A Dense SURF and Triangulation based Spatio-Temporal Feature for Action Recognition

Do Hang Nga *Keiji Yanai*

The University of Electro-Communications, Tokyo

dohang@mm.cs.uec.ac.jp, yanai@cs.uec.ac.jp

Introduction

- A method of extracting ST features
  - *An extension of method proposed by Noguchi et al.[1]*
- **Improvements:**
  - Simple yet efficient selection of interest points
  - Novel ST descriptors
- Performance on UCF-101: **62.5%**
  - Fisher Vector encoding based video representation
  - Multiclass linear SVMs

[1] A. Noguchi and K. Yanai. A surf-based spatio-temporal feature for feature-fusion-based action recognition. In ECCV WS on Human Motion: Understanding, Modeling, Capture and Animation, 2010.

Related Works

- Comparison with some methods of ***extracting local ST features*** *based on* ***interest points***

  i.e.: 3D-SIFT, 3D-HOG, STIP, Trajectory[2]

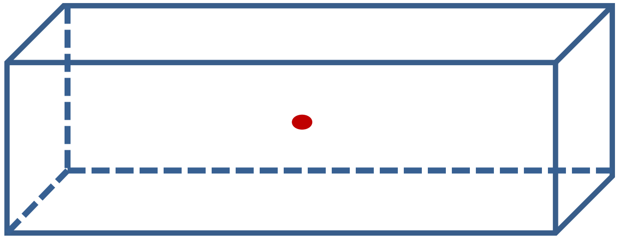
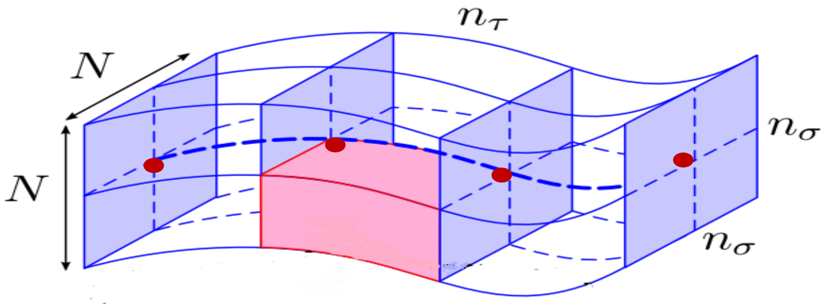
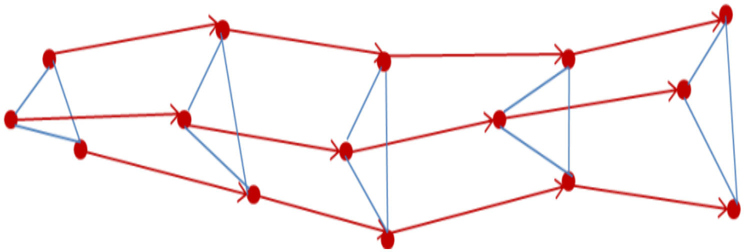
  - **Selection of interest points**
  - **Vectorization**

[2] H. Wang, A. Klaser, C. Schmid, and C-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, 103(1):60–79, 2013.

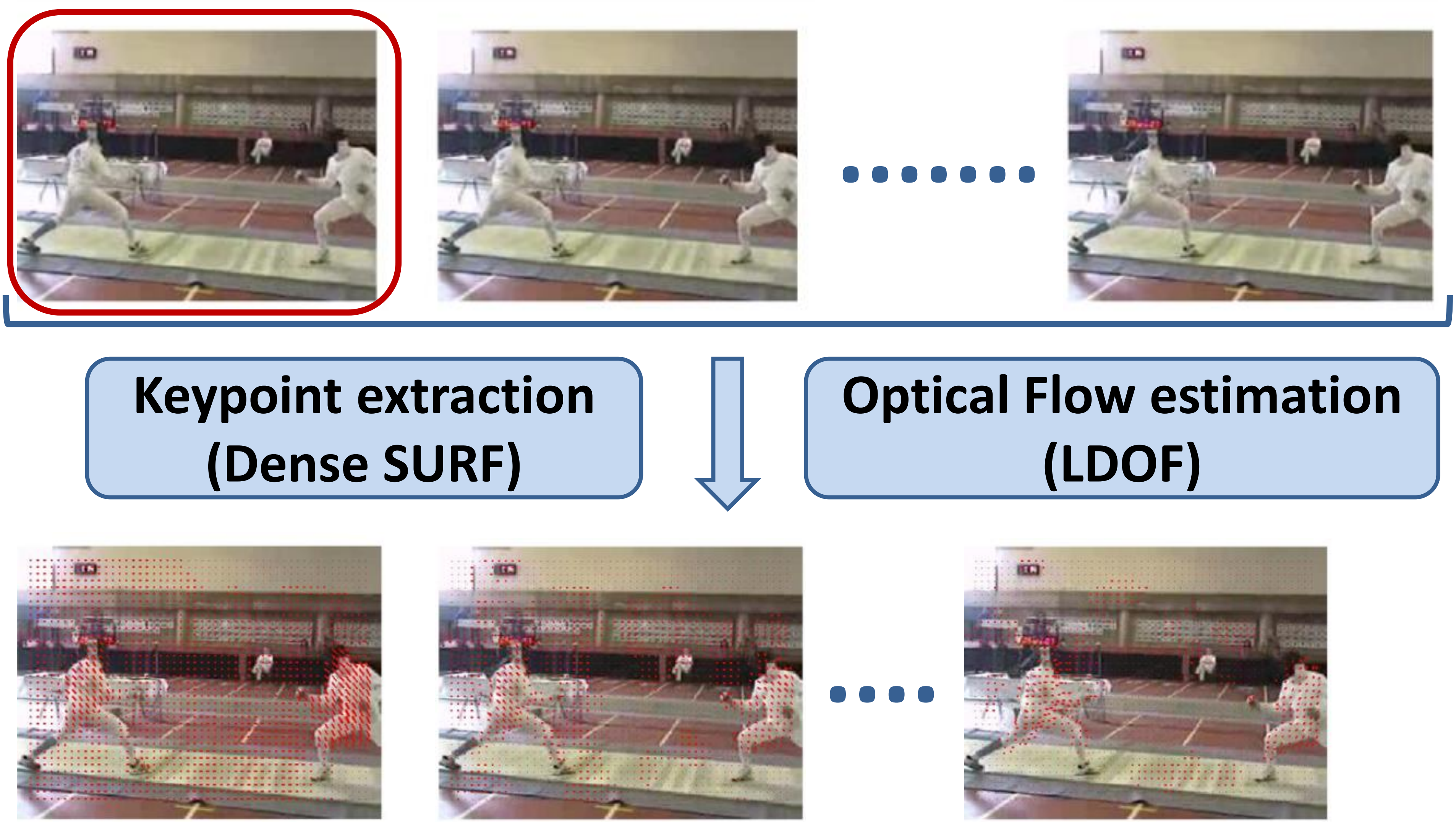
Related Works

	Point Selection
3D-SIFT	Random
3D-HOG	Harris operator
STIP	
Trajectory	Dense sampling
Proposed	**Dense sampling + Flow**

