EEE422/6082 Computational Vision

Action Recognition

Ling Shao

Action Recognition

- What is action?
- Global descriptors
  - Motion Energy Images and Motion History Images
  - Optical flow based descriptors
- Local descriptors
  - Space-time interest points + Local 3D path descriptors

What is Action?

- Action: Atomic motion(s) that can be clearly distinguished and usually has a semantic association (e.g. sitting down, running).
- Activity: An activity contains several actions performed in succession (e.g. dining, meeting a person).
- Event is a composite of activities (e.g. football match, traffic accident).

Motion Template based Approach

- Goal:
  - To create a compact template representation of an action over time.
  - Construct a vector-image suitable for matching other instances of actions.
- Reference:
  - Bobick, A.F.; Davis, J.W., "*The recognition of human movement using temporal templates*," *IEEE PAMI* , vol.23, no.3, pp.257-267, 2001

Motion Energy Image (MEI)

- MEI is a binary Image indicating the spatial distribution of a motion over a time duration.

$$E_\tau(x,y,t) = \bigcup_{i=0}^{\tau-1} B(x,y,t-i)$$

B: binary image indicating motion locations

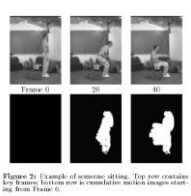


Figure 2: Example of someone sitting. Top row contains key frames; bottom row is cumulative motion images starting from Frame 0.

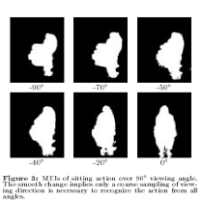
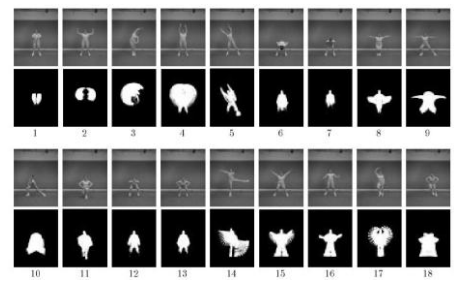


Figure 3: MEIs of sitting action over 90° viewing angle. The smooth change implies only a coarse sampling of viewing direction is necessary to recognize the action from all angles.

Motion Energy Images



Motion History Images (MHI)

- MEI indicates the location of motion only
- Motion History Image (MHI) represents how motion i moving, In an MHI, pixel intensity is a function of the motion history at that location, where brighter values correspond to more recent motion.

$$H_\tau(x,y,t) = \begin{cases} \tau & B(x,y,t) = 1 \\ \max(0, H_\tau(x,y,t-1) - 1) & B(x,y,t) = 0 \end{cases}$$

- Descriptor: Build a 2-component vector image by combining MEI and MH Images

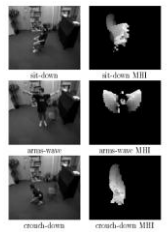


Figure 4: Action moves along with their MHIs used in a real-time system.

Descriptor and Matching

- Compute the 7 Hu moments
  - Refer to the paper with more details
- Model the 7 moments each action class with a Gaussian distribution (diagonal covariance)
- Given a new action instance: measure the Mahalanobis distance to all classes. Pick the nearest one.

Computing Moments

- Image Moments

$$M_{ij} = \sum_x \sum_y x^i y^j I(x,y)$$

- Translation Invariant Moments

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x,y)$$

$$\bar{x} = \frac{M_{10}}{M_{00}} \qquad \bar{y} = \frac{M_{01}}{M_{00}}$$

- Scale Invariant Moment

$$\eta_{ij} = \frac{\mu_{ij}}{\mu_{00}^{\left(1+\frac{i+j}{2}\right)}}$$

Computing Moments

- Seven Moments (rotation, scaling, translation invariant); the first two: i+j = 2, the rest i+j=3.

$$I_1 = \eta_{20} + \eta_{02}$$
$$I_2 = (\eta_{20} - \eta_{02})^2 + (2\eta_{11})^2$$
$$I_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$$
$$I_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$$
$$I_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$$
$$I_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$$
$$I_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2].$$

