The BEHAVE video dataset: ground truthed video for multi-person behavior classification

S. Blunsden, R. B. Fisher
Institute of Perception Action and Behaviour,
School of Informatics, University of Edinburgh
scott.blunsden@jrc.it

May 26, 2009

Abstract

Although there is much research on behaviour recognition in time-varying video, there are few ground truthed datasets for assessing multi-person behavioral interactions. This short paper presents the BEHAVE project's dataset, which has around 90,000 frames of humans identified by bounding boxes, with interacting groups classified into one of 6 different behaviors. An example of its use is also presented.

1 Introduction

In the past 10 years, there has been an explosion of research into the analysis of video data, particularly aimed at the detection of 'abnormal' human behavior (the definition of abnormal is usually defined on a paper by paper basis). The state of the art in this research has reached the point where human targets can generally be reliably detected and tracked in all but extreme conditions (poor lighting, severe and sustained occlusion). With that success, research has been concentrating on analysis of individual behaviors [5].

What has not received as much research effort so far is recognising the behavior of groups of people. Some notable examples are European handball play classification [2], American football play classification [12], basketball play classification [17] and in a more general surveillance context by Hakeen and Shah [10].

The key to making progress in a problem are potential algorithms and publically available benchmark datasets for researchers to compare algorithms. There are several potentially useful algorithmic frameworks for group behavior classification, *e.g.* Hidden Markov Models, Coupled Hidden Markov Models [16] and Conditional Random Field models [3, 4]. In the case of video sequence analysis, ground-truthed video sequences are essential. Unfortunately, they are also very time-consuming to produce if they are annotated to a reasonable level of detail. In the experience of our group, one hour of video (with about 90,000 frames), takes about 6 person-months of time for annotation at the level of individual bounding boxes and frame-by-frame behavior. Hence, such datasets are not commonly available.

This short paper presents the BEHAVE project's dataset (Section 3), which has about 90,000 frames, with humans identified by bounding boxes, and interacting groups classified into one of 6 different behaviors. An an example of its use is given in Section 5.

We are not aware of any published users of the BEHAVE dataset other than [3, 4], but the entry URL for the dataset [1] has had 5509 page accesses since October 2007, so we expect that there will be additional publications soon .

2 Related datasets

There are a number of video datasets with some form of ground truth. Most datasets are focused on target detection and tracking, or individual behavior. We review these first. Then we discuss briefly several major datasets suitable for group behavior research. Additional datasets can be found at the Cantata Video and Image Datasets Index at
`http://www.multitel.be/cantata/`.

2.1 Individual Behavior

1. **CLEAR**: The CLEAR: Classification of Events, Activities and Relationships [6] workshops produced annotated ground truth data for target detection and tracking, with a small amount of acoustic event data.

2. **i-LIDS**: Imagery Library for Intelligent Detection Systems. This dataset [11] has several hours of data about people and vehicles, including difficult lighting situations, but the ground truth is at the level of the whole clip (*e.g.* a person is entering a doorway during these frames). The focus is on security surveillance, *e.g.* sterile zones, abandoned items, etc).

3. **KTH Action Database**: The KTH "Recognition of human actions" database [13] is for recognition of instantaneous human activity, including walking, jogging, running, boxing, hand waving and hand clapping.

4. **PETS**: There have been many test challenge datasets for the PETS (Performance Evaluation of Tracking and Surveillance) series of workshops, which are indexed at: `http://www.cvg.rdg.ac.uk/slides/pets.html`. These are primarily videos of people and vehicles, with most ground truth concerning target position and instantaneous behavior.

5. **SCEPTRE**: The SCEPTRE [19] database (Service to Evaluate the Performance of Tracking and Recognition of Events) has about 5 minutes of European football (soccer) data, with hidden annotations for the players which are used for algorithm evaluation. It is unclear if game play is included in the hidden ground truth well as player position and identification.

6. **USF Sports**: The University of South Florida - Sports Action Dataset [22] contains about 10,000 frames of short clips of different sports activities, such as golf, gymnastics, skateboarding, football/soccer, horse riding, judo, etc, with target bounding boxes.

7. **ViHASi**: The ViHASi: Virtual Human Action Silhouette Data database [24] has multiple viewpoint video data of silhouettes of synthetic humans undertaking a variety of instantaneous activities. Twenty actions are recorded such as hanging onto a bar, jumping over object, jump-kick, etc.

