Learning Realistic Human Actions from Movies

I. Laptev, M. Marszałek, C. Schmid and B. Rozenfeld. CVPR 2008.

Presented by: Islam Beltagy

Girish Malkarnenkar

Experiment presentation for CS 395T

9th November 2012

Actions in movies

- Realistic variation of human actions
- Many classes and many examples per class





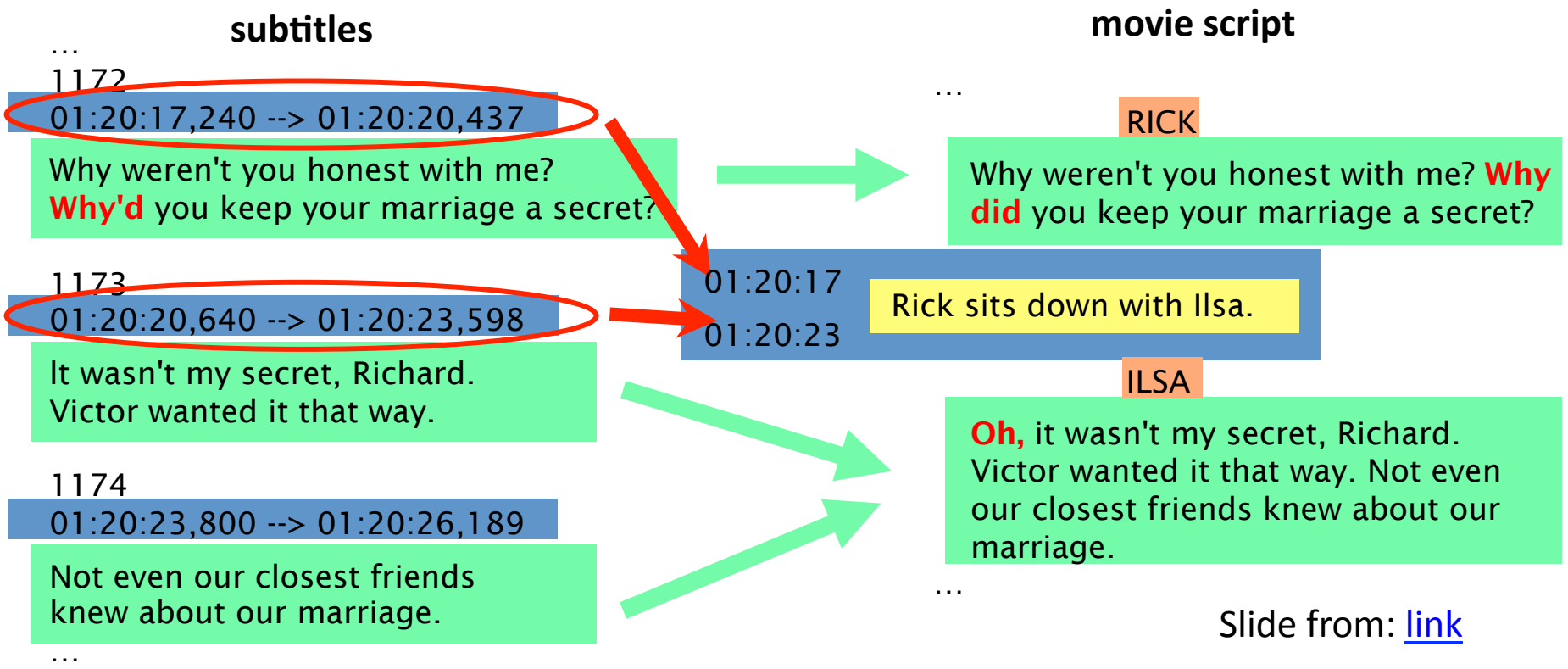
Problems:

- Typically only a few class-samples per movie
- Manual annotation is very time consuming

Slide from: link

Automatic video annotation using scripts [Everingham et al. BMVC06]

