FaSTIP: A New Method for Detection and Description of Space-Time Interest Points for Human Activity Classification

Soumitra Samanta and Bhabatosh Chanda

Indian Statistical Institute, Kolkata
soumitra_r@isical.ac.in, chanda@isical.ac.in

Human activity Analysis

- Due to applications in surveillance, video indexing and automatic video navigation, human activity analysis is quite a hot topic in Computer vision.

[1] Aggarwal and Ryoo, "Human Activity Analysis: A Review", ACM Computing Surveys, 2011

Human activity Analysis

- Due to applications in surveillance, video indexing and automatic video navigation, human activity analysis is quite a hot topic in Computer vision.
- Human activity analysis may be broadly classified into **two** main approaches[1]
  - **Single layered approaches**
  - **Hierarchical approaches**

[1] Aggarwal and Ryoo, "Human Activity Analysis: A Review", ACM Computing Surveys, 2011

Human activity Analysis

- Due to applications in surveillance, video indexing and automatic video navigation, human activity analysis is quite a hot topic in Computer vision.
- Human activity analysis may be broadly classified into **two** main approaches[2]
  - **Single layered approaches**
    - Spatio-temporal features
  - **Hierarchical approaches**

[2] Aggarwal and Ryoo, "Human Activity Analysis: A Review", ACM Computing Surveys, 2011

Spatio-temporal features based human activity analysis

- Spatio-temporal feature based approaches may further be grouped into **Two** categories.

Spatio-temporal features based human activity analysis

- Spatio-temporal feature based approaches may further be grouped into **Two** categories.
  - **Global feature**
    - histograms of gradient and optical flow computed over the frames (e.g., **HOG** and **HOF**)

Spatio-temporal features based human activity analysis

- Spatio-temporal feature based approaches may further be grouped into **Two** categories.
  - **Global feature**
    - histograms of gradient and optical flow computed over the frames (e.g., **HOG** and **HOF**)
  - **Local feature**
    - features computed over a neighborhood around interest point (e.g., **STIP** and **Cuboid**)

Spatio-temporal features based human activity analysis

- Spatio-temporal feature based approaches may further be grouped into **Two** categories.
  - **Global feature**
    - histograms of gradient and optical flow computed over the frames (e.g., **HOG** and **HOF**)
  - **Local feature**
    - features computed over a neighborhood around interest point (e.g., **STIP** and **Cuboid**)
- **Local feature based approach is so far the most successful.**

General structure of the human activity analysis based on local spatio-temporal features

General structure of the human activity analysis based on local spatio-temporal features

- Detect space-time interest points

General structure of the human activity analysis based on local spatio-temporal features

- Detect space-time interest points
- Describe the interest points in terms of locally computed features

General structure of the human activity analysis based on local spatio-temporal features

- Detect space-time interest points
- Describe the interest points in terms of locally computed features
- *Generate the vocabulary as bag-of-features*

General structure of the human activity analysis based on local spatio-temporal features

- Detect space-time interest points
- Describe the interest points in terms of locally computed features
- *Generate the vocabulary as bag-of-features*
- Label the feature vectors by nearest neighbor classification

General structure of the human activity analysis based on local spatio-temporal features

- Detect space-time interest points
- Describe the interest points in terms of locally computed features
- *Generate the vocabulary as bag-of-features*
- Label the feature vectors by nearest neighbor classification
- Generate the distribution of labels as the representation of video

General structure of the human activity analysis based on local spatio-temporal features

- Detect space-time interest points
- Describe the interest points in terms of locally computed features
- *Generate the vocabulary as bag-of-features*
- Label the feature vectors by nearest neighbor classification
- Generate the distribution of labels as the representation of video
- *Learn the action models or the classifiers*

General structure of the human activity analysis based on local spatio-temporal features

- Detect space-time interest points
- Describe the interest points in terms of locally computed features
- *Generate the vocabulary as bag-of-features*
- Label the feature vectors by nearest neighbor classification
- Generate the distribution of labels as the representation of video
- *Learn the action models or the classifiers*
- Classify the test video

General structure of the human activity analysis based on local spatio-temporal features

- **Detect space-time interest points**
- **Describe the interest points in terms of locally computed features**
- *Generate the vocabulary as bag-of-features*
- Label the feature vectors by nearest neighbor classification
- Generate the distribution of labels as the representation of video
- *Learn the action models or the classifiers*
- Classify the test video

Human activity analysis based on local spatio-temporal features

[3] Dollar et al., Behavior Recognition via Sparse Spatio-Temporal Features, VS-PETS, 2005

Human activity analysis based on local spatio-temporal features

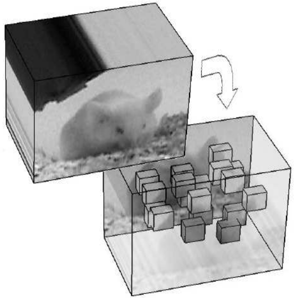
- Dollar et al.[3] have used two-dimensional Gaussian smoothing kernel in the spatial domain, and two one-dimensional Gabor filters in the temporal domain to detect the interest points.

[3] Dollar et al., Behavior Recognition via Sparse Spatio-Temporal Features, VS-PETS, 2005

Human activity analysis based on local spatio-temporal features

- Dollar et al.[3] have used two-dimensional Gaussian smoothing kernel in the spatial domain, and two one-dimensional Gabor filters in the temporal domain to detect the interest points.

- They try to capture **salient periodic** motion.
- **Feature**
  - Color / intensity
  - Gradient
  - Optical flow



[3] Dollar et al., Behavior Recognition via Sparse Spatio-Temporal Features, VS-PETS, 2005

Human activity analysis based on local spatio-temporal features (cont.)