Related Works

	Vectorization
3D-SIFT 3D-HOG, STIP	
Trajectory	
Proposed	

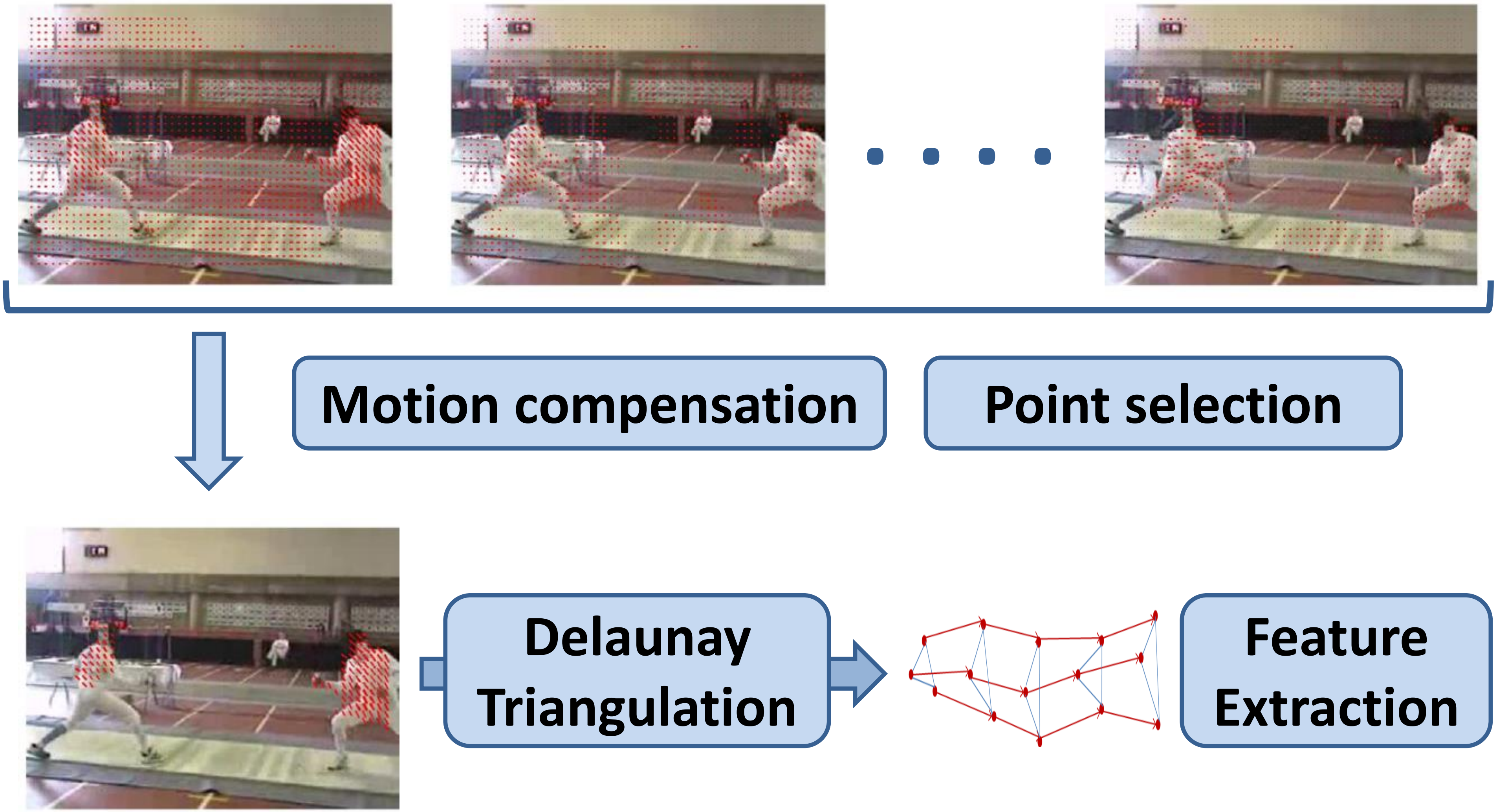
Overview of Our Method



Keypoint extraction (Dense SURF)

Optical Flow estimation (LDOF)

