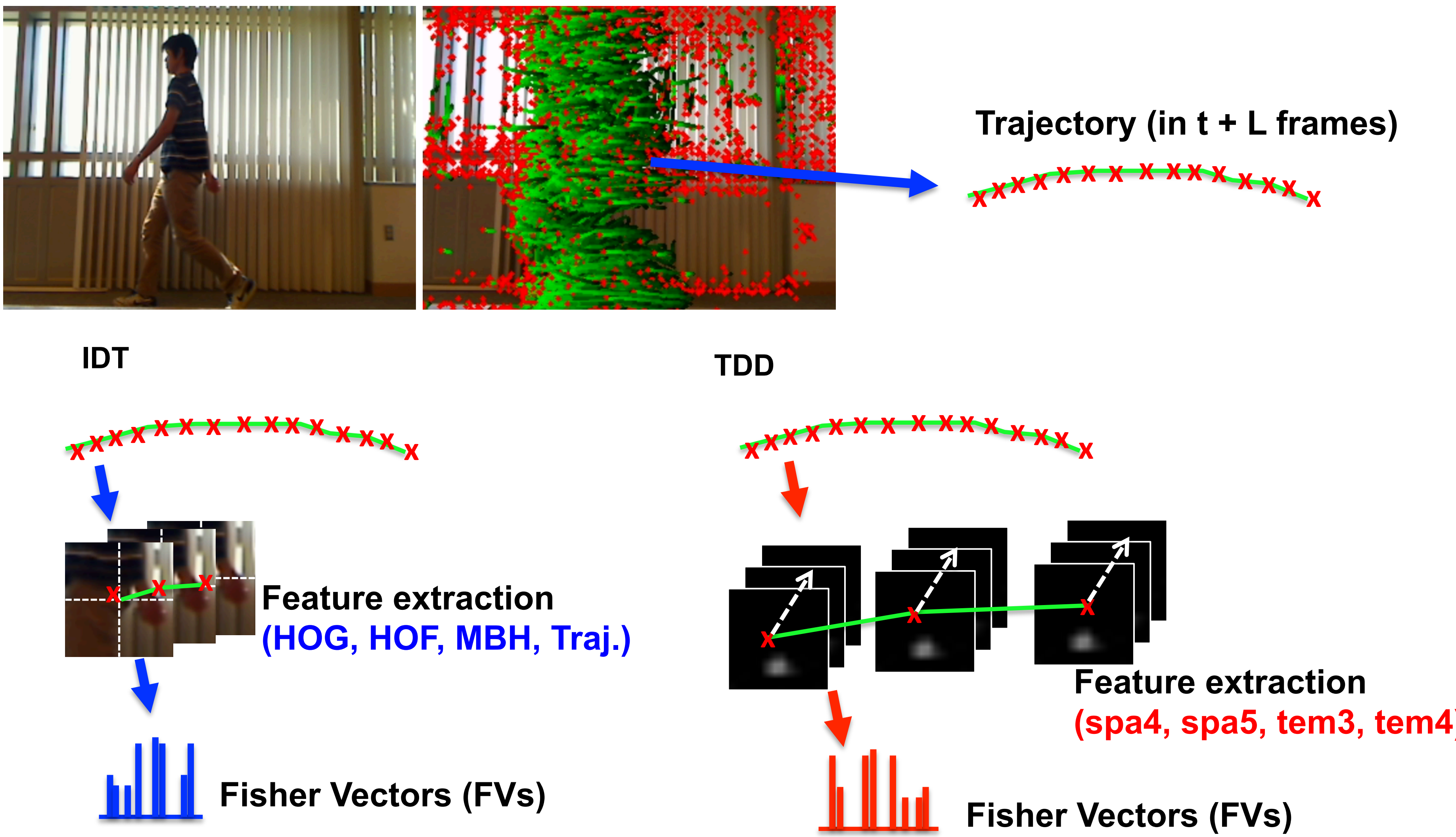
手法

- IDTのフローの特徴記述を畳み込みマップで置き換え
- ハンドクラフトと深層学習のいいとこ取りを実現

Wang, L., Qiao, Y. and Tang, X. "Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

TDDのフレームワーク

TDDとIDTの比較



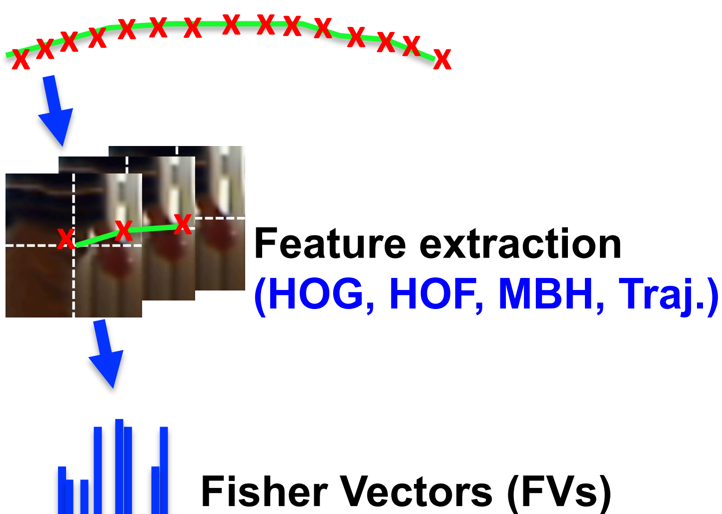
もう少し詳しく

特徴抽出が異なる

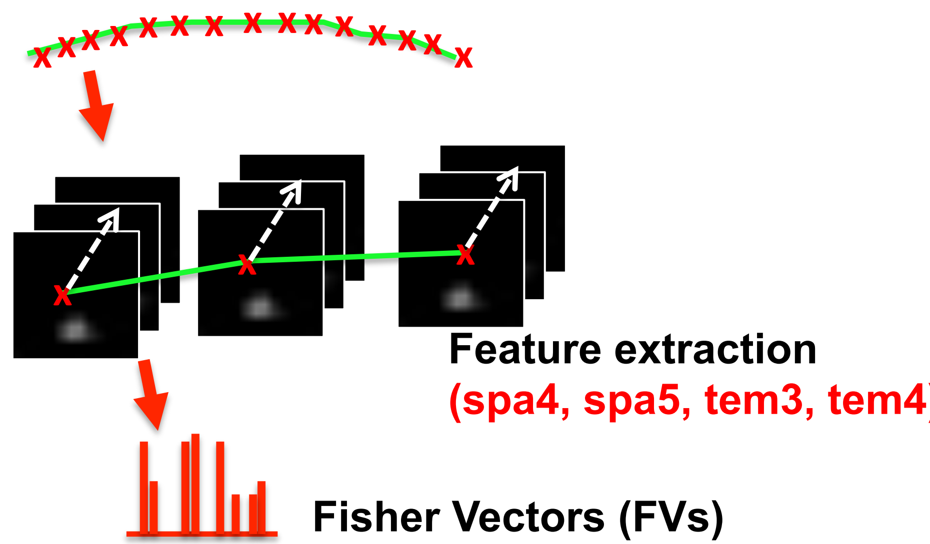
- IDT：ハンドクラフト特徴の抽出
  - サンプリング点(下図のx)の周囲から局所特徴を抽出
- TDD：畳み込みマップから値を抽出
  - サンプリング点からチャネル方向に値を抽出
  - “特徴次元数” = “チャネル数”

$$D(T_k, \widetilde{C}^a_m) = \sum_{p=1}^{P} \widetilde{C}^a_m(\overline{(r_m \times x_p^k)}, \overline{(r_m \times y_p^k)}, z_p^k),$$