- Laptev et al.[4] have detected interest points by extending the two-dimensional **Harris corner** to three-dimension

[4] Laptev et al. On Space-Time Interest Points, IJCV, 2005

Human activity analysis based on local spatio-temporal features (cont.)

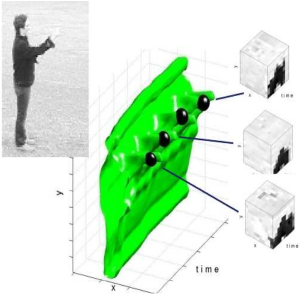
- Laptev et al.[4] have detected interest points by extending the two-dimensional **Harris corner** to three-dimension
- They formed a 3×3 spatio-temporal second-moment matrix of first order spatial and temporal derivatives

[4] Laptev et al. On Space-Time Interest Points, IJCV, 2005

Human activity analysis based on local spatio-temporal features (cont.)

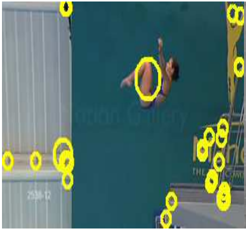
- Laptev et al.[4] have detected interest points by extending the two-dimensional **Harris corner** to three-dimension
- They formed a 3×3 spatio-temporal second-moment matrix of first order spatial and temporal derivatives

- Features are computed from a volume around each interest point divided into a grid of cells
- For each cell a 4-bin histogram of oriented gradient (**HOG**) and 5-bin histogram of oriented optical flow (**HOF**) are computed and concatenated to generate the feature vector.

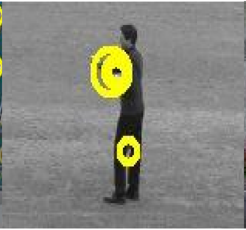


[4] Laptev et al. On Space-Time Interest Points, IJCV, 2005

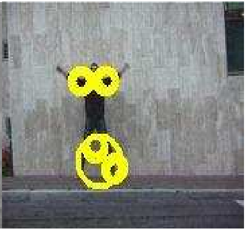
Drawbacks



UCF sports (lifting)



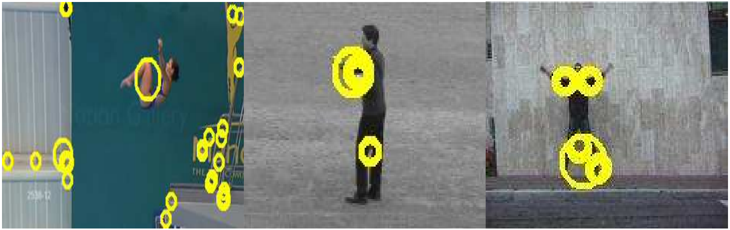
KTH (boxing)



Weizmann (pjump)

The points show using Laptev STIP.

Drawbacks

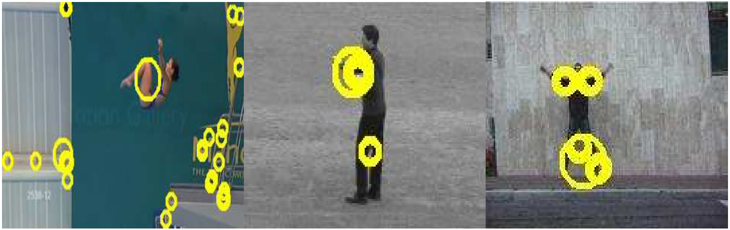


UCF sports (lifting) KTH (boxing) Weizmann (pjump)

The points show using Laptev STIP.

- Less sensitive to smooth motion

Drawbacks

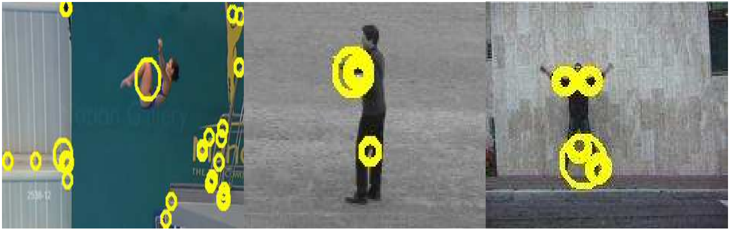


UCF sports (lifting) KTH (boxing) Weizmann (pjump)

The points show using Laptev STIP.

- Less sensitive to smooth motion
- Many points are outside the interest region

Drawbacks



UCF sports (lifting) KTH (boxing) Weizmann (pjump)

The points show using Laptev STIP.

- Less sensitive to smooth motion
- Many points are outside the interest region

To address these problems we propose a novel method based on the facet model.

Rest of the talk

Rest of the talk

- Two dimensional facet model

Rest of the talk

- Two dimensional facet model
- Proposed method

Rest of the talk

- Two dimensional facet model
- Proposed method
- Experimental evaluation

Rest of the talk

- Two dimensional facet model
- Proposed method
- Experimental evaluation
- Conclusion

Two dimensional facet model

- An image region may be approximated by piecewise bi-cubic function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ given by[5]

$$
\begin{aligned}
f(x,y) \; = \; & k_1 + k_2 x + k_3 y + k_4 x^2 + k_5 xy + k_6 y^2 + \\
& k_7 x^3 + k_8 x^2 y + k_9 x y^2 + k_{10} y^3
\end{aligned}
$$

[5] Haralick and Shapiro, "Computer and Robot Vision", Addison-Wesley Publishing Company, 1992

Two dimensional facet model

- An image region may be approximated by piecewise bi-cubic function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ given by[5]

$$\begin{aligned} f(x,y) &= k_1 + k_2x + k_3y + k_4x^2 + k_5xy + k_6y^2 + \\ &\quad k_7x^3 + k_8x^2y + k_9xy^2 + k_{10}y^3 \end{aligned}$$

  where coefficients $k_1, \ldots.., k_{10}$ are calculated by convolving the image with different two dimensional masks.

