Detecting Social Groups in Crowded Surveillance Videos Using Visual Attention

Michael.J.V.Leach
Roke Manor Research
Romsey, Hampshire, United Kingdom
michael.leach@roke.co.uk

Rolf. Baxter
School of Engineering & Physical Sciences
Heriot-Watt University, Edinburgh Campus, Edinburgh, Scotland
R.H.Baxter@hw.ac.uk

Neil.M.Robertson
School of Engineering & Physical Sciences
Heriot-Watt University, Edinburgh Campus, Edinburgh, Scotland
n.m.robertson@hw.ac.uk

Ed.P.Sparks
Roke Manor Research
Romsey, Hampshire, United Kingdom
ed.sparks@roke.co.uk

***Abstract*—In this paper we demonstrate that the current state of the art social grouping methodology can be enhanced with the use of visual attention estimation. In a surveillance environment it is possible to extract the gazing direction of pedestrians, a feature which can be used to improve social grouping estimation. We implement a state of the art motion based social grouping technique to get a baseline success at social grouping, and implement the same grouping with the addition of the visual attention feature. By a comparison of the success at finding social groups for two techniques we evaluate the effectiveness of including the visual attention feature. We test both methods on two datasets containing busy surveillance scenes. We find that the inclusion of visual interest improves the motion social grouping capability. For the Oxford data, we see a 5.6% improvement in true positives and 28.5% reduction in false positives. We see up to a 50% reduction in false positives in other datasets. The strength of the visual feature is demonstrated by the association of social connections that are otherwise missed by the motion only social grouping technique.**

***Keywords*-Video surveillance; Computer aided analysis; Machine vision**

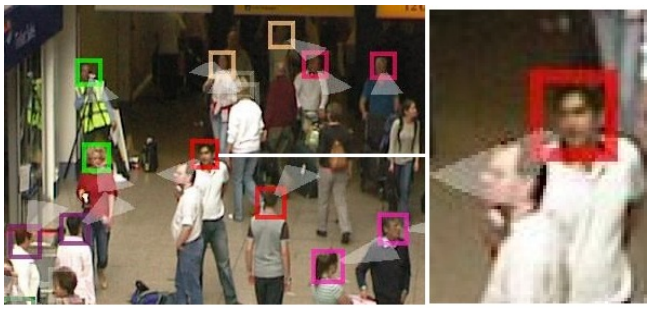
I. Introduction

Human behavior analysis has presented a challenging problem in autonomous surveillance due to the variety, subtlety, and obscurity of behavioral expression. To address the challenge, estimating social connectivity between individuals is a contextual feature gaining popularity in recent work. Social connectivity and grouping is used to improve tracking [1] and behavior analysis. It has been shown that with an understanding of the social context surrounding human behavior in surveillance it is possible to better interpret observed events and detect abnormal behavior [2], [3]. Using the social behavior feature is particularly relevant in crowded environments in which the motion of an individual is more constrained and social dependencies are more salient against the entropic crowd motion. Our work focuses on the use of visual attention to better classify social connections in a semi-crowded surveillance scene. Human motion information of individuals and crowds is commonly used in automatic social grouping. However, the surveillance environment can exert influence upon trajectories by channeling people, presenting queuing or waiting areas, or containing objects to interact with. These motions are ambiguous with intentional motion from social connections and as such obscure any trivial definition of social connectivity. In this work we extract a further feature; visual attention and demonstrate visual attention can be used to better identify social grouping in crowded environments. The visual attention of an individual provides an additive feature which supplements the motion based similarity used in the state of the art. The visual attention feature is not impacted in the same way as the motion features are by the scene.

With this research we aim to verify the hypothesis that pedestrian visual attention can be used to compliment motion based social group estimation. To verify our hypothesis we will implement our hybrid motion-visual attention system demonstrating better social grouping in a variety of different surveillance datasets. Comparison will be made against a hand labeled social group ground truth, assessing the efficacy of our visual attention and motion against motion alone.