		Closest Dist	Closest Move	Correct Dist	Median Dist	Rank
Test	1	1.43	4	1.44	2.35	2
	2	3.14	2	3.14	12.00	1
	3	3.08	3	3.08	8.39	1
	4	0.47	4	0.47	2.11	1
	5	6.84	5	6.84	19.24	1
	6	0.32	10	0.61	0.64	7
Test	7	0.97	7	0.97	2.03	1
	8	20.47	8	20.47	35.89	1
	9	1.05	8	1.77	2.37	4
	10	0.14	10	0.14	0.72	1
	11	0.24	11	0.24	1.01	1
	12	0.79	12	0.79	4.42	1
Test	13	0.13	6	0.25	0.51	3
	14	4.01	14	4.01	7.98	1
	15	0.34	15	0.34	1.84	1
	16	1.03	15	1.04	1.59	2
	17	0.65	17	0.65	2.18	1
	18	0.48	10	0.51	0.94	4

Table 1: Test results using one camera at 30° off frontal. Each row corresponds to one test move and gives the distance to the nearest move (and its index), the distance to the correct matching move, the median distance, and the ranking of the correct move.

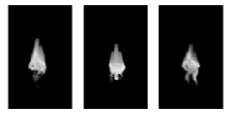


Figure 6: An example of MHIs with similar statistics. (a) Test input of move 13 at 30°. (b) Closest match which is move 6 at 0°. (c) Correct match.

Optical Flow based Approach

- **Alexei A. Efros, Alexander C. Berg, Greg Mori and Jitendra Malik Recognizing Action at a Distance, ICCV03, October 2003.**

Optical Flow based Descriptors

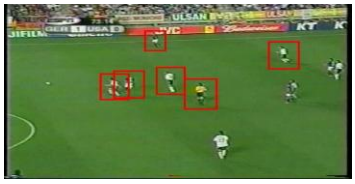
- Challenges:
  - Low resolution, Only a few pixel wide
  - Noisy data
  - Moving camera, occlusions
  - Wide range of actions (including non-periodic)
- Solution
  - Computing an optical flow template of a moving figure
  - Building descriptors from the template.
  - spatio-temporal volume for each stabilized human figure

Alexei A. Efros, Alexander C. Berg, Greg Mori and Jitendra Malik Recognizing Action at a Distance, ICCV03, October 2003.

Overview of the System

- Tracking and stabilizing
  - Result in a stabilized figure-centric motion sequence.
- Compute the spatial-temporal motion descriptor
  - Compute the Optical Flow descriptor
- Matching and Classification
- Features of Motion-based approach
  - Non-parametric; use large amount of data
  - Classify a novel motion by finding the most similar motion from the training set

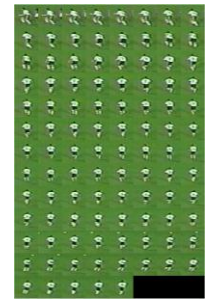
Gathering Action data



- Tracking
  - Simple correlation-based tracker
  - User-initialized

Figure-centric Representation

- Stabilized spatio-temporal volume
  - No translation information
  - All motion caused by person's limbs
    - ⇨Good news: indifferent to camera motion
    - ⇨Bad news: hard!
- Good test to see if actions, not just translation, are being captured

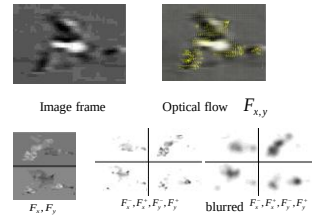


How to Describe Motion?

- Appearance
  - Not preserved across different clothing
- Gradients (spatial, temporal)
  - same (e.g. contrast reversal)
- Edges/Silhouettes
  - Too unreliable
- Optical flow
  - Explicitly encodes motion
  - Least affected by appearance
  - …but too noisy

Spatial Motion Descriptor

1. **Computing the optical flow**
2. **Separate Fx, and Fy**
3. **Half-wave rectification of each elements into four channels**
4. **Blurr motion channels**

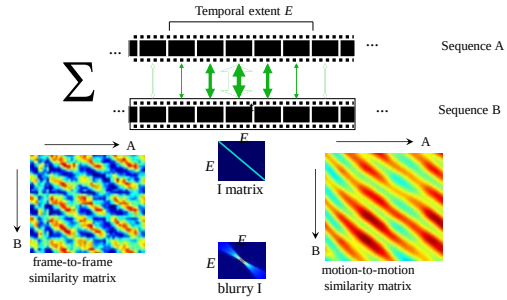


Computing Similarity

$$S(i,j) = \sum_{t \in T} \sum_{c=1}^{4} \sum_{x,y \in I} a_c^{i+t}(x,y) b_c^{j+t}(x,y)$$

- **T: time duration motion length I: spatial extents**
- **c: # of channels**
- **a,b: the blurred motion channels of each frame for two different sequences**