- Scripts available for >500 movies (no time synchronization)
  www.dailyscript.com, www.movie-page.com, www.weeklyscript.com …
- Subtitles (with time info.) are available for the most of movies
- Can transfer time to scripts by text alignment

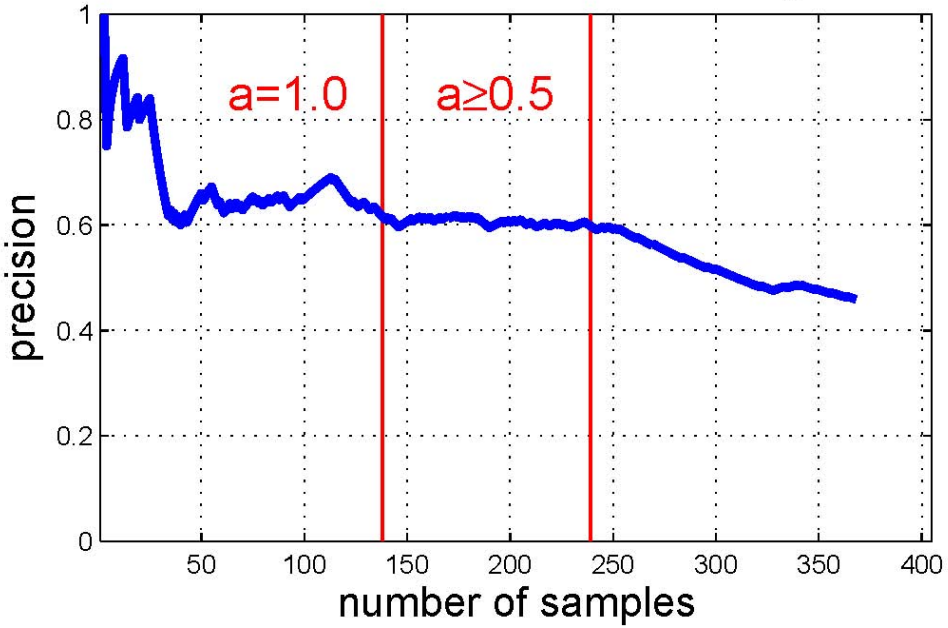


Slide from: link

Script alignment: Evaluation

- Annotate action samples *in text*
- Do automatic script-to-video alignment
- Check the correspondence of actions in scripts and movies

Evaluation of retrieved actions on visual ground truth



a: quality of subtitle-script matching

Example of a "visual false positive"

A black car pulls up, two army officers get out.

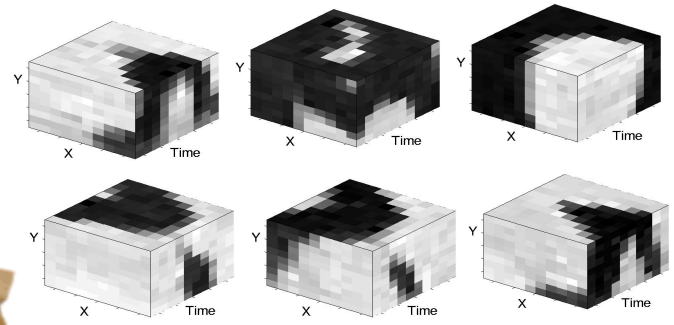
Slide from: link

Action Classification: Overview

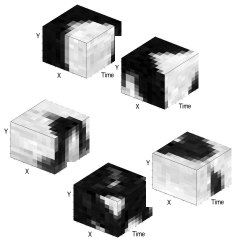
Bag of space-time features + multi-channel SVM