[5] Haralick and Shapiro, "Computer and Robot Vision", Addison-Wesley Publishing Company, 1992

Two dimensional facet model

- An image region may be approximated by piecewise bi-cubic function $f : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}$ given by[5]

$$f(x,y) = k_1 + k_2x + k_3y + k_4x^2 + k_5xy + k_6y^2 + k_7x^3 + k_8x^2y + k_9xy^2 + k_{10}y^3$$

where coefficients $k_1, \ldots\ldots, k_{10}$ are calculated by convolving the image with different two dimensional masks.

-13	2	7	2	-13
2	17	22	17	2
7	22	27	22	7
2	17	22	17	2
-13	2	7	2	-13

$[\frac{1}{175}]$ k_1

[5] Haralick and Shapiro, "Computer and Robot Vision", Addison-Wesley Publishing Company, 1992

Two dimensional facet model

- An image region may be approximated by piecewise bi-cubic function $f : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}$ given by[5]

$$f(x,y) = k_1 + k_2x + k_3y + k_4x^2 + k_5xy + k_6y^2 + k_7x^3 + k_8x^2y + k_9xy^2 + k_{10}y^3$$

where coefficients $k_1, ..., k_{10}$ are calculated by convolving the image with different two dimensional masks.

-13	2	7	2	-13
2	17	22	17	2
7	22	27	22	7
2	17	22	17	2
-13	2	7	2	-13

$[\frac{1}{175}]$ k_1

31	-5	-17	-5	31
-44	-62	-68	-62	-44
0	0	0	0	0
44	62	68	62	44
-31	5	17	5	-31

$[\frac{1}{420}]$ k_2

[5] Haralick and Shapiro, "Computer and Robot Vision", Addison-Wesley Publishing Company, 1992

Two dimensional facet model

- An image region may be approximated by piecewise bi-cubic function $f : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}$ given by[5]

$$f(x,y) = k_1 + k_2x + k_3y + k_4x^2 + k_5xy + k_6y^2 + k_7x^3 + k_8x^2y + k_9xy^2 + k_{10}y^3$$

where coefficients $k_1, ..., k_{10}$ are calculated by convolving the image with different two dimensional masks.

-13	2	7	2	-13
2	17	22	17	2
7	22	27	22	7
2	17	22	17	2
-13	2	7	2	-13

$[\frac{1}{175}]$ k_1

31	-5	-17	-5	31
-44	-62	-68	-62	-44
0	0	0	0	0
44	62	68	62	44
-31	5	17	5	-31

$[\frac{1}{420}]$ k_2

$\cdots$

-1	2	0	-2	-1
-1	2	0	-2	-1
-1	2	0	-2	-1
-1	2	0	-2	-1
-1	2	0	-2	-1

$[\frac{1}{60}]$ k_{10}

[5] Haralick and Shapiro, "Computer and Robot Vision", Addison-Wesley Publishing Company, 1992

Two dimensional facet model: corner points

- A corner point is where the gradient changes abruptly along the direction orthogonal to the gradient direction.

Two dimensional facet model: corner points

- A corner point is where the gradient changes abruptly along the direction orthogonal to the gradient direction.
- A corner response function $\theta'_{\alpha}(0,0)$ at the center (i.e., candidate pixel) may be defined as

$$\theta'_{\alpha}(0,0) = \frac{-2(k_2^2 k_6 - k_2 k_3 k_5 + k_3^2 k_4)}{(k_2^2 + k_3^2)^{\frac{3}{2}}}$$

Two dimensional facet model: corner points

- A corner point is where the gradient changes abruptly along the direction orthogonal to the gradient direction.
- A corner response function $\theta'_\alpha(0,0)$ at the center (i.e., candidate pixel) may be defined as

$$\theta'_\alpha(0,0) = \frac{-2(k_2^2 k_6 - k_2 k_3 k_5 + k_3^2 k_4)}{(k_2^2 + k_3^2)^{\frac{3}{2}}}$$

- Finally, the candidate pixel $(0,0)$ is declared as a corner point if the following two conditions are satisfied:

Two dimensional facet model: corner points

- A corner point is where the gradient changes abruptly along the direction orthogonal to the gradient direction.
- A corner response function $\theta'_\alpha(0,0)$ at the center (i.e., candidate pixel) may be defined as

$$\theta'_\alpha(0,0) = \frac{-2(k_2^2 k_6 - k_2 k_3 k_5 + k_3^2 k_4)}{(k_2^2 + k_3^2)^{\frac{3}{2}}}$$

- Finally, the candidate pixel $(0,0)$ is declared as a corner point if the following two conditions are satisfied:
  - $(0,0)$ is an edge point, and

Two dimensional facet model: corner points

- A corner point is where the gradient changes abruptly along the direction orthogonal to the gradient direction.
- A corner response function $\theta'_{\alpha}(0,0)$ at the center (i.e., candidate pixel) may be defined as

$$\theta'_{\alpha}(0,0) = \frac{-2(k_2^2 k_6 - k_2 k_3 k_5 + k_3^2 k_4)}{(k_2^2 + k_3^2)^{\frac{3}{2}}}$$

- Finally, the candidate pixel $(0,0)$ is declared as a corner point if the following two conditions are satisfied:
  - $(0,0)$ is an edge point, and
  - For a given threshold Ω, $|\theta'_{\alpha}(0,0)| > \Omega$

Propose methodology

- We extend the two-dimensional facet model to three-dimension to detect the interest points in video data.

