Fast Realistic Multi-Action Recognition using Mined Dense Spatio-temporal Features

by Andrew Gilbert, John Illingworth and Richard Bowden

Email: {a.gilbert, j.Illingworth, r.bowden}@surrey.ac.uk www.ee.surrey.ac.uk/personal/a.gilbert

Aim

To recognise and localise human actions in video sequences without pre localisation and in real time

Problems

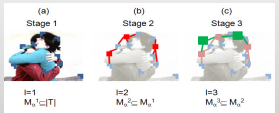
Low inter action class variation, High intra action class variation, which increases in real datasets such as Hollywood

Many approaches use sparse hand crafted spatial-temporal interest points, resulting in potential information being lost to recognition architecture

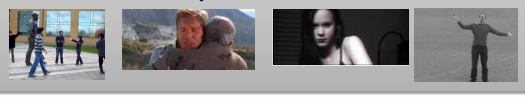
Approach

We employ an efficient data mining approach to build discriminative features using data mining that out performs competing techniques

- Hierarchically group 2D Harris corners in localised neighbourhoods
- Identify distinctive feature sub groups by data mining to form compound features
- Hierarchically group these compound features to build increasing descriptive feature



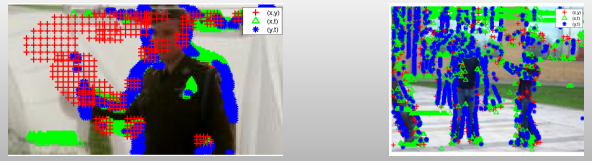
- Classification performed by matching compound features
- Tested on 3 datasets; Hollywood, Multi-KTH and KTH



Features

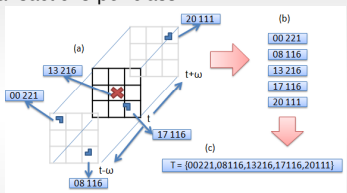
2D Harris Corner Detector applied in (x,y) (x,t) (y,t)

- Over complete set of features (1500 per frame)



Hierarchical Grouping

- Uses localised cuboid neighbourhoods to group features into Transactions vectors, T (symbolic)
- Millions of Transactions per class

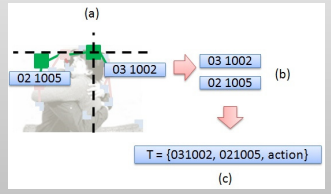


Apriori data mining identifies distinctive, descriptive feature compounds

- Efficient, 1 hr for 20GB of data,

Repeat at successive levels using feature compounds

- Group compound features in large neighbourhood



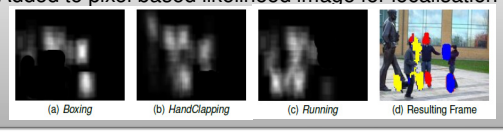
Final grouping uses relative positions in neighbourhood

- Provide scale invariance
- Final stage of mining provides action class model of compound features

Global Classifier

Frame-by Frame Voting scheme

- Detected Features matched to Learnt feature Lookup table
- Each match gives confidence of action likelihood
  - Summed for classification of sequence
  - Added to pixel based likelihood image for localisation



KTH Results

Schuldt Training/Test divisions 24fps

	box	clap	wave	jog	Run	Walk
Box	100	0	0	0	0	0
Clap	0	94	6	0	0	0
Wave	0	1	99	0	0	0
Jog	0	0	0	91	7	2
Run	0	0	0	10	89	1
Walk	0	0	0	0	6	94

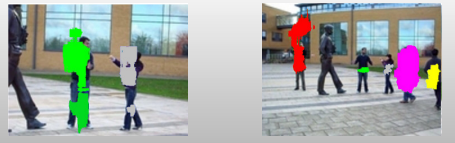
Localisation examples

Using leave-one-out test/train average precision 95%

Multi-KTH Results

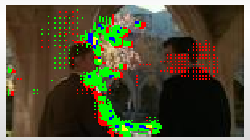
•4fps

	Clap	Wave	Box	Jog	Walk	Ave
Uemura	76%	81%	58%	51%	61%	65.4%
Mined	75%	84%	76%	50%	59%	68.8%



Hollywood Results

The compound features



Stage	Colour	No Feats
1	Red	1214
2	Green	1144
3	Blue	92
4	Yellow	61

Results - 10fps

Action	Stage 1	Stage 2	Stage 3
AnswerPhone	3.1%	25.7%	41%
GetOutCar	4.5%	38.5%	-
HandShake	2.3%	45.6%	52%
HugPerson	8.6%	42.8%	-
Kiss	43.3%	72.5%	-
SitDown	28.6%	84.6%	9%
SitUp	10.2%	29.4%	31%
StandUp	5.5	41.6	24%
Average	13.2%	53.5%	31.4%

Conclusions

Impressive state-of-the-art results, no pre localisation required

Data driven rather than hand engineered

Real-time operation, efficient training using data mining