IDT



TDD



UCF101, HMDB51における精度

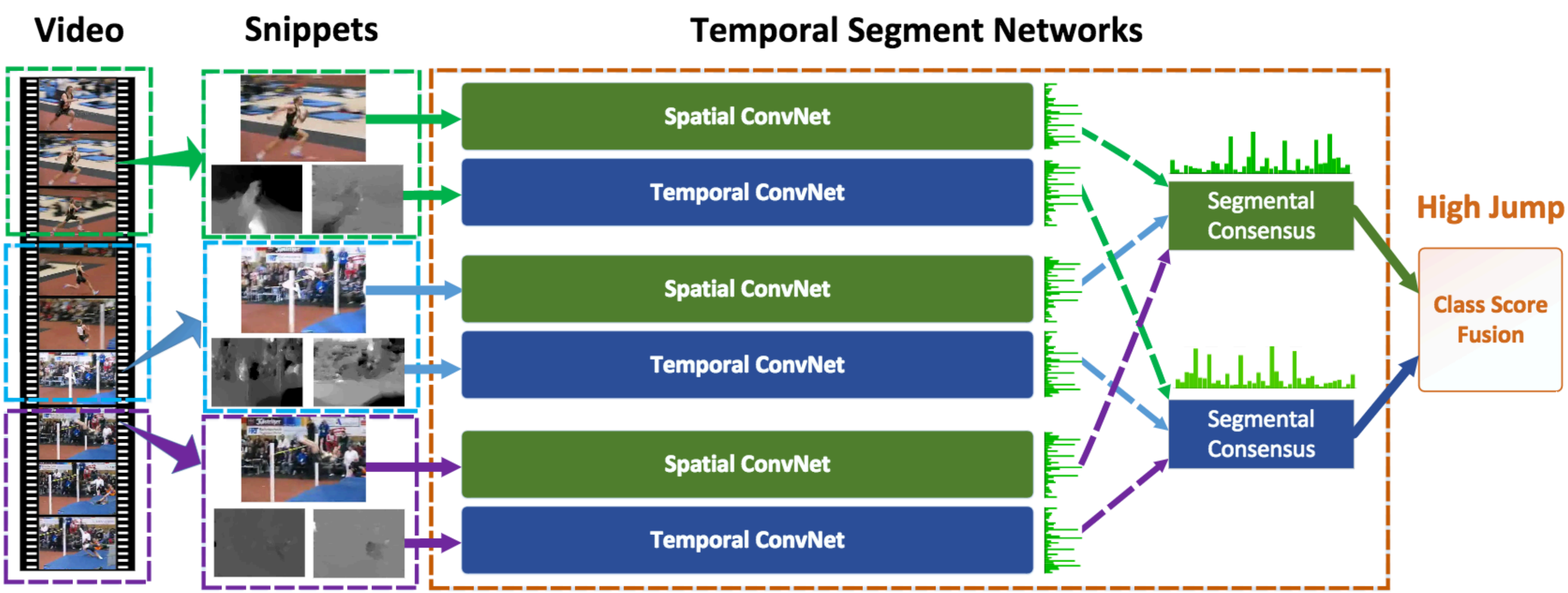
Approach	UCF101 (mAP)	HMDB51 (mAP)
STIP	43.9	20.0
DT	-	46.6
IDT	85.9	57.2
Two-stream	88.0	59.4
TDD	90.3	63.2
TDD+IDT	91.5	65.9

最新のアプローチ
*2017*年*1*月現在

Temporal Segment Network (TSN)

動画を分割したサブ動画毎の推定と統合

- フレーム単位ではなく、動画単位で学習した方が良い（目的に沿った学習）
- RGB、RGB差分、フロー画像、ワープフロー画像を入力してCNN
- 学習の初期値やBNのパラメータも実験的に決定



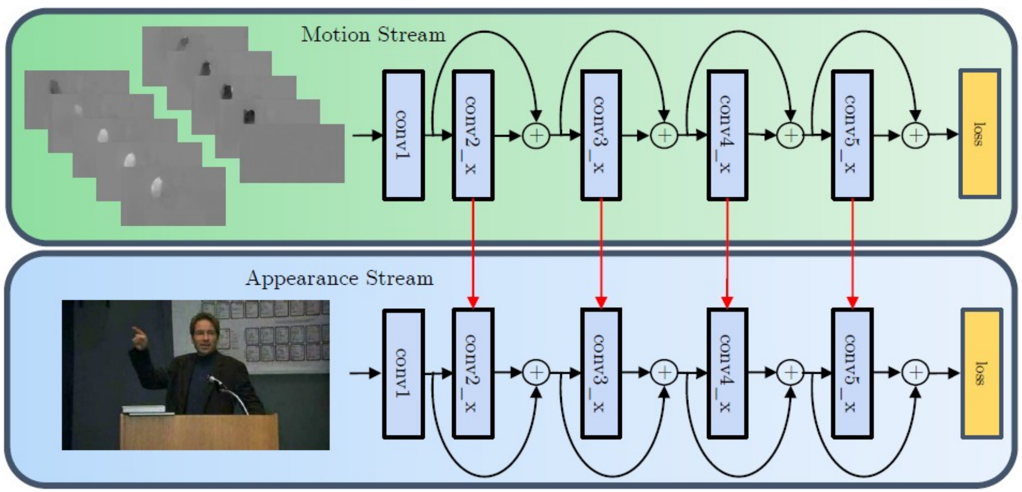
UCF101, HMDB51における精度

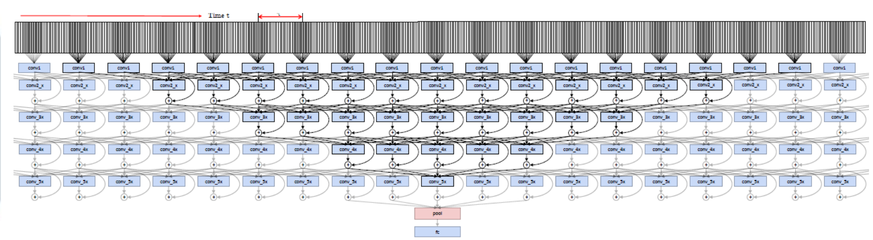
Approach	UCF101 (mAP)	HMDB51 (mAP)
STIP	43.9	20.0
DT	-	46.6
IDT	85.9	57.2
Two-stream	88.0	59.4
TDD	90.3	63.2
TDD+IDT	91.5	65.9
TSN	94.2	69.4