Propose methodology

- We extend the two-dimensional facet model to three-dimension to detect the interest points in video data.
- We estimate the video data as a tri-cubic function $f : \mathbb{N} \times \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}$ over a neighborhood of each point in the space-time domain given by

$$\begin{aligned} f(x, y, t) \quad = \quad & k_1 + k_2 x + k_3 y + k_4 t + k_5 x^2 + k_6 y^2 + k_7 t^2 + \\ & k_8 xy + k_9 yt + k_{10} xt + k_{11} x^3 + k_{12} y^3 + k_{13} t^3 \\ & + k_{14} x^2 y + k_{15} x y^2 + k_{16} y^2 t + k_{17} y t^2 + k_{18} x^2 t \\ & + k_{19} x t^2 + k_{20} xyt \end{aligned}$$

Propose methodology

- We extend the two-dimensional facet model to three-dimension to detect the interest points in video data.
- We estimate the video data as a tri-cubic function $f : \mathbb{N} \times \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}$ over a neighborhood of each point in the space-time domain given by

$$
\begin{aligned}
f(x, y, t) \quad = \quad & k_1 + k_2 x + k_3 y + k_4 t + k_5 x^2 + k_6 y^2 + k_7 t^2 + \\
& k_8 xy + k_9 yt + k_{10} xt + k_{11} x^3 + k_{12} y^3 + k_{13} t^3 \\
& + k_{14} x^2 y + k_{15} xy^2 + k_{16} y^2 t + k_{17} yt^2 + k_{18} x^2 t \\
& + k_{19} xt^2 + k_{20} xyt
\end{aligned}
$$

- We derive twenty different masks to calculate the coefficients $k_1, \ldots.., k_{20}$ by simple convolution with those masks over the neighborhood of the candidate point.

Three dimensional facet model for video data

- Calculate the rate of change of directional derivative of f in the direction orthogonal to the derivative direction.

Three dimensional facet model for video data

- Calculate the rate of change of directional derivative of f in the direction orthogonal to the derivative direction.
- Let $\overrightarrow{T}$ be the unit vector along the gradient of $f(x, y, t)$ at any point (x, y, t), then

$$\overrightarrow{T}(x, y, t) = \frac{1}{d}(f_x, f_y, f_t), \text{ where } d = \sqrt{f_x^2 + f_y^2 + f_t^2}$$

Three dimensional facet model for video data

- Calculate the rate of change of directional derivative of f in the direction orthogonal to the derivative direction.
- Let $\overrightarrow{T}$ be the unit vector along the gradient of $f(x, y, t)$ at any point (x, y, t), then

$$\overrightarrow{T}(x, y, t) = \frac{1}{d}(f_x, f_y, f_t), \text{ where } d = \sqrt{f_x^2 + f_y^2 + f_t^2}$$

- For a function f, the normal $\overrightarrow{N}$ to the gradient vector $\overrightarrow{T}$ is given by

$$\overrightarrow{N}(x, y, t) = \bigtriangledown^2 f - [\bigtriangledown^2 f \cdot \overrightarrow{T}]\overrightarrow{T}$$

where

$$\bigtriangledown^2 = \left(\frac{\partial^2}{\partial x^2}, \frac{\partial^2}{\partial y^2}, \frac{\partial^2}{\partial z^2}\right)$$

Three dimensional facet model for video data

- Calculate the rate of change of directional derivative of f in the direction orthogonal to the derivative direction.
- Let $\overrightarrow{T}$ be the unit vector along the gradient of $f(x, y, t)$ at any point (x, y, t), then

$$\overrightarrow{T}(x, y, t) = \frac{1}{d}(f_x, f_y, f_t), \text{ where } d = \sqrt{f_x^2 + f_y^2 + f_t^2}$$

- For a function f, the normal $\overrightarrow{N}$ to the gradient vector $\overrightarrow{T}$ is given by

$$\overrightarrow{N}(x, y, t) = \bigtriangledown^2 f - [\bigtriangledown^2 f \cdot \overrightarrow{T}]\overrightarrow{T}$$

where

$$\bigtriangledown^2 = \left(\frac{\partial^2}{\partial x^2}, \frac{\partial^2}{\partial y^2}, \frac{\partial^2}{\partial z^2}\right)$$

- **So to detect interest point we need to calculate $\overrightarrow{T}' \cdot \overrightarrow{N}$.**

Three dimensional facet model for video data (Cont.)

- Consider a straight line passing through the origin and any point on that line be $(\rho \sin\theta \sin\phi, \rho \sin\theta \cos\phi, \rho \cos\theta)$.

Three dimensional facet model for video data (Cont.)

- Consider a straight line passing through the origin and any point on that line be $(\rho \sin\theta \sin\phi, \rho \sin\theta \cos\phi, \rho \cos\theta)$.
- Let $\overrightarrow{T}'_{\theta,\phi}(\rho) = [T_1'(\rho), T_2'(\rho), T_3'(\rho)]$ be the directional derivative of $\overrightarrow{T}$ in the direction (θ, ϕ) (where $'$ indicates derivative with respect to ρ).

$$\begin{aligned} T_1'(\rho) &= \tfrac{d}{d\rho}[\tfrac{f_x(\rho)}{d}] \\ &= \tfrac{A(\rho) f_y - B(\rho) f_t}{d^3} \end{aligned}$$

where

$$A(\rho) = f_x' f_y - f_x f_y', \text{ and } B(\rho) = f_x f_t' - f_x' f_t$$

Three dimensional facet model for video data (Cont.)