Overview of Our Method



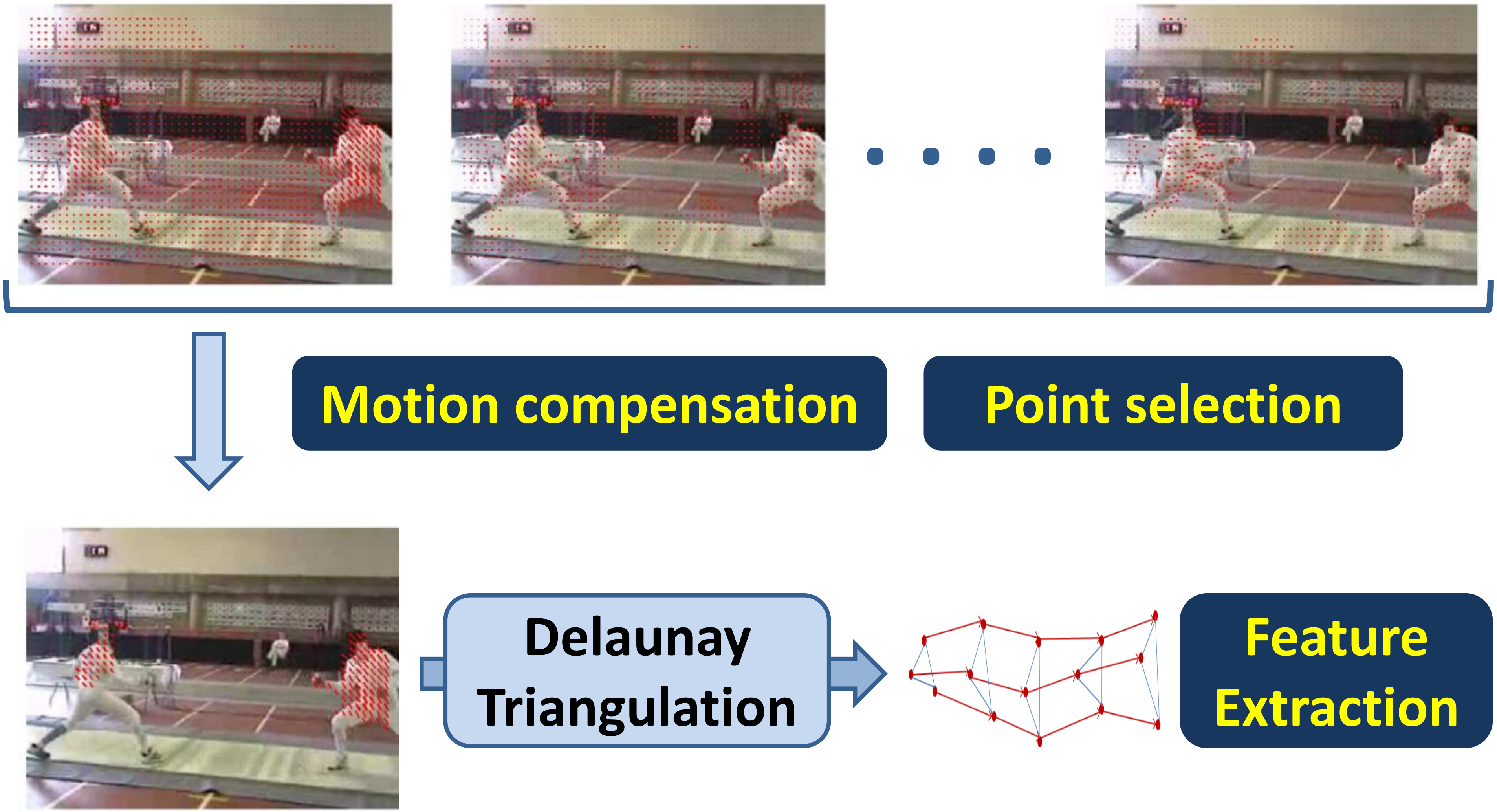
Motion compensation

Point selection

Delaunay Triangulation

Feature Extraction

Overview of Our Method



Motion compensation

Point selection

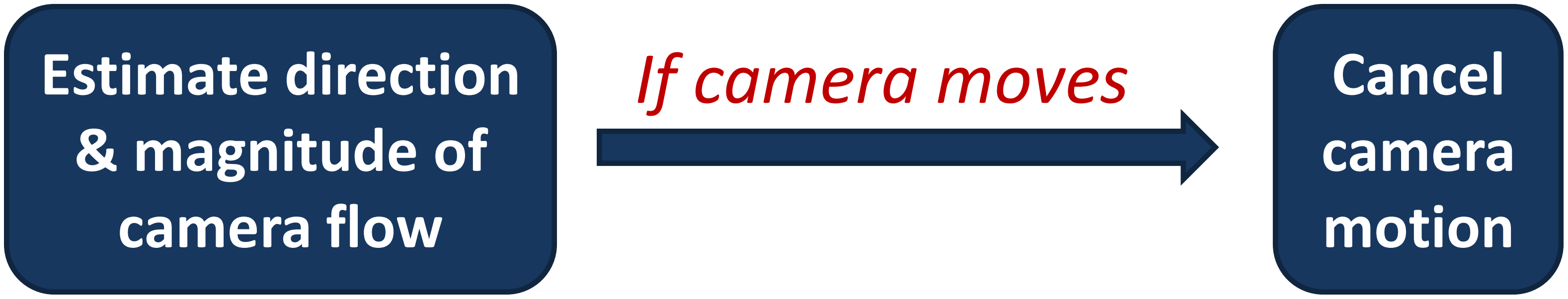
Delaunay Triangulation

Feature Extraction

Motion Compensation

- **Baseline** (Noguchi *et al.*'s work):
  - no motion compensation
  - no features extracted from videos with camera motion

- **Ours**: simple yet efficient motion compensation

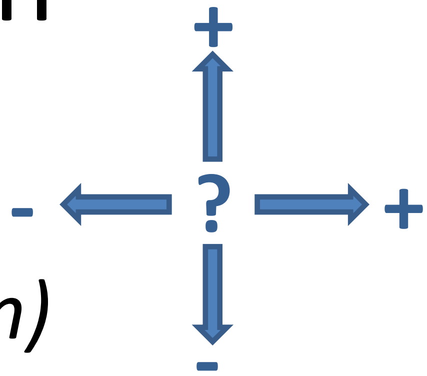


Estimate direction & magnitude of camera flow

If camera moves

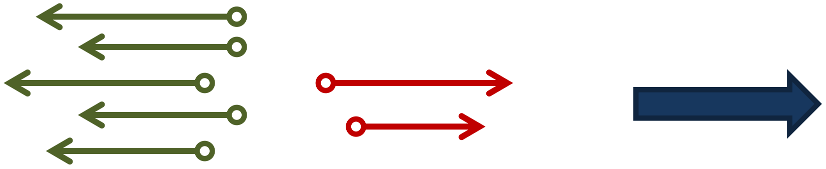
Cancel camera motion

Motion Compensation



- **Estimation of camera motion**

  e.g.: x direction (same for y direction)



Camera moved to the left

$$\boldsymbol{f^x}_{\boldsymbol{camera}} = \frac{\sum_i^P |f^x{}_i|}{P} \qquad s.t. \begin{cases} f^x{}_i < 0 \\ |f^x{}_i| > \varepsilon \end{cases}$$

$f^x{}_i$: horizontal optical flow of point i

$f^x{}_{camera}$: horizontal magnitude of camera flow

Motion Compensation

- **Compensation of motion of keypoints**

e.g.: x direction (same for y direction)

$$\boxed{\boldsymbol{f^{x}}_{\boldsymbol{k}} = \boldsymbol{f^{x}}_{\boldsymbol{k}} - \boldsymbol{\lambda f^{x}}_{\boldsymbol{camera}}}$$

$$\lambda = \begin{cases} 1 \ \ if \ camera \, is \, moving \, right \\ -1 \ \ if \ camera \, is \, moving \, left \end{cases}$$

Interest Point Selection

- **Principals of point selection**

Baseline: *at least once* flow > *fixed* threshold

Ours:

① flow > *flexible* threshold

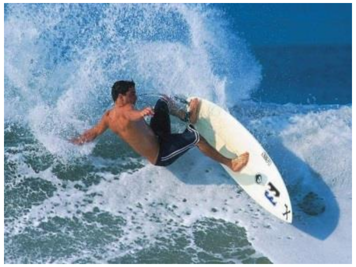
② prefer points with *more movements*

Interest point selection

- *Why motion threshold should be flexible?*

Because magnitude of movement varies largely from action to action

Large movements



Surfing

High jump

Ice dancing

Small movements

Apply lipstick

Typing

Shaving beard

Interest Point Selection

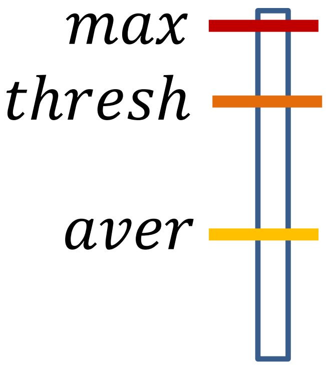
- *Why motion threshold should be flexible?*

Because magnitude of movement depends on the environment

Large movements

Small movements

Interest Point Selection



- **Estimation of motion threshold**
  - done for *every frame*
  - In each of 4 directions: x^+, x^-, y^+, y^-
  - e.g. motion threshold for frame fr in x^+ direction

$$\boxed{\boldsymbol{thresh}_{\boldsymbol{fr}_{\boldsymbol{x}^+}} = \boldsymbol{aver}_{\boldsymbol{fr}_{\boldsymbol{x}^+}} + \boldsymbol{\alpha}\left(\boldsymbol{max}_{\boldsymbol{fr}_{\boldsymbol{x}^+}} - \boldsymbol{aver}_{\boldsymbol{fr}_{\boldsymbol{x}^+}}\right)}$$

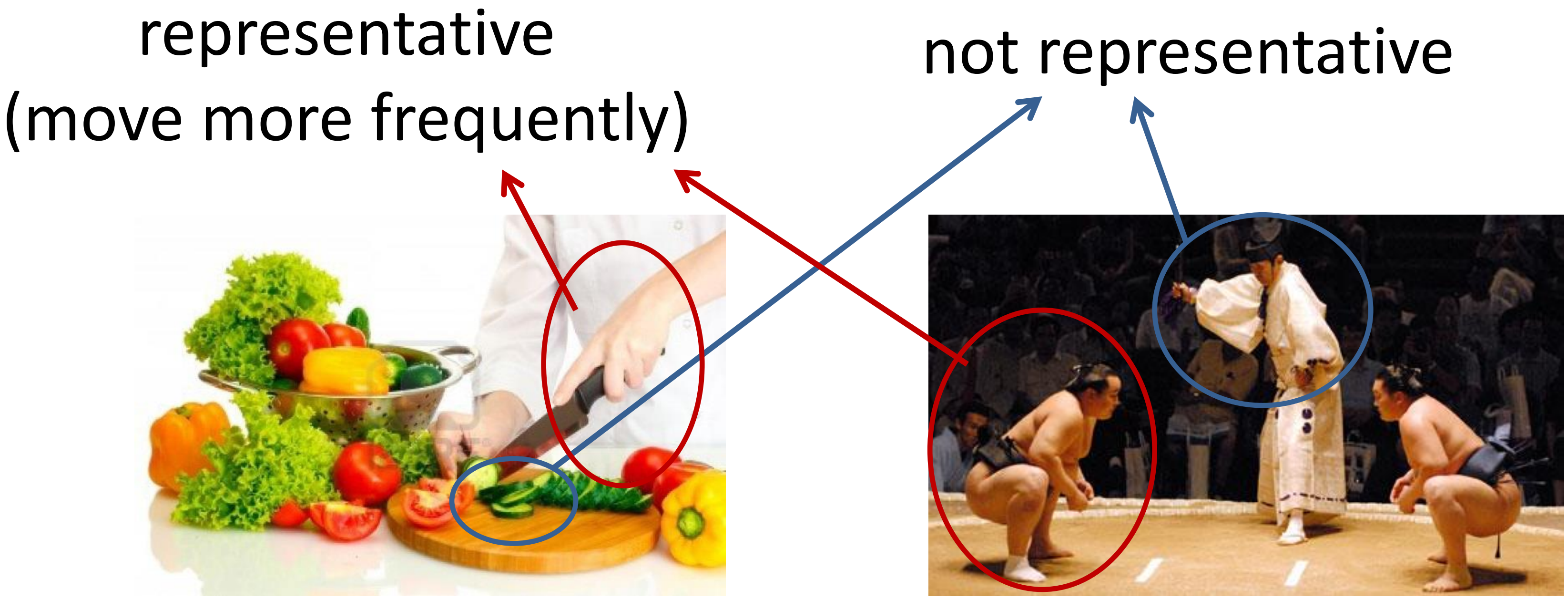
$if\ \ \boldsymbol{f^{x^+}}_{\boldsymbol{i}} \geq \boldsymbol{thresh}_{\boldsymbol{fr}_{\boldsymbol{x}^+}}, then\ point\ i\ is\ moving\ right$

$if\ \boldsymbol{thresh}_{\boldsymbol{fr}_{\boldsymbol{x}^+}} < \varepsilon, then\ no\ point\ is\ moving\ right$

Interest point selection

- *Why points with more movements should be preferred?*

<u>Because they are more representative</u>



representative
(move more frequently)