2.2 Group Behavior

1. **CAVIAR**: The CAVIAR [9, 15] video dataset has about 100,000 frames of data, of which about 5000 frames involve some form of group activity. There were 27 separate group activity instances, such as joining, separating, walking together or fighting.

2. **CVBASE**: The CVBASE 2006 [7] sports video downloads (covering basketball, team handball, squash) have about 30 minutes total of video with annotation of player position and current group play.

3. **ETISEO**: The ETISEO database [8] contains 85 videos sequences ground truthed with the Viper-GT [23] tool, primarily recording target position, but also some annotation of individual instantaneous activity (*e.g.* walking), some activity of an individual in relation to a group (e.g. tailgating) or as groups (*e.g.* enters a special zone).

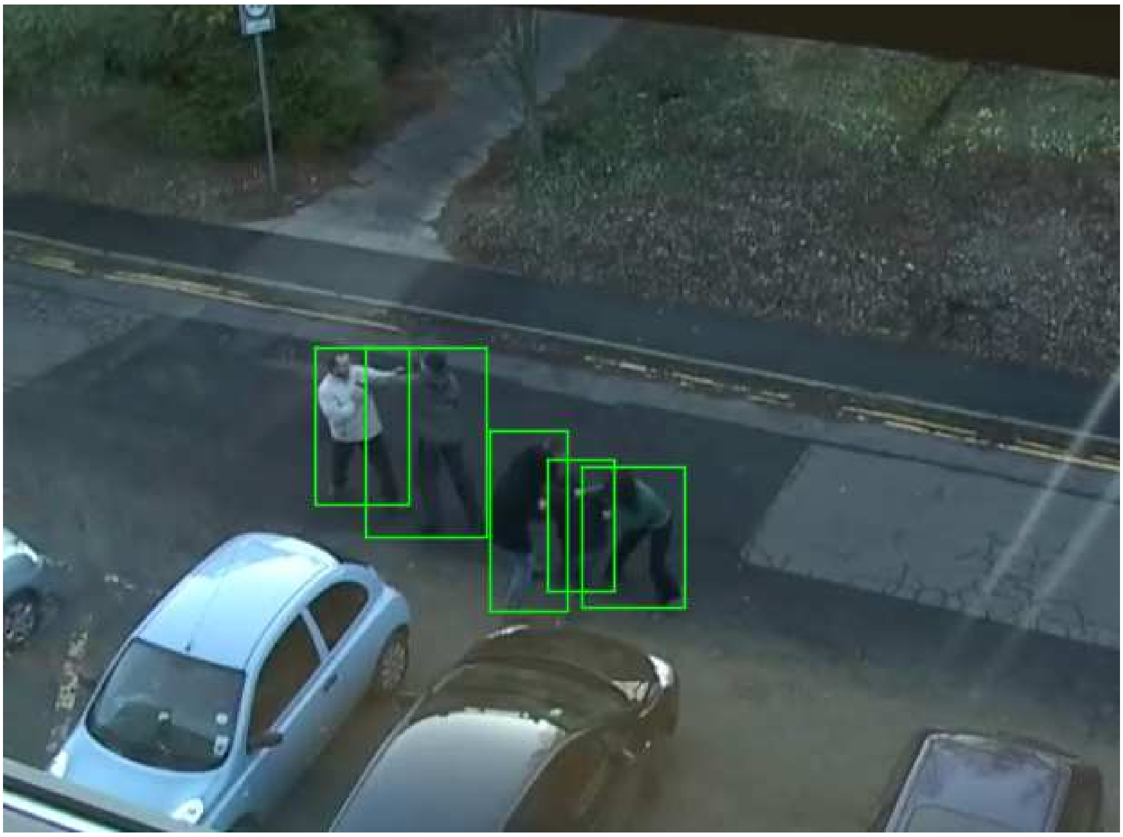

Figure 1: Example of video frame with marked bounding boxes.

3 Details of the Dataset

The BEHAVE video dataset consists of 4 video clips, downloadable as either 4 WMV videos (approximately 300 Mb in total) or as 76800 individual frames (approximately 10 GB in 8 files). The video and images were recorded at 25 frames per second using a commercial tripod-mounted camcorder. The resolution is 640x480.