- Similarly

$$
\begin{aligned}
T_2'(\rho) &= \frac{C(\rho)f_t - A(\rho)f_x}{d^3} \\
T_3'(\rho) &= \frac{B(\rho)f_x - C(\rho)f_y}{d^3}
\end{aligned}
$$

where

$$C(\rho) = f_y' f_t - f_y f_t'$$

Three dimensional facet model for video data (Cont.)

- Similarly

$$
\begin{aligned}
T_2'(\rho) &= \frac{C(\rho)f_t - A(\rho)f_x}{d^3} \\
T_3'(\rho) &= \frac{B(\rho)f_x - C(\rho)f_y}{d^3}
\end{aligned}
$$

  where

$$C(\rho) = f_y' f_t - f_y f_t'$$

- Let $\overrightarrow{N}_{\theta,\phi}(\rho) = [N_1(\rho), N_2(\rho), N_3(\rho)]$ be a normal to gradient vector $\overrightarrow{T}_{\theta,\phi}(\rho)$ at the point $(\rho \sin\theta \sin\phi,\ \rho \sin\theta \cos\phi, \rho\cos\theta)$.

Three dimensional facet model for video data (Cont.)

- Similarly

$$\begin{aligned} T_2'(\rho) &= \frac{C(\rho)f_t - A(\rho)f_x}{d^3} \\ T_3'(\rho) &= \frac{B(\rho)f_x - C(\rho)f_y}{d^3} \end{aligned}$$

where

$$C(\rho) = f_y' f_t - f_y f_t'$$

- Let $\overrightarrow{N}_{\theta,\phi}(\rho) = [N_1(\rho), N_2(\rho), N_3(\rho)]$ be a normal to gradient vector $\overrightarrow{T}_{\theta,\phi}(\rho)$ at the point $(\rho \sin\theta \sin\phi,\ \rho \sin\theta \cos\phi, \rho\cos\theta)$.
- Then we have

$$\begin{aligned} N_1(\rho) &= f_{xx} - \frac{f_x}{d^2}(f_x f_{xx} + f_y f_{yy} + f_t f_{tt}) \\ &= \frac{D(\rho)f_y - E(\rho)f_t}{d^2} \end{aligned} \quad (1)$$

where

$$D(\rho) = f_{xx}f_y - f_x f_{yy},\ \text{and}\ E(\rho) = f_x f_{tt} - f_{xx} f_t \quad (2)$$

Three dimensional facet model for video data (Cont.)

- Similarly,

$$N_2(\rho) = \frac{F(\rho)f_t - D(\rho)f_x}{d^2} \quad (3)$$

$$N_3(\rho) = \frac{E(\rho)f_x - F(\rho)f_y}{d^2} \quad (4)$$

where

$$F(\rho) = f_{yy}f_t - f_y f_{tt} \quad (5)$$

Three dimensional facet model for video data (Cont.)

- Similarly,

$$N_2(\rho) = \frac{F(\rho)f_t - D(\rho)f_x}{d^2} \quad (3)$$

$$N_3(\rho) = \frac{E(\rho)f_x - F(\rho)f_y}{d^2} \quad (4)$$

where

$$F(\rho) = f_{yy}f_t - f_y f_{tt} \quad (5)$$

- Let $\Theta_{\theta,\phi}(\rho)$ be the rate of change of gradient in the direction orthogonal to the gradient of f at any point $(\rho \sin\theta \sin\phi, \rho \sin\theta \cos\phi, \rho \cos\theta)$. Then

$$\begin{aligned} \Theta_{\theta,\phi}(\rho) &= \overrightarrow{T}' \cdot \overrightarrow{N} \\ &= \frac{AD + BE + CF}{d^3 d'} \end{aligned} \quad (6)$$

where

$$d'^2 = N_1^2 + N_2^2 + N_3^2 \quad (7)$$

Three dimensional facet model for video data (Cont.)

- At origin (i.e., at the candidate pixel over the neighborhood of which the function f is estimated) we calculate the rate of change of gradient of f along orthogonal direction by putting $\rho = 0$ in the equation (6) as

$$\Theta_{\theta,\phi}(0) = \frac{A(0)D(0)+B(0)E(0)+C(0)F(0)}{d^3(0)d'(0)} \tag{8}$$

Three dimensional facet model for video data (Cont.)

- At origin (i.e., at the candidate pixel over the neighborhood of which the function f is estimated) we calculate the rate of change of gradient of f along orthogonal direction by putting $\rho = 0$ in the equation (6) as

$$\Theta_{\theta,\phi}(0) = \frac{A(0)D(0)+B(0)E(0)+C(0)F(0)}{d^3(0)d'(0)} \quad (8)$$

- Now from equation (13) we have

$$\begin{aligned} f_x(0) &= k_2, \quad f_{xx}(0) = 2k_5 \\ f_y(0) &= k_3, \quad f_{yy}(0) = 2k_6 \\ f_t(0) &= k_4, \quad f_{tt}(0) = 2k_7 \end{aligned} \quad (9)$$

and

$$\begin{aligned} f_x'(0) &= 2k_5 \sin\theta \sin\phi + k_8 \sin\theta \cos\phi + k_{10} \cos\theta \\ f_y'(0) &= 2k_6 \sin\theta \cos\phi + k_8 \sin\theta \sin\phi + k_9 \cos\theta \\ f_t'(0) &= 2k_7 \cos\theta + k_9 \sin\theta \cos\phi + k_{10} \sin\theta \sin\phi \end{aligned} \quad (10)$$

Three dimensional facet model for video data (Cont.)