A. Related Work

The estimation of social groups in surveillance has a focused primarily on motion features. To estimate social groupings Ge et al. uses a proximity and velocity metric to associate individuals into pairs, iteratively adding additional individuals to groups using the Hausdorff distance as a measure of closeness [5]. Yu et al. implements a graph cuts based system which uses the feature of proximity alone [6]. However, modeling social groups by positional information alone is prone to finding false social connections when individuals are within close proximity due to environmental influences such as queuing. Oliver et al. uses a Coupled HMM to construct a-priori models of group events such as Follow-reach-walk together, or Approach-meet-go separately [7]. Certain actions are declared group activities and



(a) (b)

Figure 1. Image (a) illustrates an example from the PETS 2007 dataset [4] of our tracking output, the social groups (designated by colored bounding boxes) and the extracted gaze direction estimates (illustrated by field of view cones). image (b) is a zoomed subsection showing the gaze direction field of view of a person in the image.

thus groups can be constructed from individuals via mutual engagement in a grouping action. However, a more recent development in automatic social grouping seeks to model social interaction using the visual interest of the tracked individuals. The use of an individuals visual attention is significant as it uses a rich feature which indicates the intention of the individual. Robertson and Reid utilize gaze direction in order to determine whether individuals are within each other's field of view [8]. Farenzena et al use an estimation of the visual focus of attention of a person as a cue to indicate social interaction [9]. Head pose is quantized into 4 different locations at each frame, and a predefined set of spatial and visual criteria determines if the conditions for a social interaction are met at each time step. A social exchange is then defined as lasting a given duration (10 seconds). In our work we bring together the motion based social paradigm with the benefit of visual information as it is demonstrated by [8], [9].

B. Initial Hypothesis Validation

Our visual interest social grouping is based on the hypothesis that socially connected people act as a source of visual interest for each other. This hypothesis makes the implicit assumption that the gazing patterns of socially connected persons differs from those that are unconnected. The validation of the underlying hypothesis was performed in two steps. In the first step, pedestrians were segmented into two groups: those with social connections, and those without. This segmentation was performed by hand. Once segmented, we calculated the deviation between travel direction and gaze direction for each pedestrian for each frame of video. Travel direction was calculated using each persons smoothed velocity over a 15 frame window using the head centroids provided by Benfold [10] and our own tracks on the PETS 2007 data [4]. In our initial validation we removed false-positive head detections and used hand-labeled gaze directions rather than utilizing algorithmic solutions. Formally, denote a persons velocity direction at frame t as θ_t^D and their gaze direction as θ_t^G. The gaze-velocity deviation can then be calculated as the absolute error $\epsilon_t = |\theta_t^D - \theta_t^G|$. The mean and variance of the deviations was then extracted for the two pedestrian groups (socially connected and unconnected) upon which further analysis was performed.

Validation Results: The analysis of gazing patterns was performed on 3 datasets: the Benfold dataset [10], the Caviar dataset [11] and the PETS 2007 dataset [4]. In each case the pedestrian detection and tracking information provided by each dataset was used. Where not supplied, additional ground-truth gaze labels were added by the authors of this paper. Statistics were extracted for 37 tracks from the caviar dataset, 372 tracks from the PETS dataset, and 170 tracks from the Benfold dataset. Figure reffig:DataExample shows example frames from PETS scene 4 highlighting socially connected and unconnected persons. We illustrate in figure reffig:GazeVelocity the extracted distributions from all datasets. One can see from the figure that for the Benfold dataset, there is little difference between the gazing patterns of the two groups. However, on the caviar dataset two distinct distributions are observed, as is also the case with the PETS dataset. For each dataset, performing the χ^2 variance test between the socially connected and unconnected deviations with a p-value of 0.05 shows that in all three datasets, the differences between the deviations for socially connected and unconnected persons are statistically significant.