ST-ResNet

Two-Stream CNNをResNetで構成し，両者の重みを共有

- TemporalNetからSpatialNetに特徴統合しながら最適化
- 時間方向の画像を入力として扱う
- 最終的に重みが統合されたST-ResNetにて動画認識





UCF101, HMDB51における精度

Approach	UCF101 (mAP)	HMDB51 (mAP)
STIP	43.9	20.0
DT	-	46.6
IDT	85.9	57.2
Two-stream	88.0	59.4
TDD	90.3	63.2
TDD+IDT	91.5	65.9
TSN	94.2	69.4
Two-stream ResNet	93.4	66.4
Two-stream ResNet+IDT	94.6	70.3

Human Action Recognition without Human
(ECCV2016WS Oral&BestPaper)

最近のDBでは背景が効いているんじゃ？

Two-stream CNNでもRGBの入力

– UCF101, HMDB51などは人物領域と比較して背景領域が大きい

– RGBを入力とした空間情報のみを用いて高い識別を実現

  • Two-stream CNNのspatial-streamだけでも70%強の識別率@UCF101

“Human Action Recognition without Human”の提案

（人を見ない人物行動認識）

Y. He, S. Shirakabe, Y. Satoh, H. Kataoka “Human Action Recognition without Human”, in ECCV 2016 Workshop on on Brave New Ideas for Motion Representations in Videos (BNMW). (Oral & Best Paper)

賀雲, 白壁奏馬, 佐藤雄隆, 片岡裕雄, “人を見ない人物行動認識”, ViEW, 2016 (ViEW若手奨励賞)

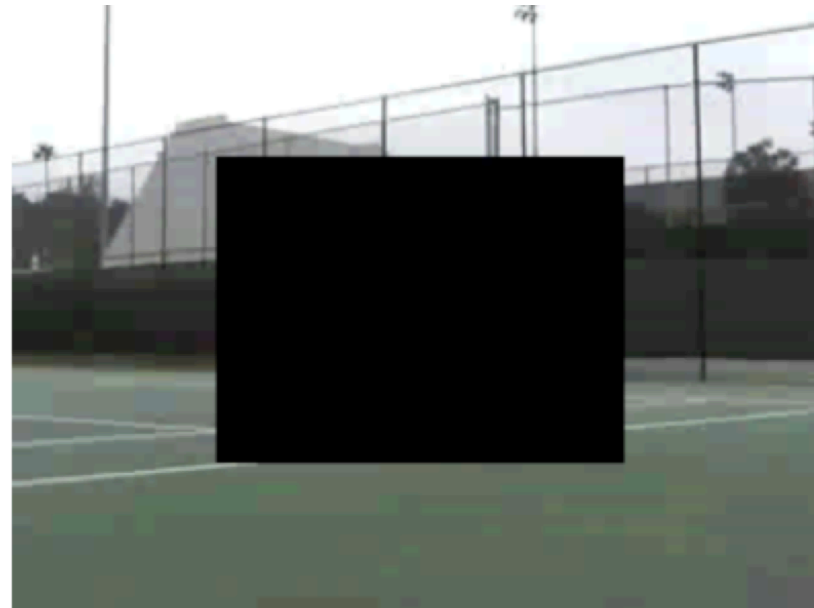
Without Human?

現在の問題設定において人物行動認識は背景領域だけあれば十分なのではないだろうか？

Motion Descriptor

↓

Tennis Swing



Motion Descriptor

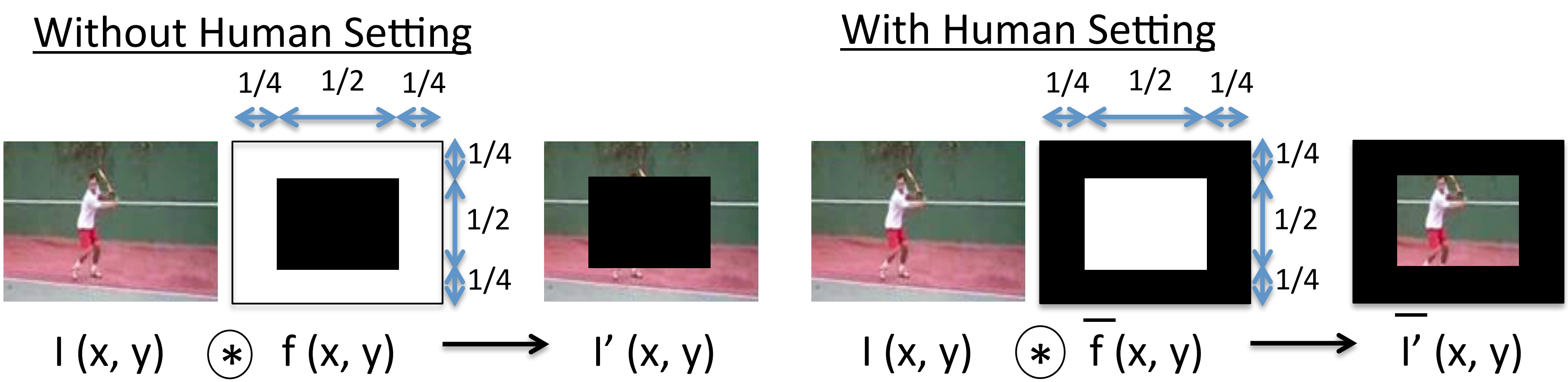
↓

Tennis Swing?

w/ and w/o Human Setting

With / Without human setting

- *Without* human setting: 中央部分が黒抜き
- *With* human setting: Without human settingのインバース