Spatio-temporal Motion Descriptor



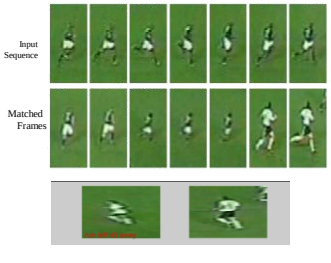
Matching and Classification

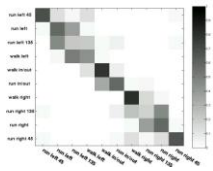
- Compute the similarity matrix
- Blurring the similarity matrix with the temporal kernel (convolution)
- For each frame of the novel sequence, the maximum score in the corresponding row of this matrix will indicate the best match to the motion descriptor centered at this frame.
- Classify this frame using a k-nearest-neighbor classier: find the k best matches from labeled data and take the majority label

Results

- **Ballet (16 Classes):**
  - Clips of motions from an instructional video.
  - Professional dancers, two men and two women.
  - Perform in mostly standard ballet moves.
- **Tennis (6 Classes):**
  - Two amateur tennis players outdoors (one player test, one player train).
  - Each player was video-taped on different days in different locations with slightly different camera positions.
  - Players about 50 pixels tall.
- **Football (8 Classes):**
  - Several minutes of a World Cup football game from an NTSC video tape.
  - Wide angle of the playing field.
  - Substantial camera motion and zoom.
  - About 30-by-30 noisy pixels per human figure.

Football Actions: Matching/Classification



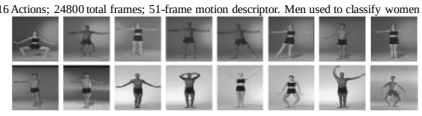


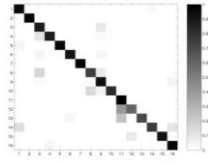
10 actions; 4500 total frames;
13-frame motion descriptor

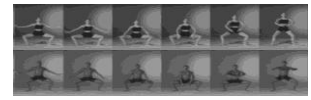
Football (K=1, T=13): [.67 .58 .68 .79 .59 .68 .58 .66]

Classifying Ballet Actions

16 Actions; 24800 total frames; 51-frame motion descriptor. Men used to classify women and vice versa.







Values on the diagonals:
Ballet (K=5, T=51): [.94 .97 .88 .88 .97 .91 1 .74 .92 .82 .99 .62 .71 .76 .92 .96]

Classifying Tennis Actions

6 actions; 4600 frames; 7-frame motion descriptor
Woman player used as training, man as testing.

Tennis (K=5, T=7): [.46 .64 .7 .76 .88 .42]

Bag of Space-time local descriptors

Reference:

[1] Ivan Laptev, On Space-Time Interest Points, International Journal of Computer Vision, v.64 n.2-3, p.107-123, 2005

[2] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, Benjamin Rozenfeld Learning realistic human actions from movies, CVPR2008

Action: Bag of STIP

- STIP: Space-Time Interest Points
- Goal: Interpretation of dynamic scenes

Common methods:
- ~~Camera stabilization~~
- ~~Segmentation~~
- ~~Tracking~~

Common problems:
- Complex & changing BG
- Appearance of new OBJ

⇒ *No global assumptions about the scene*

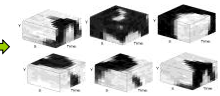
- Reference: Ivan Laptev, On Space-Time Interest Points, International Journal of Computer Vision, v.64 n.2-3, p.107-123

Action: Bag of STIP: Overview

Bag of space-time features + multi-channel SVM

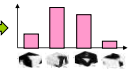
[Schuldt'04, Niebles'06, Zhang'07]

Collection of space-time patches



HOG & HOF patch descriptors

Histogram of visual words



Multi-channel SVM Classifier

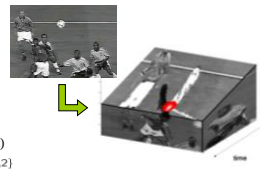
Space-Time Features: Detector

- Space-time corner detector

[Laptev, IJCV 2005]

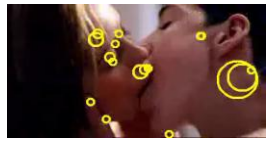
$$H = \det(\mu) + k\,\mathrm{tr}^3(\mu)$$

$$\mu = \begin{pmatrix} I_x I_x & I_x I_y & I_x I_t \\ I_x I_y & I_y I_y & I_y I_t \\ I_x I_t & I_y I_t & I_t I_t \end{pmatrix} * g(\cdot;\sigma,\tau)$$