To partially validate our assumption that socially connected individuals are a source of visual focus for each other, we analyzed the null hypothesis that socially connected and unconnected persons have the same gazing patterns. Our analysis of deviations between travel direction and gaze direction showed evidence that gazing patterns do differ between socially connected and unconnected persons. However, the degree of separation between distributions varied for each dataset, identifying the need for the weighting factor to be used when using gaze data for determining social connectivity. For all datasets the differences between the two groups were statistically significant giving support for our assumption and leading us to reject the null hypothesis.

II. Method

The motion based social grouping is grounded upon the premise that shared trajectory information implies a social dependence between two individuals. The principles of the social force model are such that socially connected individuals are more likely to move together, and thus display more similar trajectory information, and socially independent people feel a force of repulsion and are more likely to avoid moving similarly and avoid close proximity. The more entropic the underlying motion of the crowd the more salient similar social trajectories will be. We base our motion based social grouping method upon the work of Leach et al. A full description of the method is given in

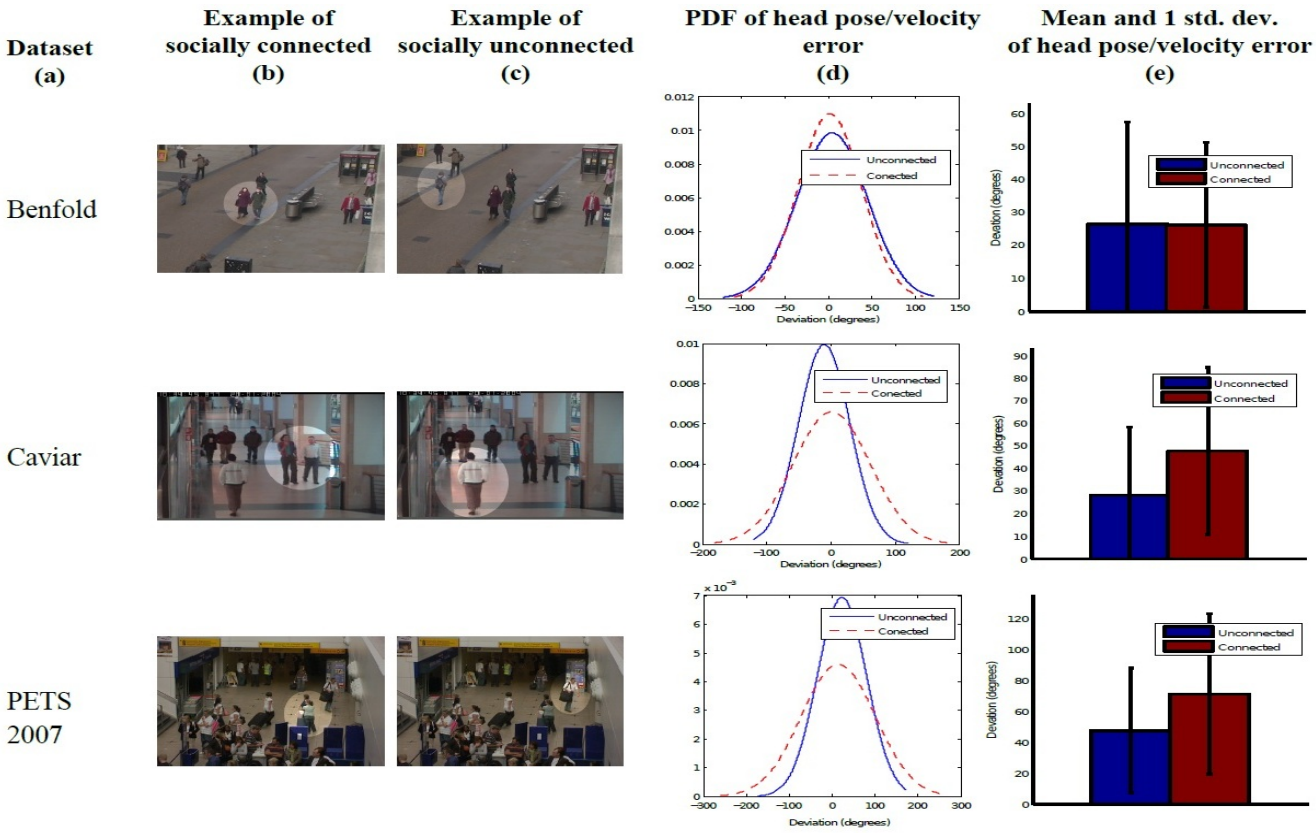


Figure 2. Example frames and extracted gaze-velocity deviation (error) statistics extracted from three datasets. Column (d) shows the normal distributions for socially connected (red) and unconnected (blue) persons. Column (e) shows the mean and standard deviation of socially connected (red) and unconnected (blue) persons.

the referenced paper [2].Their grouping method finds social similarity upon the features of direction, speed, proximity, and temporal overlap. Each feature is weighted based upon a one off training phase, such that proximity and temporal overlap have more dominance in the overall metric than direction and speed, which were found to be less important. The similarity of direction and speed are measured using the mutual information measure. The proximity and temporal overlap similarity are measured by euclidean distance. Once the similarity for each feature has been measured the four features are combined to a single similarity measure, which we use 4. Each tracked object has a similarity to every other, populating a social pairing likelihood table.

A. *Visual Interest enhanced social grouping*

To verify our hypothesis that visual interest can be used to enhance the existing motion based social grouping we incorporate gaze direction and subsequently visual interest into the social grouping model. The distinction between gaze direction and visual interest is as follows; gaze direction is the raw angle in which the person is looking, usually indicated by head pose in our data, and visual interest is an estimation of or distribution over possible regions of interest. In our case, we extract gaze direction estimates in order to further refine the estimation with knowledge of interest points and characteristics of how interest drops at the periphery or with distance, permitting a estimation of the focus of interest for any given person.