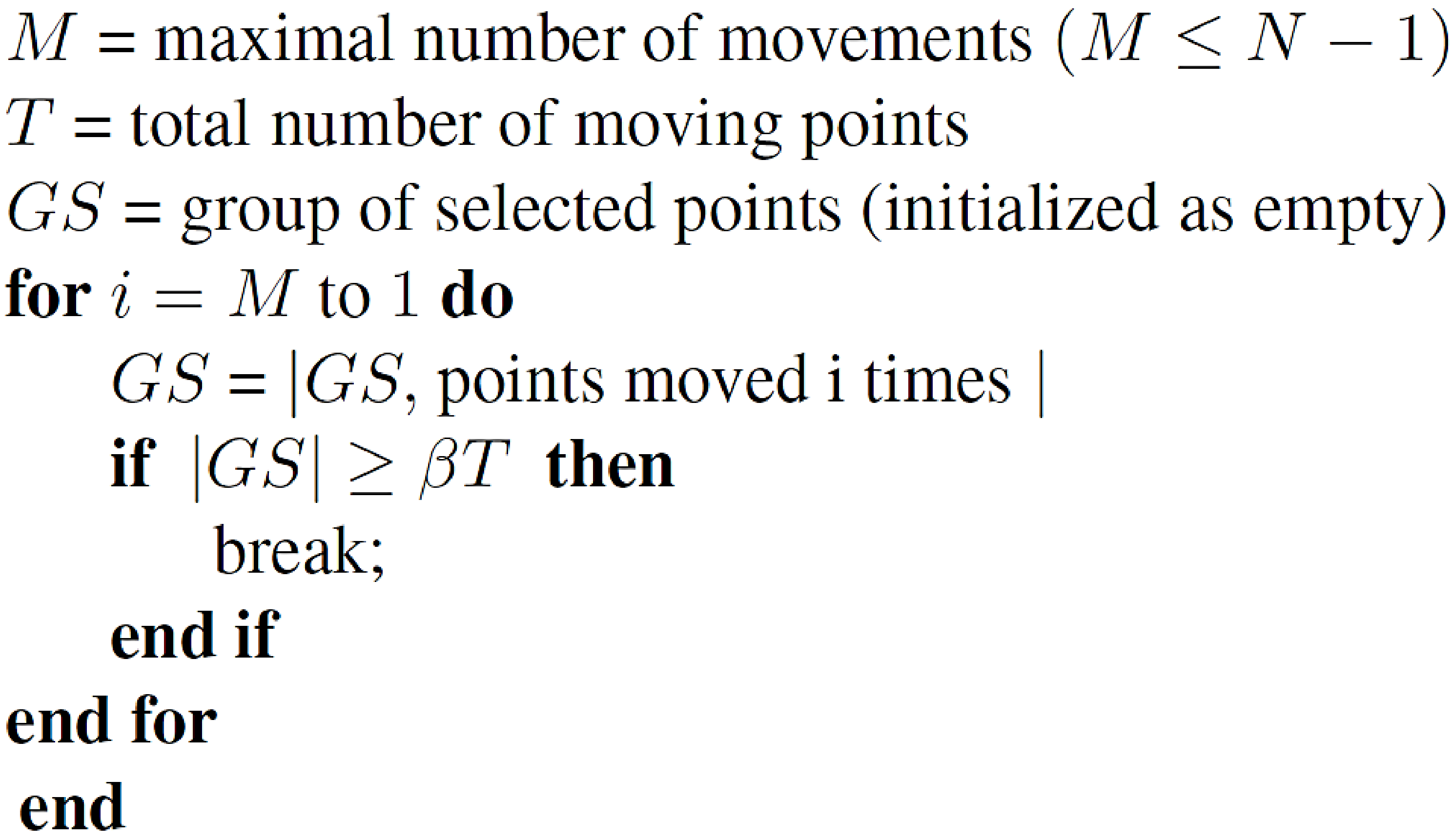
not representative

Cutting in kitchen

Sumo Wrestling

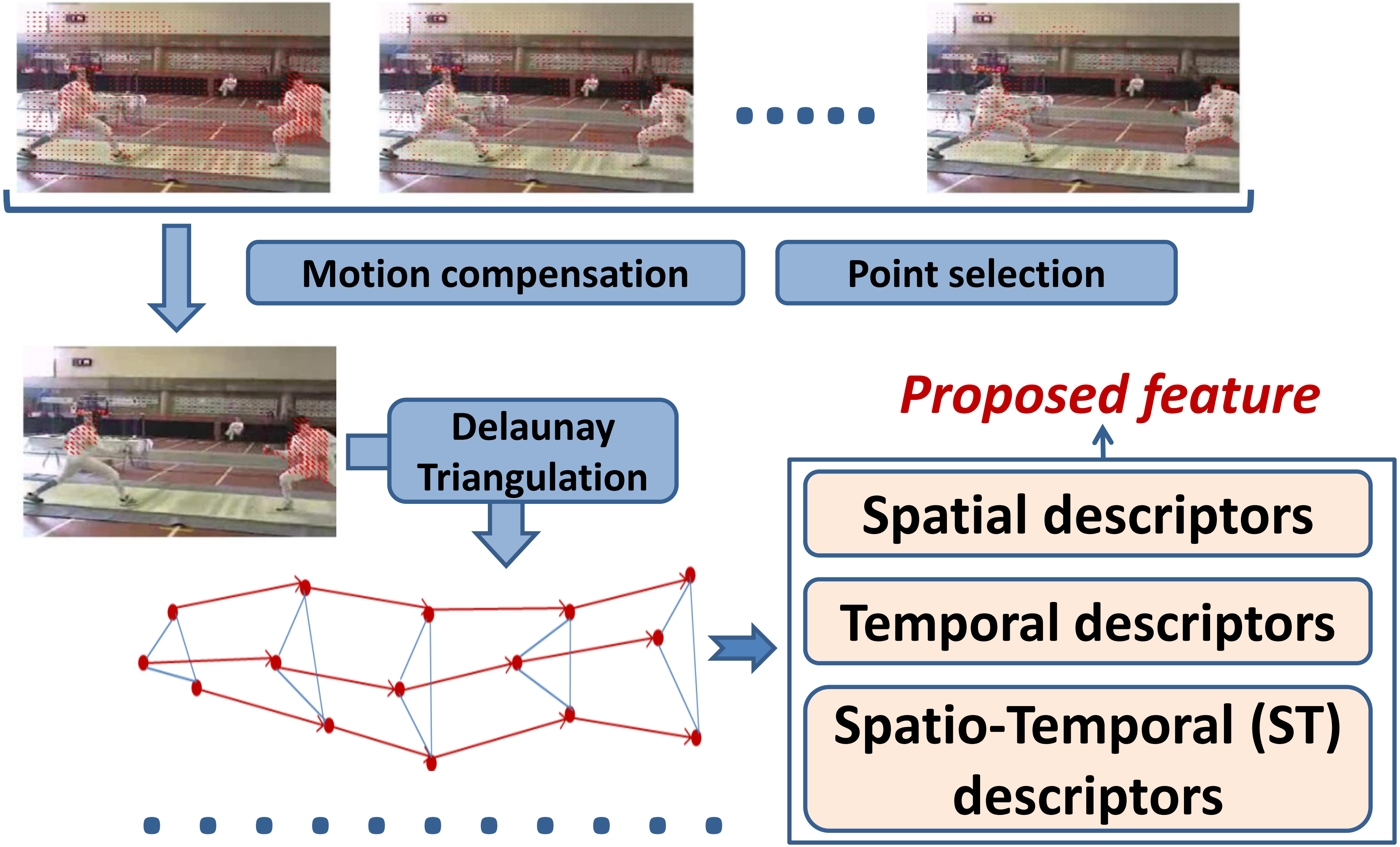
Interest Point Selection

- **Algorithm of selecting interest points**



```
$M$ = maximal number of movements ($M \leq N - 1$)
$T$ = total number of moving points
$GS$ = group of selected points (initialized as empty)
for $i = M$ to 1 do
    $GS$ = |$GS$, points moved i times |
    if $|GS| \geq \beta T$ then
        break;
    end if
end for
 end
```