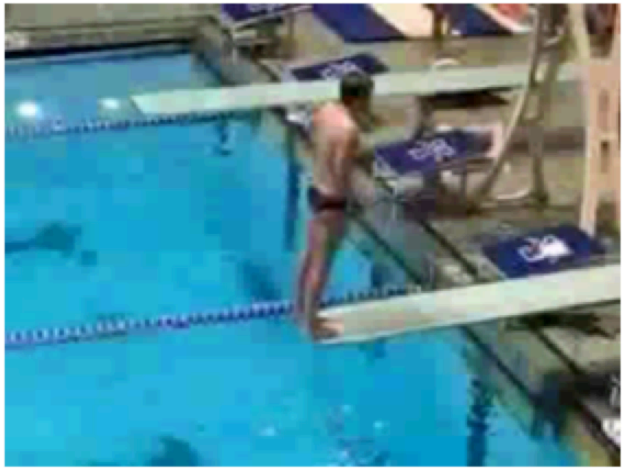
実験の設定

- ベースライン: Very deep two-stream CNN [Wang+, arXiv15]
- 二つの設定: without human and with human

Visual results (Full Image)

Answer: Bowling

Answer: TennisSwing

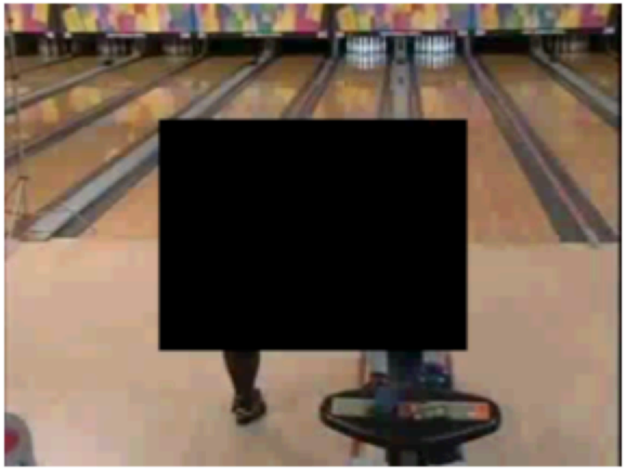


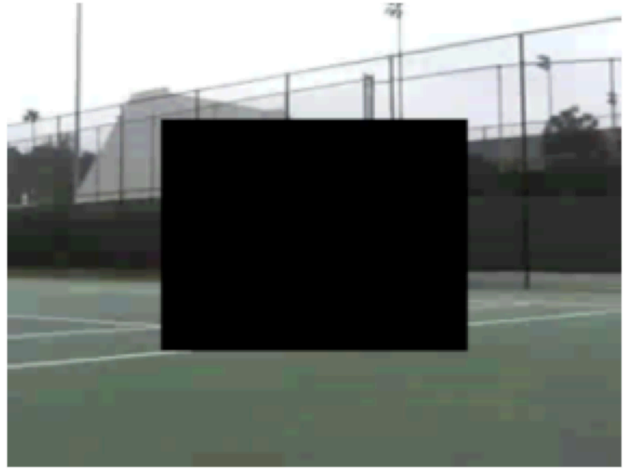
Answer: Diving

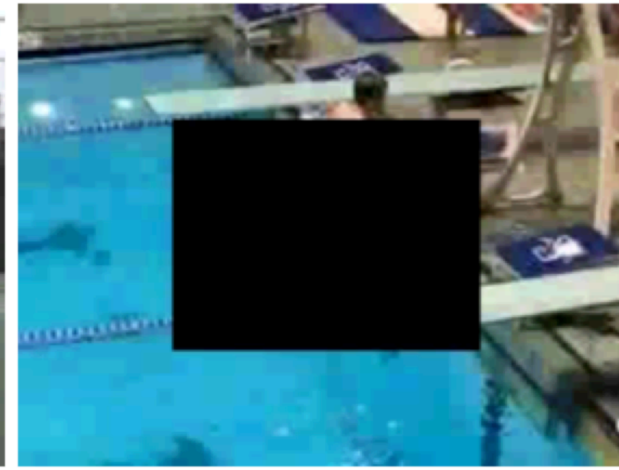
Answer: HorseRace

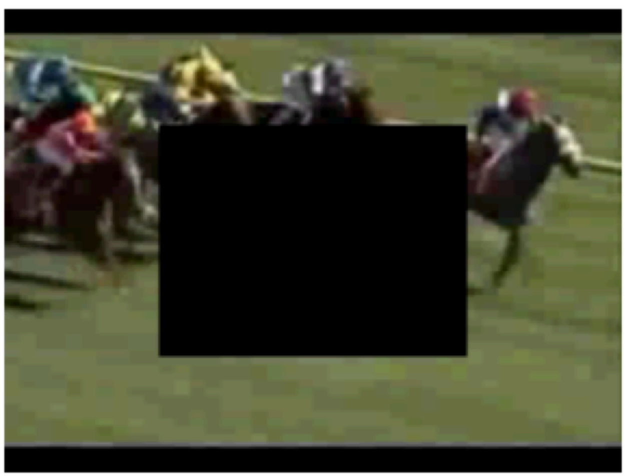
Answer: CliffDiving

Visual results (Without Human Setting)


Answer: Bowling


Answer: TennisSwing


Answer: Diving


Answer: HorseRace


Answer: CliffDiving

実験結果

@UCF101

- UCF101 pre-trained model with very deep two-stream CNN
- With/Without Human Setting

With or Without a Human	Stream	% on UCF101 (split 1)
With human	Spatial stream	51.26
	Temporal stream	40.50
	Two-stream	**56.91**
Without human	Spatial stream	45.33
	Temporal stream	26.80
	Two-stream	**47.42**

実験結果

With or Without a Human	Stream	% on UCF101 (split 1)
With human	Spatial stream	51.26
	Temporal stream	40.50
	Two-stream	**56.91**
Without human	Spatial stream	45.33
	Temporal stream	26.80
	Two-stream	**47.42**

9.49%しか差がない！

– 背景のみで約半分が認識できる

– 人を含んでいても背景領域を除去すると精度が落ちる

考察

実は，時系列CNNは純粋な人間のモーションを見ていない！

(CNNによるモーション認識では背景による寄与が大きい)

- WithとWithout Humanにおいて，両者の差が9.49%
  - 背景特徴の寄与が大きい事が判明
  - 人物特徴と背景特徴による識別率が逼迫
- Without HumanにおいてSpatial-streamの識別率が高い
  - Spatial-stream はTemporal-streamと比較して+18.53%
  - Spatial-streamはRGBによる，時系列特徴を用いていない

未来のモーション認識

洗練されたモーションを捉える事ができたら？

– 動画による教師無し学習の洗練？ [Vondrick+, CVPR16]

C. Vondrick et al. “Anticipating Visual Representations from Unlabeled Video”, in CVPR, 2016.

– 自然な動画生成 [Vondrick+, NIPS16]