We use two methods to determine gaze direction; hand annotated ground truth and automated gaze estimation. We first hand annotate each head image at each frame in order to provide a baseline gazing direction from which the error of the automatic estimation can be calculated. Furthermore this provides the means to verify our hypothesis upon ideal data prior to testing on realistic data with error. Hand annotation was achieved by a tracking a mouse pointed moved by the user to point to the current angle that the head was posed at at 10 frames per second for a single person at a time. As the head angular velocity is highly constrained the head pose is particularly predictable in the short term, and hand tracking was found to be adequate, with occasional latency when the head motion is more erratic. In total we groundtruthed 3

datasets; 2 from the PETS 2007 data, and the Oxford Data. For the Oxford dataset approximately 70,000 head images where annotated. For the PETS scene 4 near 90,000 head images where annotated. The PETS scene 0 data entailed over 50,000 head images.

To automatically determine the gaze direction we must estimate their head pose at each frame, and determine the likelihood distribution of visual interest given typical scene interest and people. Our method is identical to the work 'Unsupervised Learning of a Scene-Specific Coarse Gaze Estimator' [10] with the exception of the image classification factor. The work of Benfold uses a randomized forest of ferns to learn typical relations between pixel triplets for a given head pose angle. The randomized trees were trained in a weakly supervised fashion with examples of each head pose class being fed through the ensemble of trees such that at each end node for each tree a distribution over every class is populated showing the probability that a head image reaching this node belongs to any given head pose class. The tree is split at each branch based upon one of two types of binary decision. The first decision compares randomly chosen bins from the Histogram of Oriented Gradients (HOG) representation of the head image. The second type of decision is a Color Triplet Comparison (CTC) which samples three pixels of the head image and makes a binary decision based upon the difference between pixel A and B and the difference between B and C using the L1 Norm of the RGB vector for any given pixel pair. A full comparison of all pixel triplets would be infeasible, entailing $N^3/2$ comparisons where N is the number of pixels in a head image, each comparison representing a split in a decision tree, rendering the end nodes severely under-populated. Typically only 100 pixel triplets are sampled, which is a heavy subsampling of the entire feature space. In Benfolds work this is a necessary step in order to achieve video rate processing however as we are not constrained by processing time we opt for a slower classification technique. Rather than using pixel triplets we use pixel pairs, making a binary decision based upon whether the magnitude of pixel A is greater than that of pixel B, for each color channel independently. We build a map of all possible unique pixel comparisons, which is $3/2 * N^2$ comparisons. For any given head image the feature representing the head image is a four dimensional map detailing the pixel comparison result, 0 or 1, between each pixel and every other pixel in each color channel. To build a classifier for each class we then take the mean pixel comparison result for all head images of this class. The result of which is a 4D matrix detailing the probability of any pixel A having a greater magnitude than pixel B for a given gaze direction, quantized to 8 directions. Classification is then achieved by extracting the binary features from a query image and multiplying through by the class 4D probability matrix, giving a score of how well the query image fits the head pose class. This provides a probability distribution over all gaze directions for the head in the query image. For a sequence of query images we then smooth over the distributions using forward backward smoothing as set up in the original Benfold work.