- Dense scale sampling (no explicit scale selection)

$$(\sigma^2,\tau^2) = \mathcal{S}\times\mathcal{T},\ \mathcal{S} = 2^{\{2,\ldots,6\}},\mathcal{T} = 2^{\{1,2\}}$$



Space-Time Features: Detector

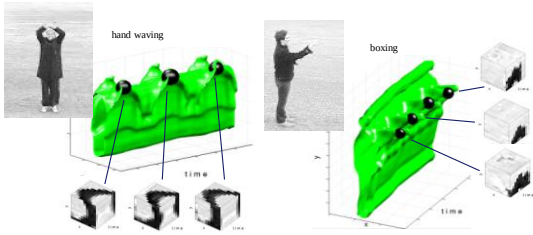
No global assumptions ⇒ Consider local spatio-temporal neighborhoods

hand waving

boxing

Space-time

No global assumptions ⇒ Consider local spatio-temporal neighborhoods



Space-Time Interest Points

What neighborhoods to consider?

Distinctive neighborhoods ⇒ High image variation in *space* and *time* ⇒ Look at the distribution of the gradient

Space-time gradient $\nabla L = (L_x, L_y, L_t)^T$

$$L_\xi(\cdot;\Sigma) = f(\cdot) * g_\xi(\cdot;\Sigma)$$

$$g_\xi(\bar{x};\Sigma) = \partial_\xi \left(\frac{e^{-\frac{1}{2}p^T\Sigma^{-1}p}}{2\pi\sqrt{\det\Sigma}} \right)$$

Covariance $\Sigma = \begin{pmatrix} c_{xx} & c_{xy} & c_{xt} \\ c_{xy} & c_{yy} & c_{xt} \\ c_{xt} & c_{yt} & c_{tt} \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \tau^2 \end{pmatrix}$

Spatial scale σ, temporal scale τ

Space-Time interest points

Distribution of ∇L within a local neighborhood

Second-moment matrix $\mu(\cdot;\Sigma) = \nabla L(\cdot;\Sigma)(\nabla L(\cdot;\Sigma))^T * g(\cdot;s\Sigma)$

$$= \begin{pmatrix} \mu_{xx} & \mu_{xy} & \mu_{xt} \\ \mu_{xy} & \mu_{yy} & \mu_{yt} \\ \mu_{xt} & \mu_{yt} & \mu_{tt} \end{pmatrix}$$

High variation of ∇L ⇒ large eigenvalues of μ

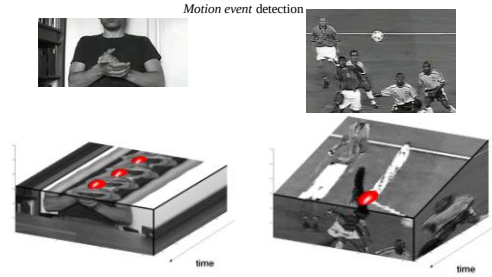
⇒ Local maxima of H over (x,y,t)

$$H(p;\Sigma) = \det(\mu(p;\Sigma)) + k\,\mathrm{trace}^3(\mu(p;\Sigma))$$
$$= \lambda_1\lambda_2\lambda_3 - k(\lambda_1+\lambda_2+\lambda_3)^3$$

(similar to Harris operator [Harris and Stephens, 1988])

Space-Time Interest Points

Motion event detection

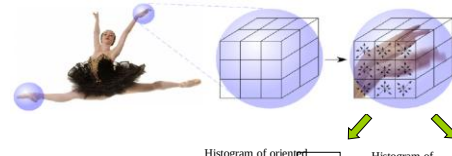


Space-Time interest points

Motion event detection: complex background

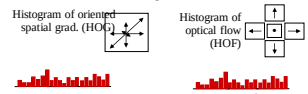
Space-Time Features: Descriptor

Multi-scale space-time patches from corner detector



Histogram of oriented spatial grad. (HOG)

Histogram of optical flow (HOF)



3x3x2x4bins **HOG** descriptor

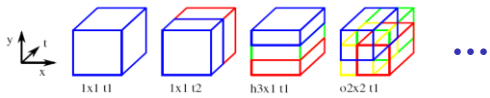
3x3x2x5bins **HOF** descriptor

Ref: Ivan Laptev, Marcin Marszalek, Cordelia Schmid, Benjamin Rozenfeld, Learning realistic human actions from movies, CVPR 2008