- θ and ϕ are defined based on orthogonal vector ($\overrightarrow{N}$) as

$$\theta = \tan^{-1}\left(\frac{\sqrt{N_1^2 + N_2^2}}{N_3}\right) \text{ and } \phi = \tan^{-1}\left(\frac{N_1}{N_2}\right) \tag{11}$$

Three dimensional facet model for video data (Cont.)

- θ and ϕ are defined based on orthogonal vector ($\overrightarrow{N}$) as

$$\theta = \tan^{-1}\left(\frac{\sqrt{N_1^2 + N_2^2}}{N_3}\right) \text{ and } \phi = \tan^{-1}\left(\frac{N_1}{N_2}\right) \qquad (11)$$

- The point $(0, 0, 0)$ is declared as a space-time interest point if the following two conditions are satisfied:

Three dimensional facet model for video data (Cont.)

- θ and ϕ are defined based on orthogonal vector ($\overrightarrow{N}$) as

$$\theta = \tan^{-1}\left(\frac{\sqrt{N_1^2 + N_2^2}}{N_3}\right) \text{ and } \phi = \tan^{-1}\left(\frac{N_1}{N_2}\right) \qquad (11)$$

- The point $(0, 0, 0)$ is declared as a space-time interest point if the following two conditions are satisfied:
  - The point $(0, 0, 0)$ is a spatio-temporal bounding surface point, and

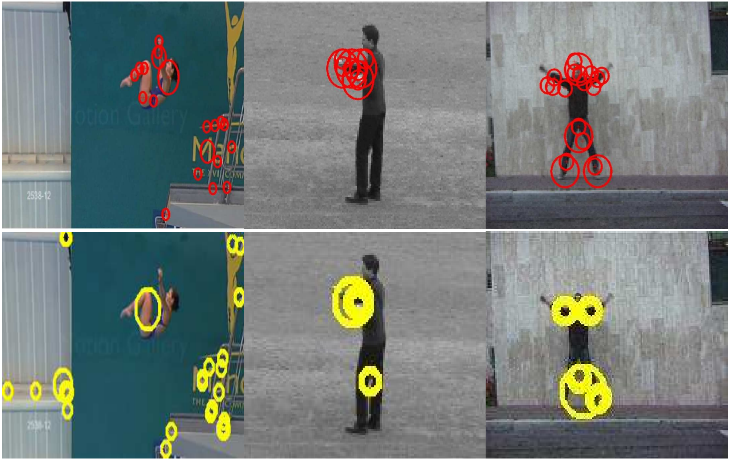
Three dimensional facet model for video data (Cont.)

- θ and ϕ are defined based on orthogonal vector ($\overrightarrow{N}$) as

$$\theta = \tan^{-1}\left(\frac{\sqrt{N_1^2 + N_2^2}}{N_3}\right) \text{ and } \phi = \tan^{-1}\left(\frac{N_1}{N_2}\right) \qquad (11)$$

- The point $(0, 0, 0)$ is declared as a space-time interest point if the following two conditions are satisfied:
  - The point $(0, 0, 0)$ is a spatio-temporal bounding surface point, and
  - For a given threshold Ω, $|\Theta_{\theta,\phi}(0)| > \Omega$

Interest points in video data



UCF sports (lifting) KTH (boxing) Weizmann (pjump)

The points show on the first row using proposed FaSTIP method and second row using Laptev STIP.

Interest point description

- Consider a volume of size $\Delta x \times \Delta y \times \Delta t$ around a interest point

Interest point description

- Consider a volume of size $\Delta x \times \Delta y \times \Delta t$ around a interest point
- Divide the volume into $\eta_x \times \eta_y \times \eta_t$ cells

Interest point description

- Consider a volume of size $\Delta x \times \Delta y \times \Delta t$ around a interest point
- Divide the volume into $\eta_x \times \eta_y \times \eta_t$ cells
- Apply the three-dimensional wavelet transform on each cell up to a desired number of levels

Interest point description

- Consider a volume of size $\Delta x \times \Delta y \times \Delta t$ around a interest point
- Divide the volume into $\eta_x \times \eta_y \times \eta_t$ cells
- Apply the three-dimensional wavelet transform on each cell up to a desired number of levels
- At each level one cell contains low frequency component and the rest seven high frequency components

Interest point description

- Consider a volume of size $\Delta x \times \Delta y \times \Delta t$ around a interest point
- Divide the volume into $\eta_x \times \eta_y \times \eta_t$ cells
- Apply the three-dimensional wavelet transform on each cell up to a desired number of levels
- At each level one cell contains low frequency component and the rest seven high frequency components
- At each cell we calculate the sum of magnitude of positive and negative values (separately) and concatenate them to form a feature vector

Interest point description

- Consider a volume of size $\Delta x \times \Delta y \times \Delta t$ around a interest point
- Divide the volume into $\eta_x \times \eta_y \times \eta_t$ cells
- Apply the three-dimensional wavelet transform on each cell up to a desired number of levels
- At each level one cell contains low frequency component and the rest seven high frequency components
- At each cell we calculate the sum of magnitude of positive and negative values (separately) and concatenate them to form a feature vector
- The low frequency components of each cell is added and are concatenated to form a another vector

Interest point description

- Consider a volume of size $\Delta x \times \Delta y \times \Delta t$ around a interest point
- Divide the volume into $\eta_x \times \eta_y \times \eta_t$ cells
- Apply the three-dimensional wavelet transform on each cell up to a desired number of levels
- At each level one cell contains low frequency component and the rest seven high frequency components
- At each cell we calculate the sum of magnitude of positive and negative values (separately) and concatenate them to form a feature vector
- The low frequency components of each cell is added and are concatenated to form a another vector
- Finally get the feature vector of length $\eta_x \eta_y \eta_t \times (14 \times L + 1)$

Interest point description (cont.)