C. Vondrick et al. “Generating Videos with Scene Dynamics”, in NIPS, 2016.

Transitional Action Recognition
(BMVC2016)

行動予測の問題点

早期行動認識

- 行動開始時点から観測開始，十分な手がかりが得られないと認識不可
- 行動認識とほぼ同じ？

行動予測

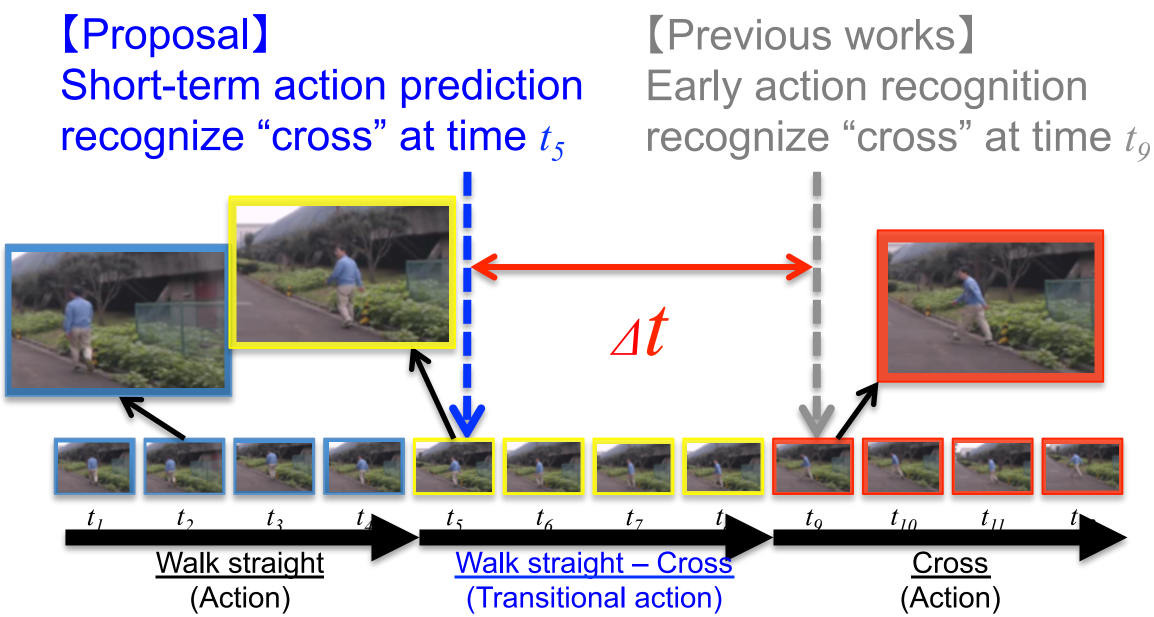
- 行動が生起する前から予測するので不安定

手法	設定
行動認識	$f(F^{A}_{1\ldots t}) \rightarrow A_t$
早期行動認識	$f(F^{A}_{1\ldots t-L}) \rightarrow A_t$
行動予測	$f(F^{A}_{1\ldots t}) \rightarrow A_{t+L}$

提案手法の問題設定

2つの行動間に遷移行動 (TA; Transitional Action)を挿入

- 予測のためのヒントがTAに含有: 早期行動認識より時間的に早く認識
- TAの認識が即ち次行動の予測: 行動予測より安定した予測



提案手法の問題設定

2つの行動間に遷移行動 (TA; Transitional Action)を挿入

- 予測のためのヒントがTAに含有: 早期行動認識より時間的に早く認識
- TAの認識が即ち次行動の予測: 行動予測より安定した予測

手法	設定
行動認識	$f(F^{A}_{1\ldots t}) \rightarrow A_t$
早期行動認識	$f(F^{A}_{1\ldots t-L}) \rightarrow A_t$
行動予測	$f(F^{A}_{1\ldots t}) \rightarrow A_{t+L}$
遷移行動認識	$f(F^{TA}_{1\ldots t}) \rightarrow A_{t+L}$