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

Collection of space-time patches

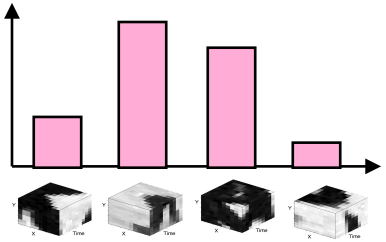


Visual vocabulary



HOG & HOF patch descriptors

Histogram of visual words



Multi-channel SVM Classifier

Slide from: link

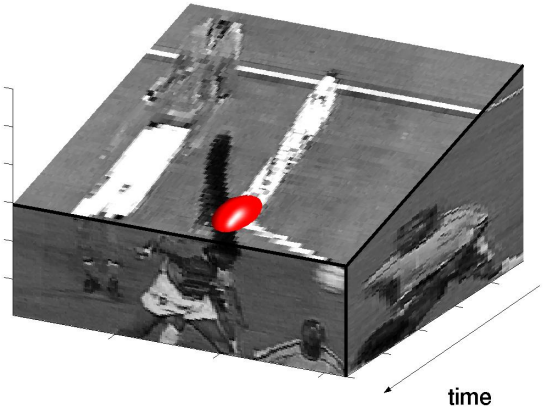
Space-Time Features: Detector

- Space-time corner detector

  [Laptev, IJCV 2005]

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

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



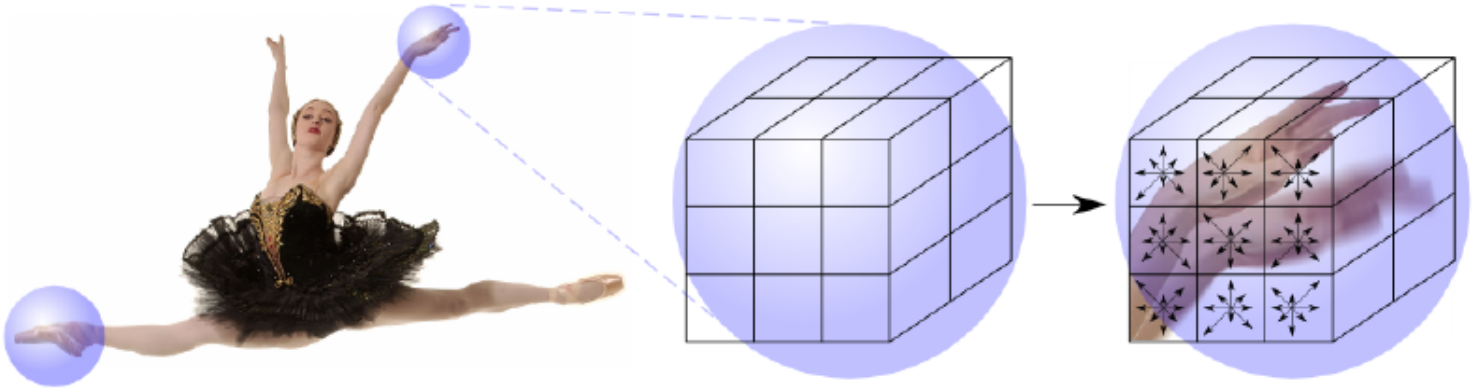
- Dense scale sampling (no explicit scale selection)

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

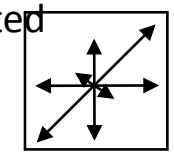
Slide from: link

Space-Time Features: Descriptor

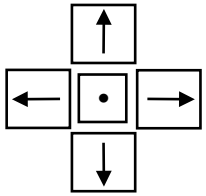
Multi-scale space-time patches from corner detector



Histogram of oriented spatial grad. (HOG)



Histogram of optical flow (HOF)