- For our experiment $\Delta x = \Delta y = 16\sigma$ and $\Delta t = 8\tau$
  - where σ and τ represent the spatial and temporal scales respectively

Interest point description (cont.)

- For our experiment $\Delta x = \Delta y = 16\sigma$ and $\Delta t = 8\tau$
  - where σ and τ represent the spatial and temporal scales respectively
- Divide the neighborhood into **8 cells** ($\eta_x = \eta_y = \eta_t = 2$)

Interest point description (cont.)

- For our experiment $\Delta x = \Delta y = 16\sigma$ and $\Delta t = 8\tau$
  - where σ and τ represent the spatial and temporal scales respectively
- Divide the neighborhood into **8 cells** $(\eta_x = \eta_y = \eta_t = 2)$
- Apply three-dimensional wavelet transform up to **2 levels**

Interest point description (cont.)

- For our experiment $\Delta x = \Delta y = 16\sigma$ and $\Delta t = 8\tau$
  - where σ and τ represent the spatial and temporal scales respectively
- Divide the neighborhood into **8 cells** ($\eta_x = \eta_y = \eta_t = 2$)
- Apply three-dimensional wavelet transform up to **2 levels**
- Finally, describe each interest points by a feature vector of length **232**

Experimental evaluation

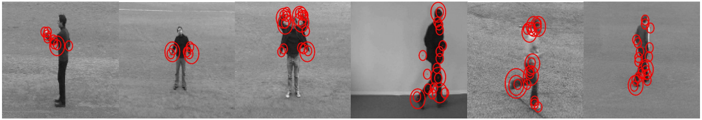
- We have tested our method on three state-of-the-art human action dataset: UCF sports, KTH and Weizmann

Experimental evaluation

- We have tested our method on three state-of-the-art human action dataset: UCF sports, KTH and Weizmann
- UCF sports dataset contain 10 sports activities: diving, golf swinging, kicking (a ball), weight-lifting, horse riding, running, skating, swinging (on the floor), waking and swinging (at the high bar)

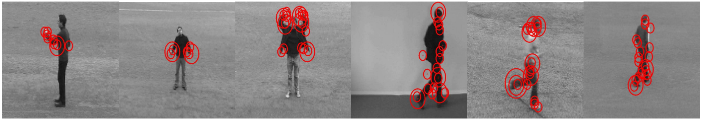
Experimental evaluation (cont.)

- KTH dataset consists of six common human activities: boxing, hand clapping, hand waving, jogging, running and walking

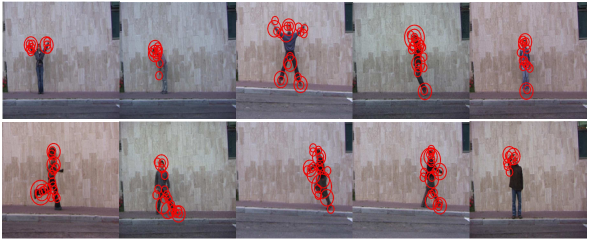


Experimental evaluation (cont.)

- KTH dataset consists of six common human activities: boxing, hand clapping, hand waving, jogging, running and walking



- Weizmann data has ten classes: two-hands waving, bending, jumping jack, jumping, jumping in place, running, sideways, skipping, walking and one-hand waving



Experimental evaluation (cont.)

- For each dataset, we randomly select different number of points to build the vocabulary

Experimental evaluation (cont.)

- For each dataset, we randomly select different number of points to build the vocabulary
- We use multi-channel non-linear SVM with a χ^2-kernel [7] for classification

Experimental evaluation (cont.)

- For each dataset, we randomly select different number of points to build the vocabulary
- We use multi-channel non-linear SVM with a χ^2-kernel [7] for classification
- Run the classier for different vocabulary size and report the result for optimal vocabulary size for each dataset

Experimental results on UCF sports dataset

- Randomly select 100000 points to build the vocabulary

Experimental results on UCF sports dataset

- Randomly select 100000 points to build the vocabulary
- We use leave-one-out cross validation strategy and get 87.33% accuracy with 1200 as optimal vocabulary size

Experimental results on UCF sports dataset

- Randomly select 100000 points to build the vocabulary
- We use leave-one-out cross validation strategy and get 87.33% accuracy with 1200 as optimal vocabulary size

Approach	Year	Accuracy(%)
Rodriguez et al. [11]	2008	69.20
Yeffet & Wolf [15]	2009	79.30
Wang et al. [14]	2009	85.60
Kovashka & Grauman [6]	2010	87.27
Wang et al. [13]	2011	88.20
Guha & Ward [5]	2012	83.80
Our approach		87.33

Comparison of results with the state-of-the-art for UCF sports dataset

Experimental results on KTH dataset

- Randomly select 200000 points to build the vocabulary

[6] Laptev et al., On Space-Time Interest Points, IJCV, 2005

Experimental results on KTH dataset

- Randomly select 200000 points to build the vocabulary
- We follow the author suggested[6] training, validation and test data partition and obtain average accuracy of 93.51%.