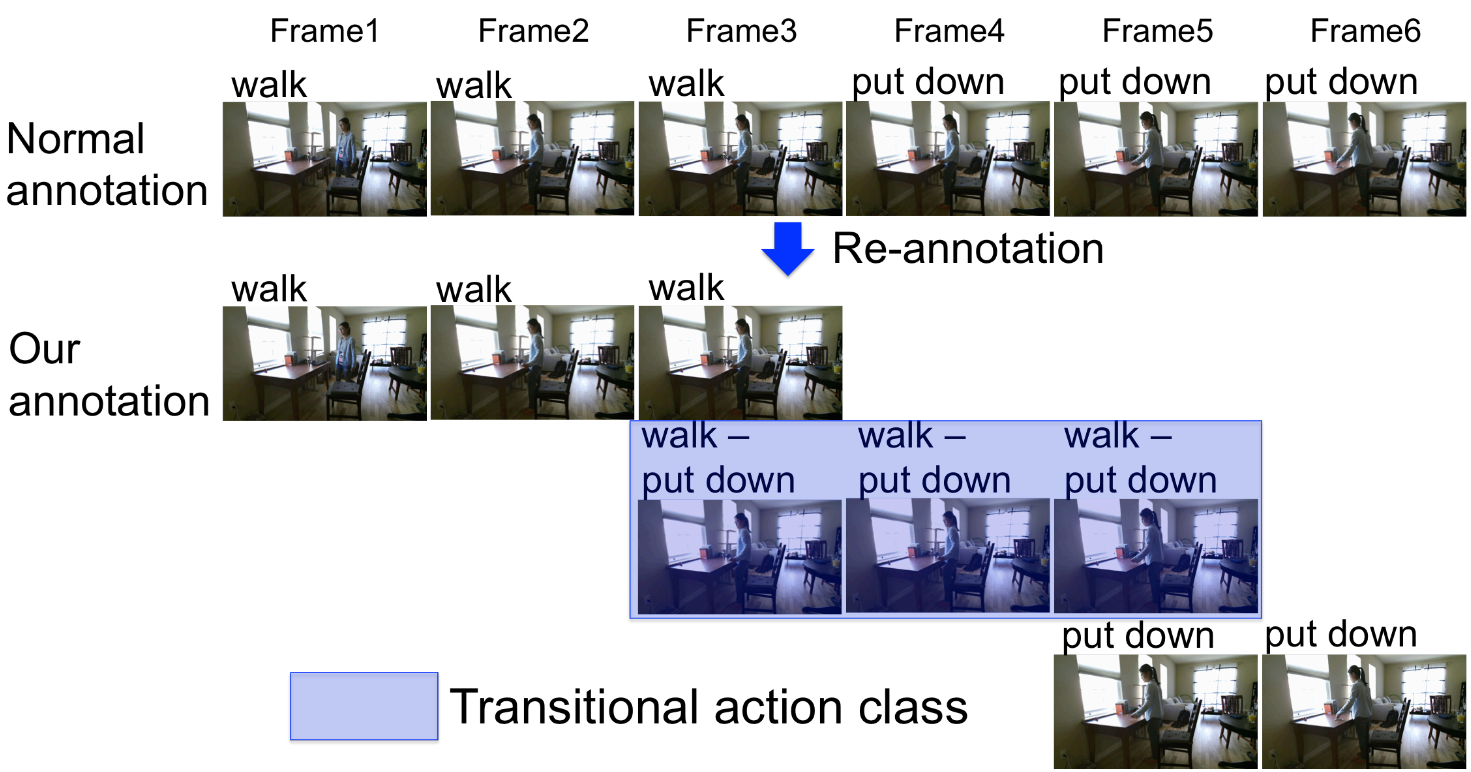
Transitional Action (TA)の詳細

TAのアノテーション

– 通常行動 (NA; Normal Action)に数枚オーバーラップしてラベル付

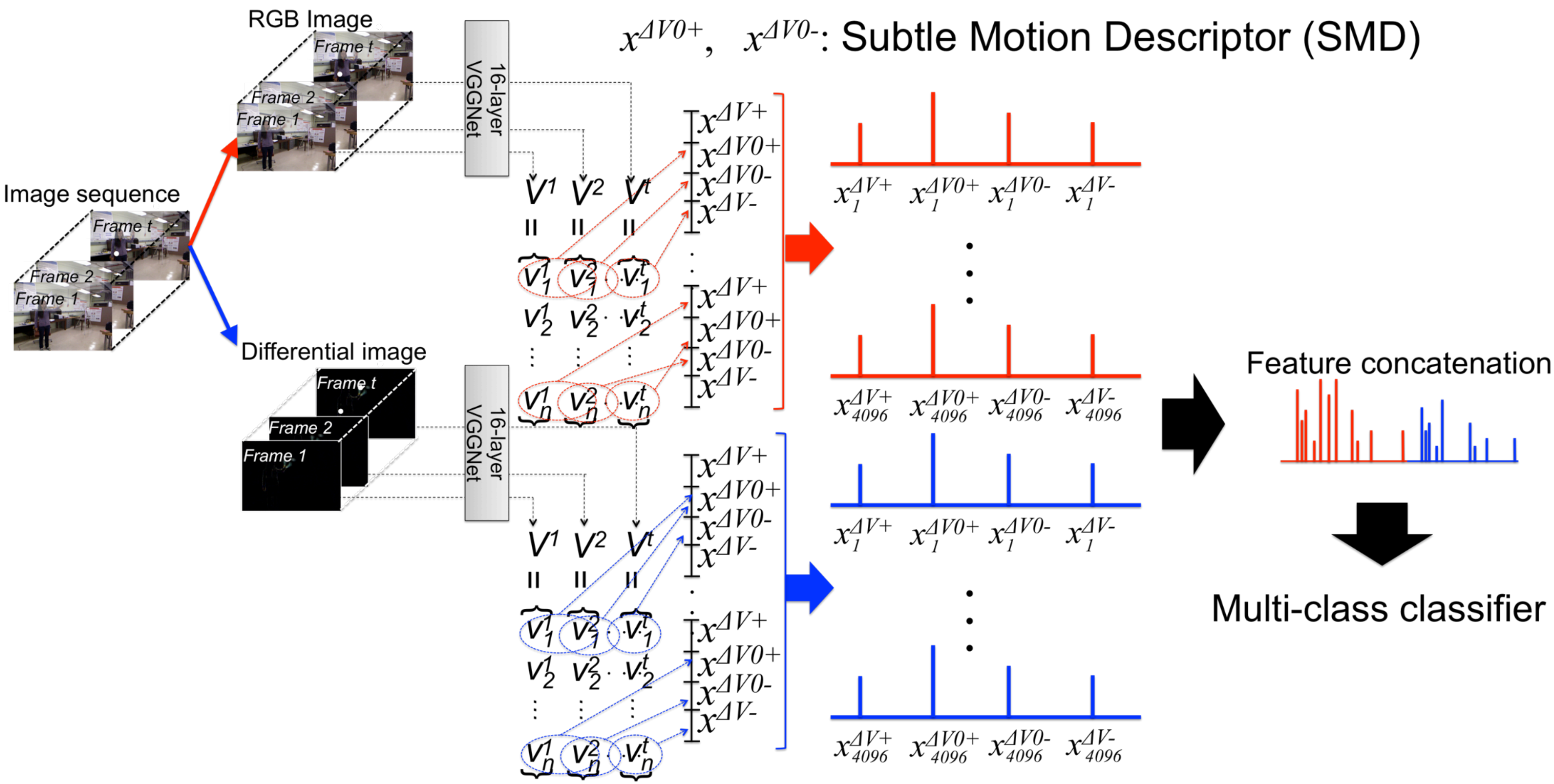
TA認識の困難さ

– NAとTAの行動特徴が混在



Subtle Motion Descriptor (SMD)の提案

NA~TAの変動時の行動を捉えるための詳細な特徴

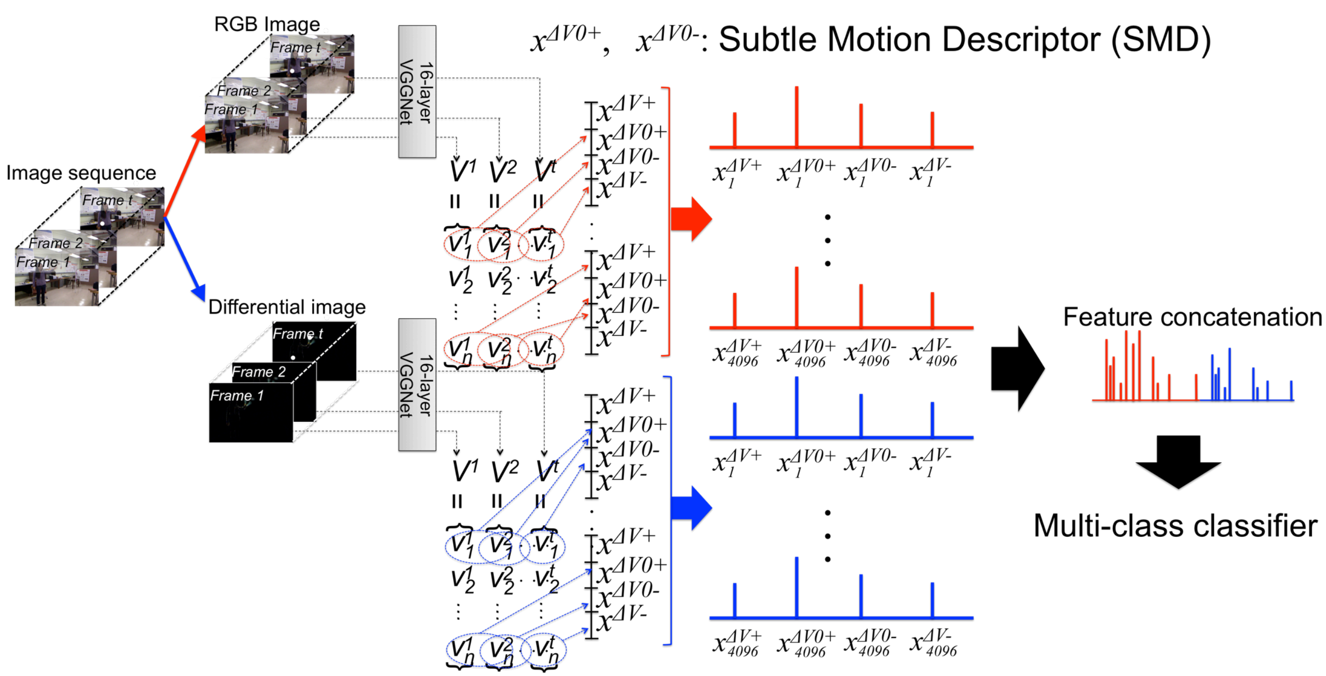


Subtle Motion Descriptor (SMD)

深層畳み込みNN (VGGNet-16)の中間層から特徴を取り出す

- 最初の全結合層
- $N = 4{,}096$

$$V^t = [v_1^t, v_2^t, \ldots, v_N^t] \tag{2}$$

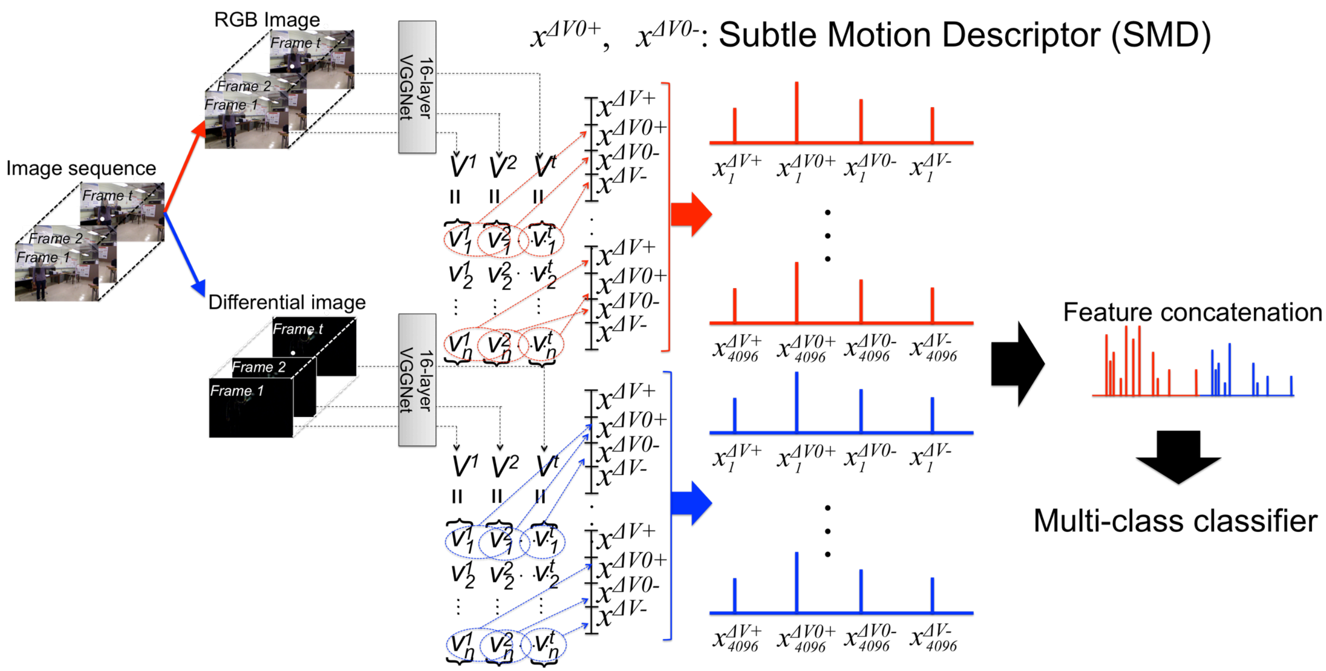


Subtle Motion Descriptor (SMD)

各要素の時系列差分ΔV^tを計算

- (Frame *t*) – (Frame *t-1*)

$$\Delta V^t = [(v_1^t - v_1^{t-1}), (v_2^t - v_2^{t-1}), \ldots, (v_N^t - v_N^{t-1})] \quad (3)$$



Subtle Motion Descriptor (SMD)

時系列差分ΔV^tの量子化ベクトル

- 正負にて分割
- 閾値(*TH*)を参照し，ゼロ付近の特徴を収集 (→特にこれをSMDと呼ぶ)
- *TH*は$TH = 0.05$と設定

$$x_i^{\Delta V^+} = \Sigma_{t=t_s}^{t_e} h_i^+(t), x_i^{\Delta V^{0^+}} = \Sigma_{t=t_s}^{t_e} h_i^{0^+}(t) \qquad (4)$$