3x3x2x4bins **HOG** descriptor

3x3x2x5bins **HOF** descriptor

Public code available at www.irisa.fr/vista/actions

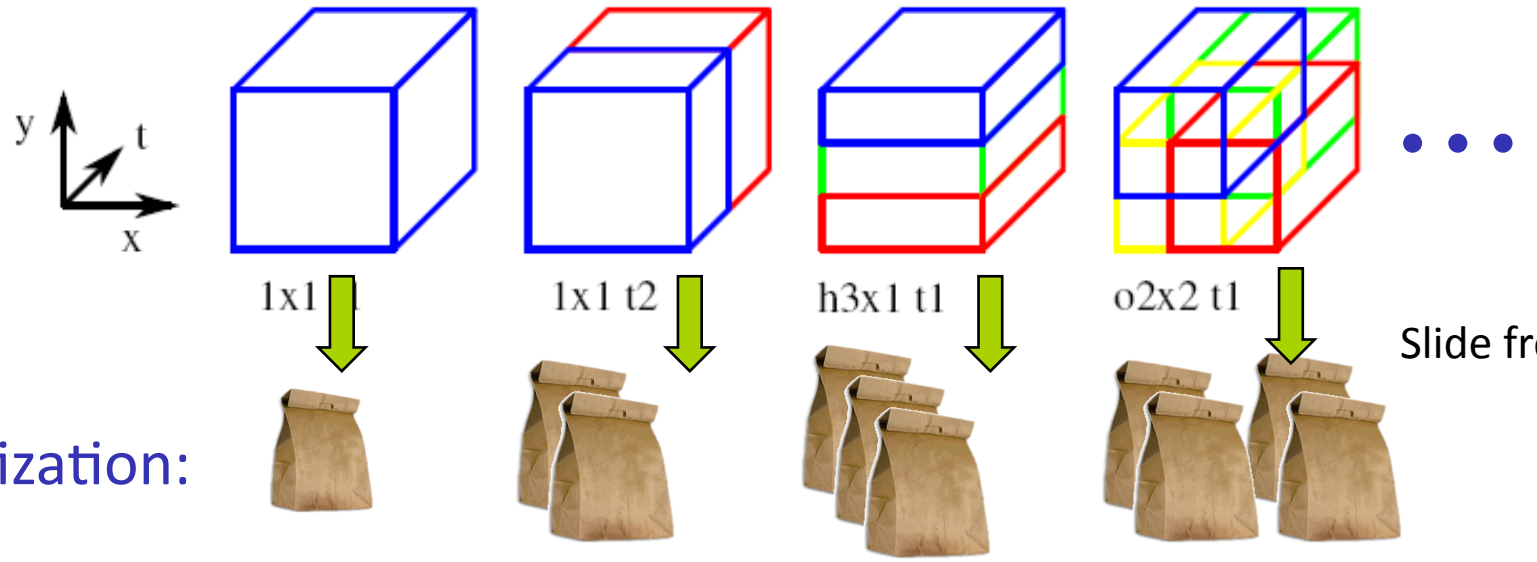
Slide from: link

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

Figure: Examples of a few spatio-temporal grids



Quantization:

Slide from: link

Multi-channel chi-square kernel

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

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

- 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

Slide from: link

STIP in Action!

- Link to a 2min video showing the author's CVPR 2008 paper results [notice the subtitle dialogue and human action/screenplay information]