Spatio-temporal bag-of-features

We use global spatio-temporal grids

- In the spatial domain:
  - 1x1 (standard BoF)
  - 2x2, o2x2 (50% overlap)
  - h3x1 (horizontal), v1x3 (vertical)
  - 3x3
- In the temporal domain:
  - t1 (standard BoF), t2, t3



Multi-channel chi-square kernel

We use SVMs with a multi-channel chi-square kernel for classification

$$K(H_i, H_j) = \exp\Big(-\sum_{c \in C} \frac{1}{A_c} D_c(H_i, H_j)\Big)$$

- Channel c is a combination of a detector, descriptor and a grid
- $D_c(H_i, H_j)$ is the chi-square distance between histograms
- A_c is the mean value of the distances between all training samples
- The best set of channels C for a given training set is found based on a greedy approach

Combining channels

Task	HoG BoF	HoF BoF	Best chan.	Best comb.
KTH multi-class	81.6%	89.7%	91.1%	91.8%
Action AnswerPhone	13.4%	24.6%	26.7%	32.1%
Action GetOutCar	21.9%	14.9%	22.5%	41.5%
Action HandShake	18.6%	12.1%	23.7%	32.3%
Action HugPerson	29.1%	17.4%	34.9%	40.6%
Action Kiss	52.0%	36.5%	52.0%	53.3%
Action SitDown	29.1%	20.7%	37.8%	38.6%
Action SitUp	6.5%	5.7%	15.2%	18.2%
Action StandUp	45.4%	40.0%	45.4%	50.5%

Table: Classification performance of different channels and their combinations

⇒
- It is worth trying different grids
- It is beneficial to combine channels

Evaluation of Spatio-temporal Grids

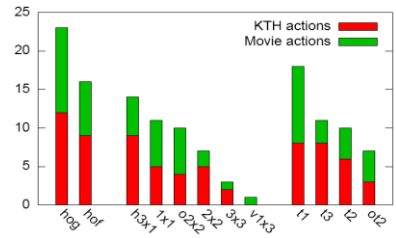


Figure: Number of occurrences for each channel component within the optimized channel combinations for the KTH action dataset and our manually labeled movie dataset

Comparison to the state-of-the-art

Figure: Sample frames from the KTH actions sequences, all six classes (columns) and scenarios (rows) are presented

Comparison to the state-of-the-art

Method	Schuldt et al.	Niebles et al.	Wong et al.	Nowozin et al.	ours
Accuracy	71.7%	81.5%	86.7%	87.0%	91.8%

Table: Average class accuracy on the KTH actions dataset

	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	.99	.01	.00	.00	.00	.00
Jogging	.04	.89	.07	.00	.00	.00
Running	.01	.19	.80	.00	.00	.00
Boxing	.00	.00	.00	.97	.00	.03
Waving	.00	.00	.00	.00	.91	.09
Clapping	.00	.00	.00	.05	.00	.95

Table: Confusion matrix for the KTH actions

Training noise robustness

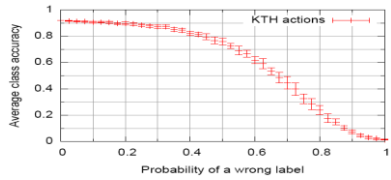


Figure: Performance of our video classification approach in the presence of wrong labels

- Up to $p=0.2$ the performance decreases insignicantly
- At $p=0.4$ the performance decreases by around 10%

Action recognition in real-world videos

Figure: Example results for action classification trained on the automatically annotated data. We show the key frames for test movies with the highest confidence values for true/false pos/neg

Action recognition in real-world videos

- Note the suggestive FP: hugging or answering the phone
- Note the dicult FN: getting out of car or handshaking

Action recognition in real-world videos

	Clean	Automatic	Chance
AnswerPhone	32.1%	16.4%	10.6%
GetOutCar	41.5%	16.4%	6.0%
HandShake	32.3%	9.9%	8.8%
HugPerson	40.6%	26.8%	10.1%
Kiss	53.3%	45.1%	23.5%
SitDown	38.6%	24.8%	13.8%
SitUp	18.2%	10.4%	4.6%
StandUp	50.5%	33.6%	22.6%

Table: Average precision (AP) for each action class of our test set. We compare results for clean (annotated) and automatic training data. We also show results for a random classifier (chance)