Each interacting person has a bounding box (several non-interacting people who passed through the recording area were not marked up). Altogether, 125 instances of people were marked up for a total of 83545 bounding boxes.

The BEHAVE ground truth was constructed using the Viper-GT [23] ground-truthing tool, which encodes target positions in an XML variant. A sample frame with overlaid target bounding boxes is shown in Figure 1. The tracking information is only available for one of the two views. The single jpeg images of the video are also only given for the view where tracking information is available.

The position ground truth is supplemented by the group behavioural description, *e.g.*:

```
ID1     ID2     Start   End     Label
[2]     [0,1]   ;60296  ;60349  ;Approach
```

which says that group ID1 with person 2 is ‘Approach’ed by Group ID2 with persons 0 & 1 during frames 60296 to 60349.

Supplementing the tracking and behavior data is a set of measured scene points that allow generation of a ground plane homography.

The interactions consist of 2 to 5 people interacting as a group, or as two groups interacting. There are 10 types of group behavior that were annotated, given in Table 1 with (number of

Behavior	Brief	Instances	Frames
Type	Description		
InGroup	The people are in a group and not moving very much	35	14683
Approach	Two people or groups with one (or both) approaching the other	25	2272
WalkTogether	People walking together	43	6694
Meet	Two or more people meeting one another	1	27
Split	Two or more people splitting from one another	23	2529
Ignore	Ignoring of one another	2	597
Chase	One group chasing another	10	216
Fight	Two or more groups fighting	19	1751
RunTogether	The group is running together	4	335
Following	Being followed	1	92
Total		163	29196

Table 1: Number of interactions by type

instances, number of frames).

4 Example of an interaction

Figure 2 shows the evolution of part of a walking together sequence. Within the supplemental group behavioural description file the action would be represented as:

```
ID1    ID2     Start    End      Label
[0,1]          ;50359   ;50643   ;WalkTogether
```

This shows the ids of the persons which are given in the xml file. The xml file contains the position of the persons bounding box (as illustrated). In this sequence the two people are labelled as walking together.

The sequence shown in figure 3 corresponds to a more complex example. Here there are two separate fighting interactions occurring. Here it is shown that the two persons on the right (purple and blue boxes) are fighting and separately there is a fighting interaction occurring between the three people on the left of the screen. Within the file this information is represented as:

```
ID1 ID2        Start    End     Label
[0] [4]        ;60423   ;60635   ;Fight
[1] [2,5]      ;60423   ;60683   ;Fight
```

5 Example of use

This section presents a brief example of how one can perform classification upon the dataset. Here we present results of using a hidden Markov model (HMM) to classify the data. First the features which are used for classification are described.

6 Features

6.1 Movement Based Features

Movement plays an important role in recognising interactions. The speed of an individual is calculated as shown in equation (1). The double vertical bar ($\|.\|$) represents a vector $L2$ norm as given by $\|\mathbf{x}\| = \sqrt{\mathbf{x}_1^2 + \mathbf{x}_2^2 +, \dots, +\mathbf{x}_n^2}$, where $\mathbf{x}_n$ refers to the n^{th} component of the vector $\mathbf{x}$.

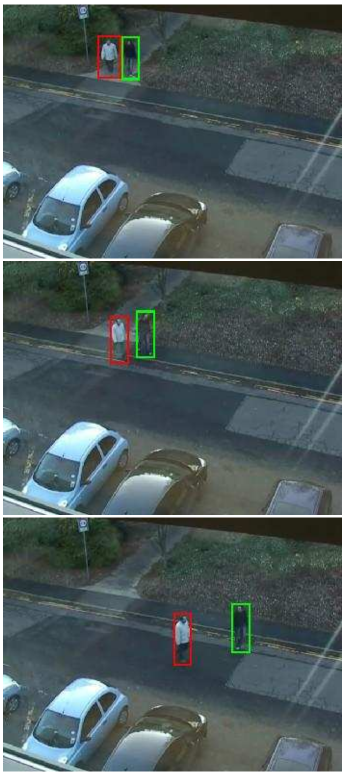

Figure 2: Example of a walking sequence. The two people are walking through the scene together.