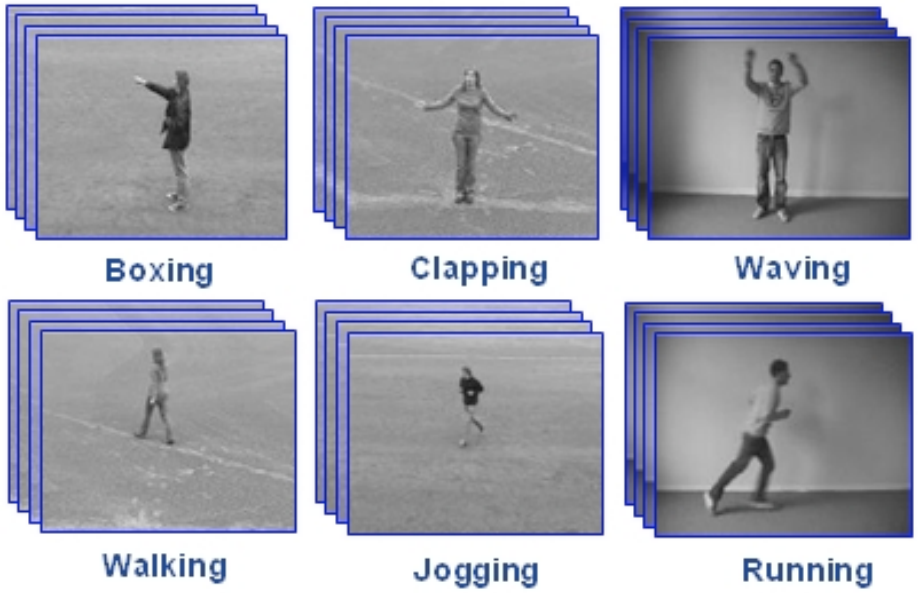
Examples of STIP detections

- AnswerPhone
- GetOutCar
- HugPerson
- Kiss
- SitDown

For the Hollywood Dataset, **STIPs are calculated only for specified start & end frames** from the annotations file & not for the whole clip, unlike the KTH action clips…

Experimental Dataset 1: KTH Actions

- **6 classes** of 100 clips each [64 training & 36 testing]
- **Same size/split as used in the CVPR 2008 paper**



Link

KTH Dataset examples

- Boxing
- Hand-Clapping
- Hand-Waving
- Jogging
- Running
- Walking

KTH Training & Testing split are based on making sure that the **same person (actor)** doesn't appear in both training & testing!

Between which 2 categories do you expect the most confusion in a 6 way multi-classification task?

Experimental Dataset 2: Hollywood

- Selected a **subset** of the dataset used in the paper
- **4 classes** with 18 videos each [9 training & 9 testing]

Hollywood Dataset examples

- GetOutCar
- HandShake
- Kiss
- Stand-Up

Hollywood Training & Testing split are based on making sure that clips from the **same movie** don't appear in both training & testing!

Between which 2 categories do you expect the most confusion in a 4 way multi-classification task?

Experiment 1: HoG & HoF

- **Goal**: See the effect of HoG, HoF and HoG +HoF on KTH & Hollywood
- Did a simple bag of features approach over the full video
- 100k features randomly sampled from the total of ~300k (HoG | HoF | HoG+HoF) descriptors) to form 4000 clusters
- Used kchi2 kernel for SVM based multi-classification (one against one)

Classification Accuracy

Dataset (classes*tests per class)	HoG	HoF	HoG+HoF
KTH (6*36=216)	69.44% (150)	**81.94%** (177)	79.17% (171)
Hollywood (4*9=36)	**44.44%** (16)	30.56% (11)	33.33% (12)

Discussion: KTH v/s Hollywood…

- Reason behind higher multi-classification accuracy achieved on **KTH (~82%)** than on **Hollywood (~44%)**?

- **KTH is "easier" than Hollywood**: homogenous background + choreographed actions
- Hollywood dataset: variability in scale/ viewpoint/background

Discussion: HOG v/s HOF

•Similar to the results obtained in the paper

•**HoG performs better for Hollywood** perhaps because **HoG captures context & image content better than HoF** and these play an important role in realistic settings

•Simple actions (like in KTH) can be well represented by their motion only (i.e. HoF)

Data	HOG	HOF
KTH	69.44	**81.94**
Hollywood	**44.44**	30.56

Discussion: HoG+HoF

- Combining HoG and HoF didn't help a lot over either.
- I used a simple 1x1x1 BoF approach for binning (just a single channel)

- Paper explores better combinations based on various binning/spatio-temporal grids & combines the best channels using a greedy approach and a multi channel SVM

Best KTH Confusion Matrix [HoF]

%	BOXING	CLAPPING	WAVING	JOGGING	RUNNING	WALKING
BOXING	39	**47**	0	0	0	14
CLAPPING	0	**100**	0	0	0	0
WAVING	0	11	89	0	0	0
JOGGING	0	0	0	81	13	**16**
RUNNING	0	0	0	**17**	83	0
WALKING	0	0	0	0	0	**100**

81.94%

Examples of confusion in KTH

- Detected correctly as boxing
- Detected correctly as clapping
- Detected wrongly as clapping (true= boxing)

- Notice the leg motion in boxing that helps differentiation from clapping. Most of the errors happen for cases when this leg motion is missing.