$$x_i^{\Delta V^-} = \Sigma_{t=t_s}^{t_e} h_i^-(t), x_i^{\Delta V^{0^-}} = \Sigma_{t=t_s}^{t_e} h_i^{0^-}(t) \qquad (5)$$

where

$$\begin{cases} h_i^+(t) = |\Delta v_i^t| \ \ (\Delta v_i^t > TH) \\ h_i^{0^+}(t) = |\Delta v_i^t| \ \ (0 < \Delta v_i^t < TH) \\ h_i^{0^-}(t) = |\Delta v_i^t| \ \ (-TH < \Delta v_i^t < 0) \\ h_i^-(t) = |\Delta v_i^t| \ \ (\Delta v_i^t < -TH) \end{cases} \qquad (6)$$

実験の設定 (データセット)

連続行動が含まれるデータセットを準備

- NTSEL [Kataoka+, ITSC2015]
  - ヒヤリハットDBの切り抜き, walk/turn/cross等の行動，人物位置あり
- UTKinect-Action [Xia+, CVPRW2012]
  - 順序が決まっている10種のNA (walk, throw, sit等)
  - TAはその間の8種 (push/pullを除く; 次ページ)
  - 人物位置なし
- Watch-n-Patch [Wu+, CVPR2015]
  - 日常行動 10種 (read, turn on monitor, leave office等)
  - 遷移行動は上位10種 (次ページ)
  - 人物位置なし

実験の設定 (TAの詳細)

@UTKinect-Action @Watch-n-Patch

Table 1. Transitional action classes and sequence number of the UTKinect-Action dataset and Watch-n-Patch dataset.