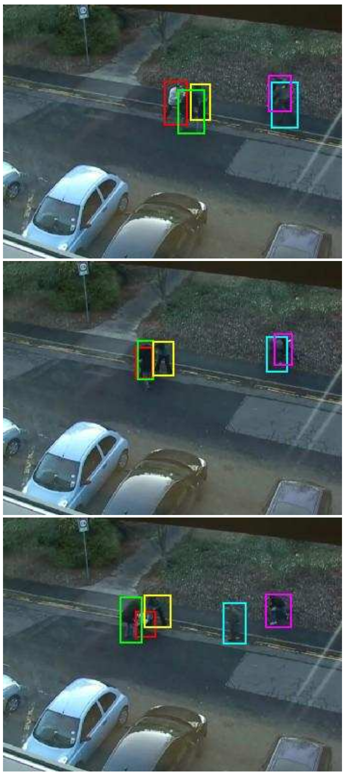

Figure 3: Example of a fighting sequence.

$$s_i^t = \frac{1}{w}\|\mathbf{p}_i^t - \mathbf{p}_i^{t-w}\| \tag{1}$$

Here $\mathbf{p}_i^t$ refers to the position of the tracked object at time t for object i. Within this work only the two dimensional ($\mathbf{p}_i^t = [x_i^t, y_i^t]$) case is considered due to tracking information being two dimensional. The w temporal offset is introduced due to the high frame rates which typify many modern video cameras. High frame rates of around 25fps can mean that taking the last frame (w=1) results in very small movements between subsequent frames which can be mostly noise.

The absolute difference in speed ($\epsilon_{[i,j]}^t$) between two tracks is also calculated ($\left|s_i^t - s_j^t\right|$). The vorticity ($\nu_t^i$) is measured as a deviation from a line. The line is calculated by fitting a line to a set of previous positions of the trajectory $\mathbf{P}_i^t = [\mathbf{p}_i^{t-w}, .., \mathbf{p}_i^t]$. At each point the orthogonal distance to the line is found. The total distance of all points are then summed and normalised by window length to give a measure of the vorticity.

6.2 Alignment Based Features

The alignment of two tracks can give valuable information as to how they are interacting. The degree of alignment is common to [21] and [16] who all make use of such information when classifying trajectory information.

To calculate the dot product the heading ($\mathbf{h}$) of the object is taken as in equation (2) and the dot product was calculated from the directions of tracks i and j.

$$\hat{\mathbf{h}}_i^t = \frac{\mathbf{p}_i^t - \mathbf{p}_i^{t-w}}{\|\mathbf{p}_i^t - \mathbf{p}_i^{t-w}\|} \tag{2}$$

$$a_{[i,j]}^t = \hat{\mathbf{h}}_i^t \cdot \hat{\mathbf{h}}_j^t \tag{3}$$

In addition to alignment the potential intersection ($\gamma_t^{i,j}$) of two trajectories is also calculated. To calculate this a simple line intersection is performed (the lines are determined by fitting a line to previous points). We then check that the people are both heading towards the point of intersection (as the lines are undirected).

6.3 Distance Based Features

Distance is a good measure for many types of interaction. For example, meeting is not possible without being in close physical proximity. First a Euclidean distance measure is used as given in equation 4.

$$d_{[i,j]}^t = \|\mathbf{p}_i^t - \mathbf{p}_j^t\| \tag{4}$$

The derivative of the distance was also calculated. This is the difference in distance at contiguous time steps. It is calculated as shown in equation 5 below.

$$\dot{d}_{[i,j]}^t = d_{[i,j]}^t - d_{[i,j]}^{t-1} \tag{5}$$

An instantaneous measure such as the distance and the derivative of the distance can both be prone to short term tracking errors. In an effort to remove this effect a window size containing w points (as in $\mathbf{P}_i^t$ in section 6.1) was averaged. The distance was calculated for every point (as in equation 4) in this window.