Best Hollywood Confusion Matrix [HoG]

%	GetOutCar	HandShake	Kiss	Stand-Up
GetOutCar	**56**	0	0	**44**
HandShake	22	11	0	**67**
Kiss	0	0	11	**89**
Stand-Up	0	0	0	**100**

44.44%

Examples of confusion in Hollywood

- Detected correctly as HandShake
- Detected correctly as StandUp
- Detected wrongly as StandUp (true= HandShake)

- StandUp could be the source of most confusion in Hollywood mainly because almost all other actions involve some component of standing up?

Experiment 2: Back to 2d descriptors…

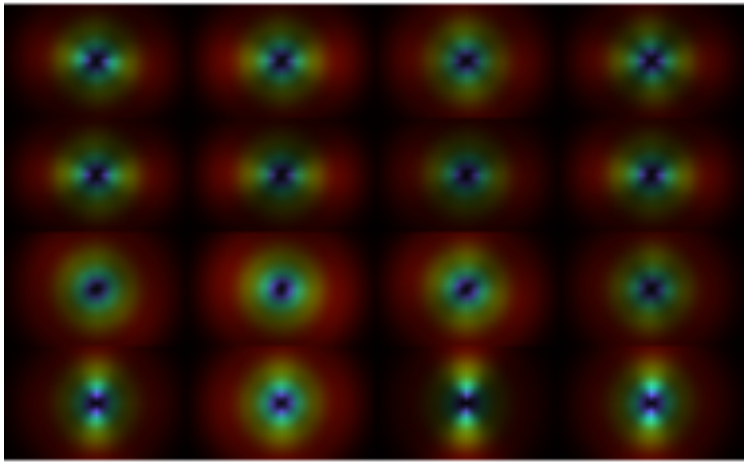
- **Goal:** To see whether a global image descriptor like GIST might help in activity classification.
- But GIST works for an image while STIP works for videos.

Video to image

1. Converted each video to frames
2. Clustered these frames [the grayscale values] into 10 clusters
3. Obtained a "**representative frame**" by considering a frame nearest to the center of the largest cluster
4. Calculated the 512 dim GIST vector on this frame

- **Classification is done by training a kchi2 kernel SVM** on the full 512 dim GIST vector (1 per video)

GIST Visualization



Examples of representative frames selected for the GetOutCar category

Examples of representative frame selected for the Running category






GIST based classification results

Data	GIST
KTH	37.96%
Hollywood	**33.33%**

v/s

Data	HOG	HOF
KTH	69.44	**81.94**
Hollywood	**44.44**	30.56

Discussion…

- The bad results of GIST based classification for KTH could be because of homogenous background in KTH clips
- GIST performed above chance (25%) for Hollywood
- Let's look at the confusion matrix for GIST based classification for Hollywood…**which category do you expect it to have helped the most?**

Hollywood Confusion Matrix [GIST]

%	GetOutCar	HandShake	Kiss	Stand-Up
GetOutCar	**89**	0	11	0
HandShake	67	22	11	0
Kiss	56	22	22	0
Stand-Up	56	22	22	0

33.33%

Discussion…

- GIST performed best for the GetOutCar category.
- Presence of car -> global information?
- GIST considers spatial structure of the image, so could perform better than a bag of words/ features approach
- Problem lies perhaps in identifying a representative key-frame?
- Scope for considering multiple frames per video & calculating GIST on them…