[6] Laptev et al., On Space-Time Interest Points, IJCV, 2005

Experimental results on KTH dataset

- Randomly select 200000 points to build the vocabulary
- We follow the author suggested[6] training, validation and test data partition and obtain average accuracy of 93.51%.
- The optimal vocabulary size is 4000

Approach	Year	Accuracy(%)
Schuldt et al. [12]	2004	71.72
Dollár et al. [3]	2005	81.17
Nowozin et al. [10]	2007	84.72
Laptev et al. [7]	2008	91.80
Niebles et al. [9]	2008	81.50
Bregonzio et al. [1]	2009	93.17
Kovashka & Grauman [6]	2010	94.53
Wang et al. [13]	2011	94.20
Our approach		93.51

Comparison of results with the state-of-the-art for KTH dataset

[6] Laptev et al., On Space-Time Interest Points, IJCV, 2005

Experimental results on Weizmann dataset

- Randomly select 30000 points to build the vocabulary

Experimental results on Weizmann dataset

- Randomly select 30000 points to build the vocabulary
- We have tested on Weizmann dataset with leave-one-out cross validation scheme and get on an average 96.67% accuracy

Experimental results on Weizmann dataset

- Randomly select 30000 points to build the vocabulary
- We have tested on Weizmann dataset with leave-one-out cross validation scheme and get on an average 96.67% accuracy

Approach	Year	Accuracy(%)
Dollár et al. [3]	2005	85.20
Gorelick et al. [4]	2007	97.80
Niebles et al. [9]	2008	90.00
Zhe Lin et al. [8]	2009	100.00
Bregonzio et al. [2]	2012	96.67
Guha & Ward [5]	2012	98.90
Our approach		96.67

Comparison of results with the state-of-the-art for Weizman dataset

Comparison with other state-of-the-art STIP points based method

- We compare our results with interest points based activity classification schemes like popular STIP[7], Cuboid[8] and achieve much better performance

---

7 Laptev et al., On Space-Time Interest Points, IJCV, 2005

8 Dollar et al., Behavior Recognition via Sparse Spatio-Temporal Features, VS-PETS, 2005

Comparison with other state-of-the-art STIP points based method

- We compare our results with interest points based activity classification schemes like popular STIP[7], Cuboid[8] and achieve much better performance

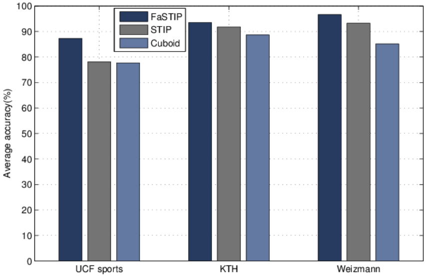


Figure: Comparison results with STIP and Cuboid

[7] Laptev et al., On Space-Time Interest Points, IJCV, 2005

[8] Dollar et al., Behavior Recognition via Sparse Spatio-Temporal Features, VS-PETS, 2005

Conclusion

Conclusion

- We present a new model for space-time interest point detection and description.

Conclusion

- We present a new model for space-time interest point detection and description.
- Experimental results shows that the performance of our system is comparable to the state-of-the-art methods.

Conclusion

- We present a new model for space-time interest point detection and description.
- Experimental results shows that the performance of our system is comparable to the state-of-the-art methods.
- Though our method marginally falls behind the best result only in a few classes but we achieves far better performance compared the other state-of-the-art STIP methods.

Conclusion

- We present a new model for space-time interest point detection and description.
- Experimental results shows that the performance of our system is comparable to the state-of-the-art methods.
- Though our method marginally falls behind the best result only in a few classes but we achieves far better performance compared the other state-of-the-art STIP methods.
- Our FaSTIP is supposed to perform better compared to STIP and Cuboid on others applications too.

THANKS

Matteo Bregonzio, Shaogang Gong, and Tao Xiang.
Recognising action as clouds of space-time interest points.
In *CVPR*, 2009.

Matteo Bregonzio, Tao Xiang, and Shaogang Gong.
Fusing appearance and distribution information of interest points for action recognition.
Pattern Recognition, 45(3):1220–1234, 2012.

Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie.
Behavior recognition via sparse spatio-temporal features.
In *VS-PETS*, October 2005.

Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri.
Actions as space-time shapes.
IEEE Trans. PAMI, 29(12):2247–2253, 2007.

Tanaya Guha and Rabab Kreidieh Ward.
Learning sparse representations for human action recognition.

IEEE Trans. PAMI, 34(8):1576–1588, 2012.

Adriana Kovashka and Kristen Grauman.
Learning a hierarchy of discriminative space-time neighborhood features for human action recognition.
In *CVPR*, June 2010.

Ivan Laptev, Marcin Marszaek, Cordelia Schmid, and Benjamin Rozenfeld.
Learning realistic human actions from movies.
In *CVPR*, 2008.

Zhe Lin, Zhuolin Jiang, and Larry S. Davis.
Recognizing actions by shape-motion prototype trees.
In *ICCV*, 2009.

Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei.
Unsupervised learning of human action categories using spatial-temporal words.
International Journal of Computer Vision, 79(3):299–318, 2008.

Sebastian Nowozin, Gk̈han Bakir, and Koji Tsuda.

Discriminative subsequence mining for action classification.
In *ICCV*, 2007.

Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah.
Action mach: A spatio-temporal maximum average correlation height filter for action recognition.
In *CVPR*, 2008.

Christian Schuldt, Ivan Laptev, and Barbara Caputo.
Recognizing human actions: A local svm approach.
In *ICPR*, 2004.

Heng Wang, Alexander Kläser, Cordelia Schmid, and Liu Cheng-Lin.
Action recognition by dense trajectories.
In *CVPR*, pages 3169–3176, June 2011.

Heng Wang, Muhammad Muneeb Ullah, Alexander Klašer, Ivan Laptev, and Cordelia Schmid.
Evaluation of local spatio-temporal features for action recognition.
In *BMVC*, 2009.

Lahav Yeffet and Lior Wolf.
Local trinary patterns for human action recognition.
In *ICCV*, pages 492–497, 2009.