$$\hat{d}_{[i,j]}^t = \frac{1}{w}\sum_{k=t-w}^{t} d_{[i,j]}^k \tag{6}$$

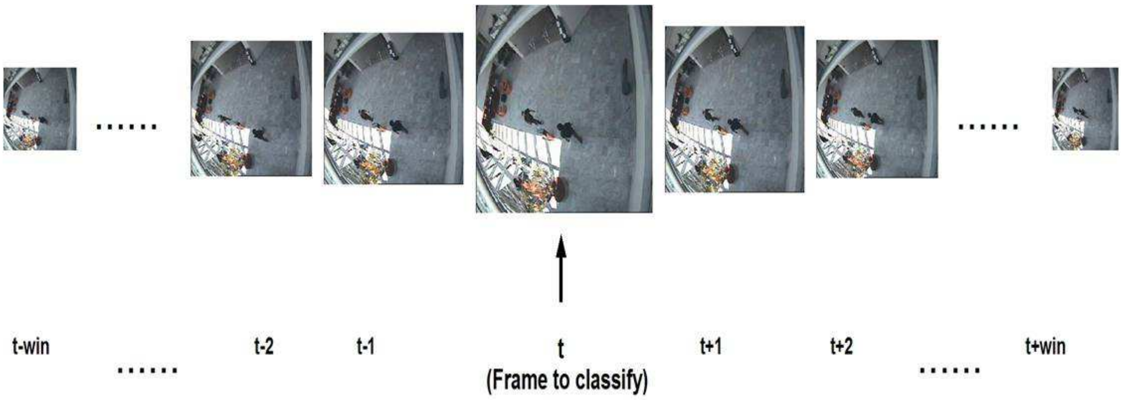

Figure 4: The frame to classify (t) uses information from ±w frames around the current frame in order to classify the frame.

6.4 Final Feature Vector

The final feature vector for each pair of people is given in equation (7).

$$\mathbf{r}_t^{i,j} = \left[s_i^t, s_j^t, \dot{s}_t^i, \dot{s}_t^j, \epsilon_{[i,j]}^t, a_{[i,j]}^t, d_{[i,j]}^t, \dot{d}_{[i,j]}^t, \nu_t^i, \nu_t^j, \gamma_t^{i,j}\right] \tag{7}$$

The vector between persons i and j at time t is made up of the speed of each person (s_i^t, s_j^t) along with the change in speed $\dot{s}_t^i, \dot{s}_t^j$. The alignment, distance and change in distance at a particular point in time is given by $a_{[i,j]}^t$, $d_{[i,j]}^t$and $\dot{d}_{[i,j]}^t$ respectively. The vorticity of a trajectory is given by ν_t^i. The possible intersection of two trajectories is represented by $\gamma_t^{i,j}$. The final vector contains 11 features. The data was normalised to have zero mean and unit standard deviation.

6.5 Observation Window Size

Throughout these experiments we investigated the role of varying the number of video frames used before making a decision as to what is happening within the frame. Figure 4 below shows how this is achieved. We used information from before and after the current frame in order to classify it. This helps with the lag problem where too much of the current decision is based upon previous frames. The window size variation is equivalent to a few seconds delay. This is not foreseen as a problem if such an approach was taken in a real surveillance application. The fact that there would be a slight lag in classification if making use of only previous information seems an appropriate trade-off for an increase in accuracy.

7 Classification

Here we demonstrate results when using a hidden Markov model (HMM). HMM's have been introduced by (among others) Rabiner [18]. The model is parameterised by a prior distribution Π with each element π_i representing $\pi_i = p(x = i)$ across all hidden states $i \in [1, .., N]$. The stationary state transition matrix $\mathbf{A}$ is referenced by $a_{i,j} = p(x_t = i | x_{t-1} = j)$. Within this work we are concerned with continuous real valued observations $(\mathbf{r}_t)$ which can be accommodated within the model by using a Gaussian mixture model to represent the observation probability distribution $p(\mathbf{r}_t | x_t = j)$.

$$b_j(\mathbf{r}_t) = \sum_{m=1}^{M} c_{j,m} N(\mathbf{r}_t, \mu_{j,m}, \mathbf{C}_{j,m}) \tag{8}$$