Feature Extraction



Motion compensation

Point selection

Delaunay Triangulation

Proposed feature

Spatial descriptors

Temporal descriptors

Spatio-Temporal (ST) descriptors

Feature Extraction

Spatial descriptors[1]:

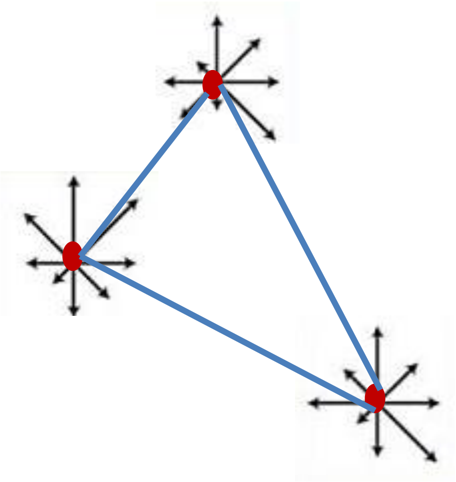
SURF descriptors (64-D) of 3 points of the triple (first frame)

concatenate

192-D descriptor

PCA

PCA-SURF (96-D)

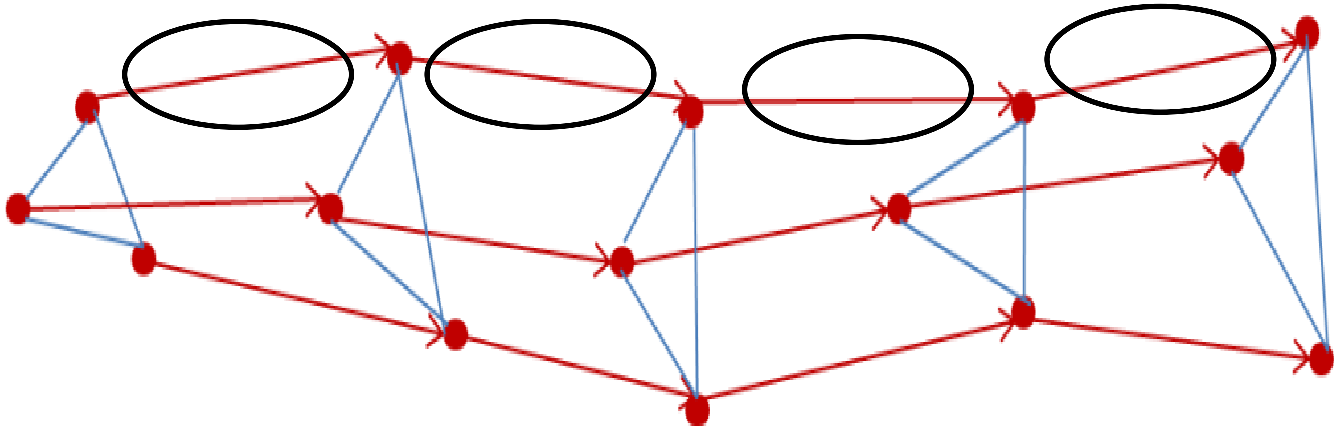


Feature Extraction

Temporal descriptors:

① Histogram of Oriented Optical Flow (**HOOF**)[3]

② Histogram of Direction of Flow (**HDF**)[1]



[3] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and binetcauchy kernels on nonlinear dynamical systems for the recognition of human actions. In Proc. of IEEE Computer Vision and Pattern Recognition, pages 1932– 1939, 2009

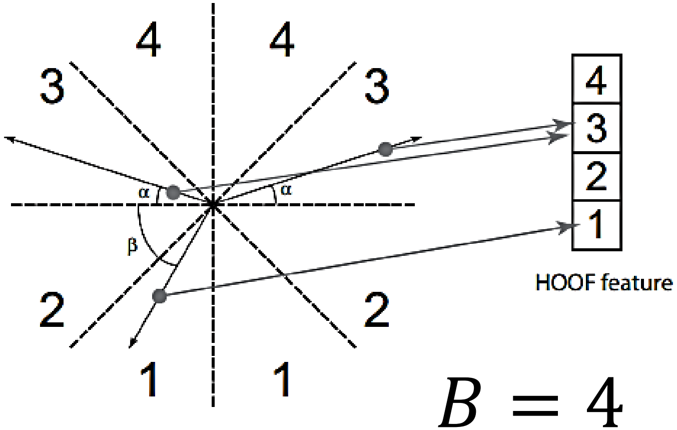
Feature Extraction

- ***HOOF* (Histogram of Oriented Optical Flow)**
  - **$3(N-1)$ flow vectors** of 3 points are binned to a B-bin histogram
  - $v = [x, y]$ with $\theta = \tan^{-1}(y)$ in the range:

$$-\frac{\pi}{2} + \pi\frac{b-1}{B_o} \leq \theta < -\frac{\pi}{2} + \pi\frac{b}{B_o}$$

will contribute by $\sqrt{x^2+y^2}$ to the sum in bin b

  - Histogram is normalized to sum up to 1

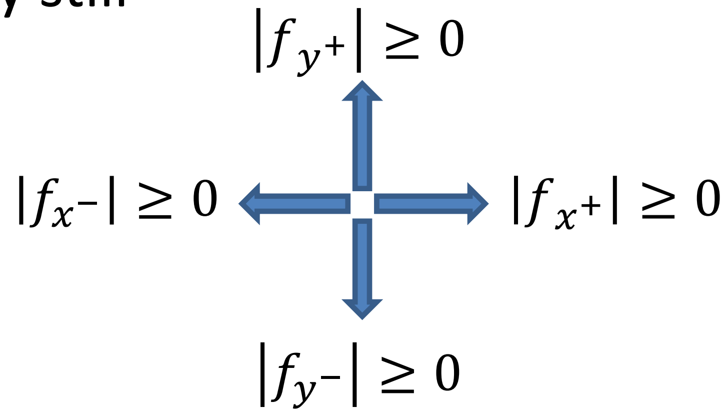


Optical Flow Vectors

$B = 4$

Feature Extraction

- ***HDF*** ***(Histogram of Direction of Flow)***
  - **$3(N-1)$ flow vectors** are binned to a 4-bin histogram based on direction of movements:
    - $|f_{x^+}| \geq 0$: moving forward or horizontally still
    - $|f_{x^-}| \geq 0$: moving backward
    - $|f_{y^+}| \geq 0$: moving up or vertically still
    - $|f_{y^-}| \geq 0$: moving down
  - Histogram is normalized to sum up to 1