And using SIFT instead of GIST…

- Similar to obtaining GIST on the representative frame, I calculated SIFT keys on the same frame per video
- Since there are (multiple + variable number of) SIFT keys per image, I clustered the SIFT keys with k = 200 clusters
- Classified using a kchi2 kernel based SVM classifier on the obtained histograms

SIFT based classification results

Data	SIFT
KTH	38.43%
Hollywood	**25.00%**

v/s

Data	HOG	HOF	GIST
KTH	69.44	**81.94**	37.96%
Hollywood	**44.44**	30.56	33.33%

STIP(HoG+HoF)+GIST+SIFT

- I also ran the same classification experiments with different combinations [via concatenation] of STIP based HoG, HoF, HoG +HoF, GIST & SIFT as my feature vectors to the SVM

- There was no improvement over the baseline STIP (HoG, HoF) performance

Discussion…

- SIFT performs slightly worse than GIST for Hollywood
- SIFT is more local while GIST is more global
- Hence more influenced by the choice of the representative frame for each video?

Experiment 3: CLUSTERING STIPs

- Apart from returning the HOG & HOG descriptors calculated at the STIP, Laptev's STIP code also returns the following:
- **x,y** co-ordinates of the STIP points
- **time** of the STIP extraction
- XY-scale or **sigma**
- T-scale or **tau**
- **Detector-confidence**

I was curious whether there was any relation between these X, Y and T values and the action classes

3 types of information

- I considered this STIP "metadata" in 3 ways:

1. X & Y co-ordinates [**xy**]
2. X & Y & Time [**xyt**]
3. X & Y & Time & Sigma & Tau & Confidence [**xyt3**]

Training a classifier based on STIP metadata

- I clustered the values from the xy | xyt | xyt3 into 100 clusters.
- Quantized into a histogram for each video by considering the assignments of each STIP detection in the video to these 100 clusters
- Trained a SVM using the chi2-kernel on these histograms and used it to predict the per class accuracy

STIP metadata based Classification

Dataset (classes*tests per class)	XY	XYT	XYT3
KTH (6*36=216)	76.39% (165)	79.17% (171)	**80.56% (174)**
Hollywood (4*9=36)	36.11% (13)	41.67% (15)	**47.22%** (17)

Versus earlier classification based on HOG/HOF extracted at those STIPs

Data	HOG	HOF
KTH	69.44	**81.94**
Hollywood	**44.44**	30.56

Discussion: Importance of STIP metadata

- Classification based just on X & Y values of the STIPs gives pretty decent results
- Adding the temporal information (T) and the other 3 dimensions, improves accuracy further
- Seems to indicate that the STIP detection does correspond to highly relevant motion/activity detection in the videos w.r.t space and time.

Discussion: Possible implementation biases…

- Could be the result of using small dataset for Hollywood? (9 training+9 testing per class)?
- I did not normalize the X,Y & T co-ordinates for the frame dimensions & clip length. But they do not differ significantly intra-dataset, so that doesn't seem a likely source of bias.
- Normalization should actually improve results?
- This would have contributed in case the KTH & Hollywood datasets were mixed.

Reason for metadata performance …

- Could be **dataset dependent**?
- The Hollywood clips tend to **have slightly longer clips for activities like kissing and shorter clips for faster activities like standing up**. Hence the time of the STIP detections could help here.

Implementation hurdles…

- Source code for the STIP calculation is not available.
- Laptev has shared **pre-compiled binaries** for 64 bit Unix on his website which have some weird OpenCV & ffmpeg dependencies
- I settled for an old windows STIP .exe to get the STIPs which was terribly slow & crashed for most videos
- If anyone has gotten the STIP 2.0 code from his site working on Mac or on 32bit Linux, please do let me know since I need the STIP code for my project too. 😐

References

- Source of KTH action dataset:

http://www.nada.kth.se/cvap/actions

- Source of STIP code: (used stip-1.1-winlinux.zip)

http://www.di.ens.fr/~laptev/download.html#stip

Questions? && Thanks!