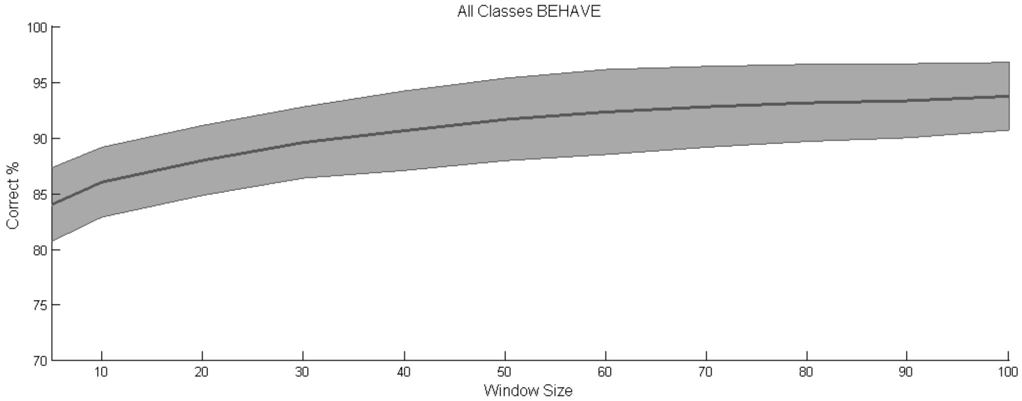

Figure 5: Overall performance on the BEHAVE dataset when using a HMM classifier. Lines show averaged results (over 50 runs) whilst the shaded regions show one standard deviation.

Here the observed data is given by $\mathbf{R}$, $c_{j,m}$ is the mixture coefficient for the m^{th} mixture in state j. N is the Gaussian distribution with mean vector $\mu_{j,m}$ and covariance $\mathbf{C}_{j,m}$ for the m^{th} mixture in state j. The mixture coefficients c_j must sum to 1. The HMM's parameters can thus be represented as $\lambda = (\Pi, \mathbf{A}, \Theta)$ where θ represents the parameters of the mixture model.

8 Results

Here the results of applying the HMM classifier to the dataset are presented. We first split the training and testing data 50/50. We classify five types of interaction provided by the datasets. The five interactions we classify are 'in group','walk together', 'fight', 'split' and 'approach'. Each are well represented in the dataset (see figure 1). Each class has its own HMM which is trained upon that class's training set. Each frame of the training set creates a vector (as given in equation 7) and the complete sequence is used to train the parameters of the HMM using expectation maximisation.

For classification a window around the current frame is used with each frame being represented by the calculated feature vector. This window is then presented to each HMM and a likelihood is produced. We classify the segment as having the same class as the HMM model with the largest likelihood. The overall classification results are presented in figure 5 and table 2.

Window Size	Performance
5	82.75 ± 3.39
10	85.98 ± 3.12
20	87.93 ± 3.13
30	89.54 ± 3.19
40	90.59 ± 3.56
50	91.6 ± 3.69
60	92.26 ± 3.80
70	92.73 ± 3.61
80	93.08 ± 3.45
90	93.27 ± 3.31
100	93.67 ± 3.02

Table 2: Average performance and variance

9 Conclusions

This paper has presented a new dataset [1] providing ground truth tracking information along with descriptions of behaviors for interacting groups. The contents and format of the dataset have been described. An example of how the dataset can be used has been presented. It is our hope that making such data publically available will stimulate other work involving multiple interactions and provide a common benchmark dataset.

References

[1] University of Edinburgh, The BEHAVE Dataset, `http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/`. Accessed May 11, 2009.

[2] S. Blunsden, R. Fisher. “Recognition of coordinated multi agent activities, the individual vs the group”, Proc. Workshop on Computer Vision Based Analysis in Sport Environments (CVBASE), pp 61-70, 2006.

[3] S. Blunsden, E. Andrade, R. Fisher. “Non Parametric Classification of Human Interaction”, Proc. 3rd Iberian Conference on Pattern Recognition and Image Analysis, Girona, pp 347-354, June 2007.

[4] S. Blunsden, “Recognition and Classification of Multi-Person Interaction”, PhD Thesis, University of Edinburgh, 2008.

[5] H. Buxton, “Learning and understanding dynamic scene activity: a review”, Image and Vision Computing, Volume 21, Issue 1, pp 125-136, 10 January 2003.

[6] CHIL and NIST, CLEAR: Classification of Events, Activities and Relationships, `http://www.clear-evaluation.org/`. Accessed November 12, 2008.

[7] Anonymous, The CVBASE 2006 Test Dataset, `http://vision.fe.uni-lj.si/cvbase06/downloads.html`. Accessed November 12, 2008.