Feature Extraction

ST descriptors:

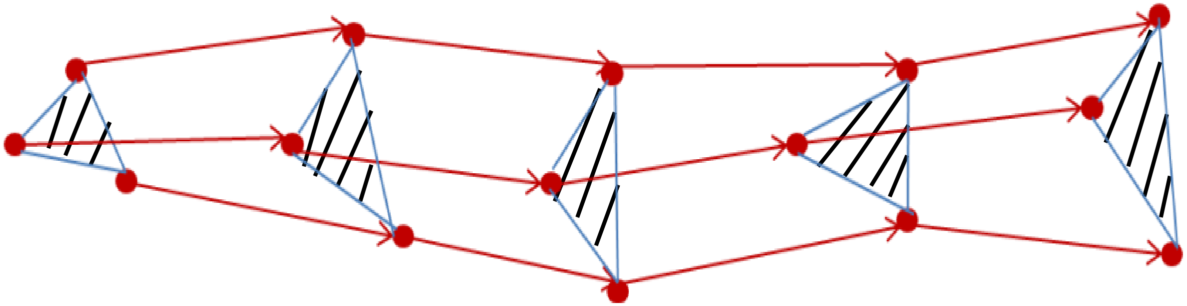
① Area of Triangle (**AT**)[1]

② **Histogram of Angle of Triangle (HAT)**

③ **Histogram of Angle of Flow (HAF)**

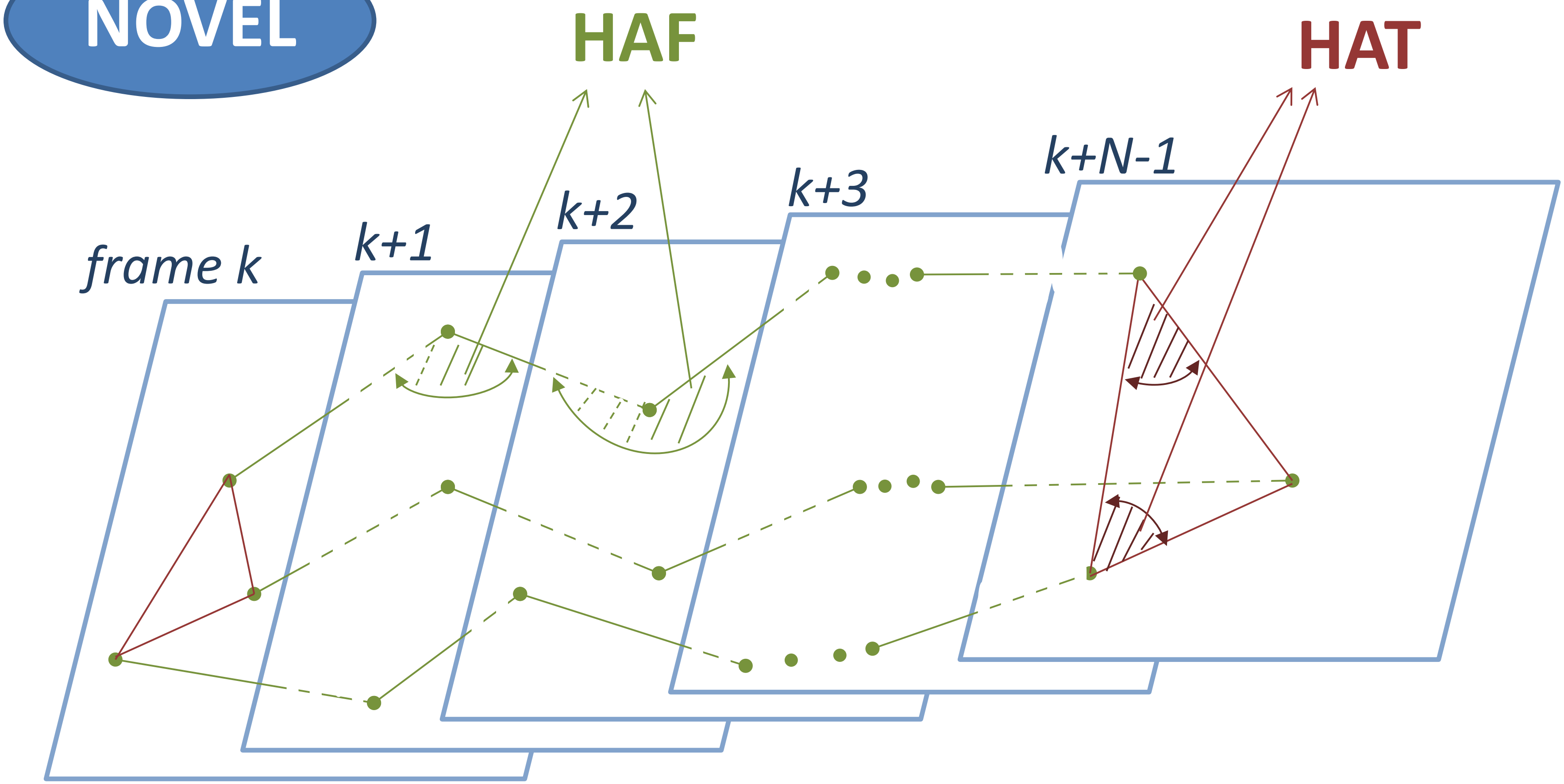
Proposed by us

- ***AT (Area of Triangle)***
  - concatenated by areas of triangles at N frames
  - normalized to sum up to 1



Feature Extraction

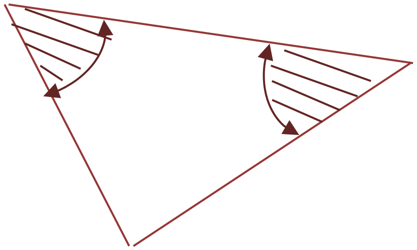
NOVEL



Feature Extraction

- ***HAT* (Histogram of Angle of Triangle)**
  - **2 smallest angles of each triangle** are

    binned to a 5 bin-histogram:
    $[0-15],[15-30],[30-45],[45-60],[60-90]$
  - Each angle is weighted by sum of magnitude of its two edges
  - The histogram is normalized to sum up to 1