Class name	Frequency	Class name	Frequency
Walk - Sit down	20	Walking - Put down item	99
Sit down - Stand up	20	take item - Leave office	93
Stand up - Pick up	20	Reading - Leave-office	77
Pick up - Carry	20	Turn on monitor - Play computer	57
Carry - Throw	18	fetch book - Reading	50
Throw - Push	18	Put down item - Reading	49
Pull - Wave hands	20	Play computer - Turn on monitor	41
Wave hands - Clap hands	20	Walking - Reading	38
		Put down item - Play computer	38
		Play computer - Turn off monitor	34

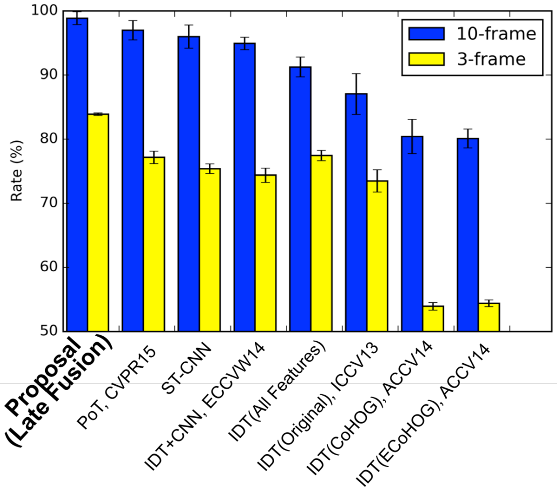
比較に用いた手法

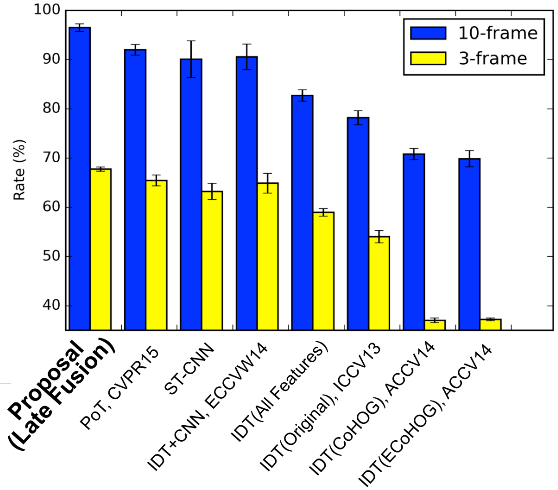
行動認識の最先端手法より

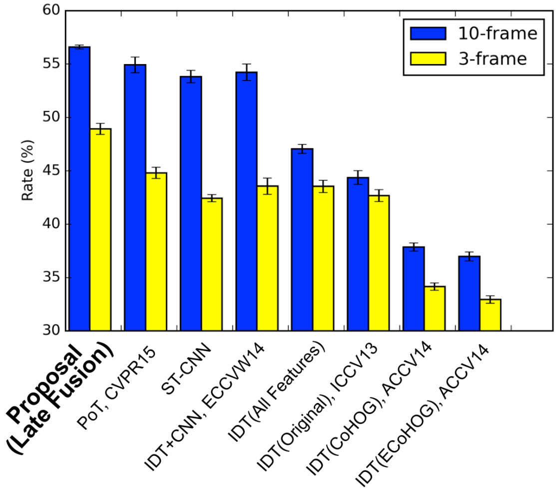
- 時系列CNN特徴
  - Pooled Time-series (PoT) [Ryoo+, CVPR2015]
  - CNN単純連結
  - CNN + IDT [Jain+, ECCVW2014]
- Optical Flowベース
  - IDT [Wang+, ICCV2013]
  - IDT + Co-Feature [Kataoka+, ACCV2014]
  - All Features in IDT

実験結果

SMD + PoT連結ベクトル (Proposal)が最高精度を達成







PoTとの比較

従来研究と比較して，SMDは有効であることを実証

- NTSEL: +1.88%, +6. 76%
- UTKinect: +4.51%, +2.30%
- Watch-n-Patch: +1.67%, +4.12%

※ 左：10 [frames]蓄積時，右：3 [frames]蓄積時

Table 2. Detailed performance rate of our proposal (late fusion) and PoT [25] on three different datasets.

	% on NTSEL		% on UT		% on WnP	
	10 frm	3 frm	10 frm	3 frm	10 frm	3 frm
Proposal (Late Fusion)	**98.88**	**83.91**	**96.51**	**67.76**	**56.60**	**48.93**
PoT, CVPR2015 [25]	97.00	77.15	92.00	65.46	54.93	44.81

遷移行動認識

本研究の新規性として,

1. 遷移行動 (TA; Transitional Action)を定義
2. TAを解くための特徴Subtle Motion Descriptor (SMD)を提案

今後の展望

今後の展望

- 純粋に対象物を見るような動画認識手法の考案
  - 現在のCNNベースの手法は特徴が混在
- 行動検出
  - 対象行動のフレーム位置を推定
  - 物体検出のように画像中の位置を推定
- 行動予測
  - 再帰的に時系列情報を参照する枠組み(e.g. RNN)も整備されている
- 大規模すぎるDB(次ページ)をいかに扱うか？
  - 計算コストを削減したモデル？
  - 教師なし/ 弱教師付き学習が発展？

YouTube 8M

Dataset Explore Download About

YouTube-8M Dataset

YouTube-8M is a large-scale labeled video dataset that consists of 8 million YouTube video IDs and associated labels from a diverse vocabulary of 4800 visual entities. It also comes with precomputed state-of-the-art vision features from billions of frames, which fit on a single hard disk. This makes it possible to train video models from hundreds of thousands of video hours in less than a day on 1 GPU!

Our goal is to accelerate research on large-scale video understanding, representation learning, noisy data modeling, transfer learning, and domain adaptation approaches for video. More details about the dataset and initial experiments can be found in our technical report.

8 Million | 0.5 Million | 1.9 Billion | 4800 | 1.8

https://research.google.com/youtube8m/

動画認識は著しく進んだように見えますが，人間と背景が混在した学習モデルを生成しているのが現状です．

大規模化するデータを有効に活用し，動画ラベル以上の解析を行えるように準備していくことが重要です．