[8] Nghiem A.-T., F. Bremond, M. Thonnat and V. Valentin. “ETISEO, performance evaluation for video surveillance systems”. Proceedings of AVSS 2007, September, 2007, London, UK. See `http://www-sop.inria.fr/orion/ETISEO/download.htm` for the data. Accessed November 12, 2008.

[9] R. B. Fisher, “The PETS04 Surveillance Ground-Truth Data Sets”, Proc. Sixth IEEE Int. Work. on Performance Evaluation of Tracking and Surveillance (PETS04), pp 1–5, Prague, May 2004. (Unrefereed).

[10] Hakeem, A. and Shah, M. “Learning, detection and representation of multi-agent events in videos” Artificial Intelligence, Vol 171, pp 586-605, 2007.

[11] UK Home Office, CCTV and imaging technology, `http://scienceandresearch.homeoffice.gov.uk/hosdb/...` `cctv-imaging-technology/video-based-detection-systems/?version=1`, Accessed November 12, 2008.

[12] S. S. Intille and A. F. Bobick, “A framework for recognizing multi-agent action from visual evidence”, Proc. Sixteenth National Conference on Artificial Intelligence, Menlo Park, CA: AAAI Press pp. 518-525., 1999

[13] Christian Schuldt, Ivan Laptev and Barbara Caputo, ”Recognizing Human Actions: A Local SVM Approach”, Proc. ICPR’04, Cambridge, UK. The data can be acquired at: `http://www.nada.kth.se/cvap/actions/`. Accessed November 12, 2008.

[14] Welling, M. “Fisher Linear Discriminant Analysis”, Technical report, University of Toronto, Kings College Road, Toronto, M5S 3G5, Canada, September, 2000.

[15] T. List, J. Bins, J. Vazquez, R. B. Fisher, “Performance Evaluating the Evaluator”, Proc. 2nd Joint IEEE Int. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, (VS-PETS), pp 129-136, Beijing, Oct 2005.

[16] Oliver, N. M. and Rosario, B. and Pentland, A. P. “A Bayesian Computer Vision System for Modelling Human interactions”, IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 22, number 8, pp 831-843, August 2000.

[17] Matej Perse, Matej Kristan, Stanislav Kovacic, Goran Vuckovic, Janez Pers, “A trajectory-based analysis of coordinated team activity in a basketball game”, Computer Vision and Image Understanding, In Press, Corrected Proof, Available online 28 March 2008.

[18] Rabiner, L. R. “A tutorial on hidden Markov models and selected applications in speech recognition”, Readings in speech recognition, Morgan Kaufmann Publishers Inc., pp 267-296, 1990.

[19] Kingston University, SCEPTRE database (Service to Evaluate the Performance of Tracking and Recognition of Events), `http://sceptre.king.ac.uk/sceptre/default.html`. Accessed November 12, 2008.

[20] Yilmaz, A., Javid, O. and Shah, M., “Object Tracking: A Survey”, ACM Computing Surveys, volume 38, pp 13-58, 2006

[21] Gigerenzer, G., Todd, P. M. and ABC Research Group, “Simple Heuristics That Make Us Smart”, Oxford University Press, Evolution and Cognition Series, 1999

[22] University of South Florida, “Sports Action Dataset”. `http://server.cs.ucf.edu/~vision/projects/...`
`action_mach/ucf_sports_actions.rar`. Accessed November 12, 2008.

[23] University of Maryland Language and Media Processing Lab, “ViPER: The Video Performance Evaluation Resource”, `http://viper-toolkit.sourceforge.net/`. Accessed November 12, 2008.

[24] H. Ragheb, S. Velastin, P. Remagnino and T. Ellis, “ViHASi: Virtual Human Action Silhouette Data for the Performance Evaluation of Silhouette-Based Action Recognition Methods”, Workshop on Activity Monitoring by Multi-Camera Surveillance Systems, Stanford, September 11, 2008. The data can be acquired at: `http://dipersec.king.ac.uk/VIHASI/`. Accessed November 12, 2008.

[25] Wallach, H. M., “Conditional Random Fields: An Introduction”, University of Pennsylvania, CIS Technical Report MS-CIS-04-21, 2004.