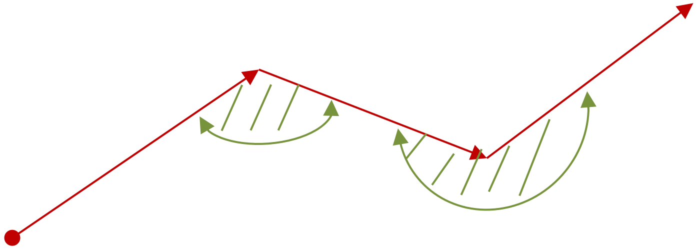


Feature Extraction

- ***HAF*** ***(Histogram of Angle of Flow)***
  - $\mathbf{3(N-1)}$ **angles** shaped by adjacent trajectories are binned similarly to HOOF

$$-\frac{\pi}{2}+\pi\frac{b-1}{B}\leq\theta<-\frac{\pi}{2}+\pi\frac{b}{B}$$

  - Each angle is weighted by sum of magnitude of its two edges
  - The histogram is normalized to sum up to 1

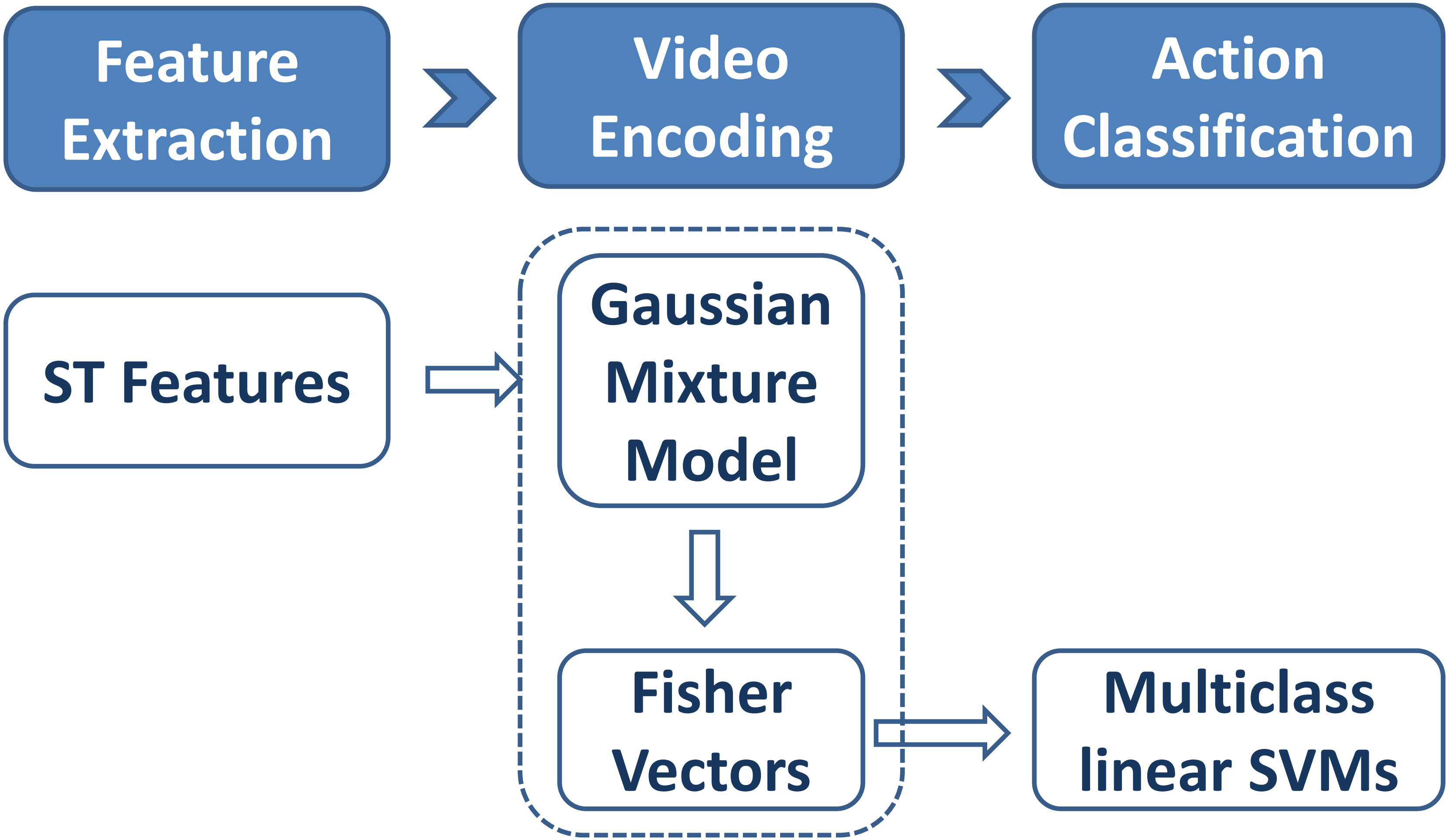


Summary of Proposed ST Descriptor

	Feature	Dimension
Spatial	PCA-SURF	96
Temporal	HOOF	$4^{(*)}$
	HDF	4
Spatio-Temporal	AT	$5^{(*)}$
	HAT	5
	HAF	$4^{(*)}$
	Proposed ST	***118***

$(*)$: ***adjustable***

Classification Framework



Feature Extraction

Video Encoding

Action Classification

ST Features

Gaussian Mixture Model

Fisher Vectors

Multiclass linear SVMs

Experiments and Results

- Dataset: UCF-101
  - ***101*** *actions,* ***13320*** *videos*
- Evaluation method: workshop *THUMOS'13* [1]
  - *3 training/test splits*
- Methods of extracting ST features to compare:
  - Baseline[1]
  - HOG, MBH (Dense Trajectories[2])
  - Proposed

[1] http://crcv.ucf.edu/ICCV13-Action-Workshop/: The First International Workshop on Action Recognition with a Large Number of Classes, in conjunction with ICCV '13

Experiments and Results

Method	Precision
Baseline[1]	38.2%
HOG	56.4%
MBH	61.6%
Proposed	**62.5%**
Combined (HOG+MBH+Proposed)	**74.7%**

Conclusions

- A method of extracting ST features as an extension of Noguchi *etal.*'s method[1]
- Better performance than dense trajectory based features[2]
  - *complementary to [2]*
- Future works:
  - handle more complicated camera motion
  - combine with other features