(a) (b)

Figure 3. Illustration of mutual visual attention (a) and visual correlation (b). In both cases the red line represents variable θ^G_{it}, the gaze direction for person i. The bright green line is the angle compared to in the two metrics. For image (a) this is θ_{ijt} the direction from person i to j for mutual attention, and for image (b) the green line represents θ^G_{jt} the gazing direction for person j for visual correlation.

We theorize that there are two ways socially connected individuals can demonstrate social interaction using gazing direction; correlated direction of gaze, and looking at each other. The former occurs in cases when two individuals are actively looking at the *same thing*, which requires communication to coordinate, however it could be coincidental when an event or object has drawn both of their attention. The latter event, when two individuals are looking at *each other*, implies communication - that they are the object of attention for each other. It is at least unusual for two socially unconnected individuals to look at each for a prolonged period of time. Following from this reasoning, there are two events we wish to measure. These are, how similar the gaze direction of two individuals are and the amount of time gaze direction is directed towards each other. Figure 3 illustrates the two examples of mutual visual attention (a) and visual correlation (b). In both cases the red line represents variable θ^G_{it}, the gaze direction for person i. The bright green line is the angle compared to in the two metrics. For image (a) this is θ_{ijt} the direction from person i to j for mutual attention, and for image (b) the green line represents θ^G_{jt} the gazing direction for person j for visual correlation. We wish to exclude cases where two individuals are looking in the same direction due to walking in that direction. The work of Benfold [10] showed that pedestrians spend the majority of their time looking in the direction of travel. To avoid highly scoring correlated gaze direction due to two people looking in the same direction of travel we introduce a weight which represents our confidence that the direction of gaze is due to visual interest other than direction of travel. The weighting is greater for those with a gaze direction off their current

direction of travel. The visual correlation weight coefficient is given by:

$$\omega_{ijt} = |\theta^G_{it} - \theta^D_{it}||\theta^G_{jt} - \theta^D_{jt}| \quad (1)$$

Where θ^G_{it} is the gazing direction for person i at frame t, and θ^D_{it} is the direction of travel. Thus we score people with attention towards either side stronger than those who are looking in the direction of travel, dropping linearly. This is justified on the assumption that a social source of attention is more likely when not looking in the direction of travel; backed up by the preliminary hypothesis verification. If there is no current direction of travel then this weight is always 1. Similarly, we introduce a weighting for visual mutual attention. We weight the measure of visual interest between two individuals by proximity. The further away someone is the less confident we are they are a social focus of attention. The mutual visual attention weight is given by:

$$\lambda_{ijt} = 1 - \frac{\sqrt{x^2_{ijt} + y^2_{ijt}}}{X} \quad (2)$$

Where x_{ijt} and y_{ijt} is the x and y distance between person i and person j at frame t and X is the width of the scene; the maximal distance between two people. Thus we model the probability of interest between person i and person j as falling linearly with distance. We then define the total Visual Interest feature Λ_{ijt} between person i and j at any given frame as the product of two Gaussian distributions encompassing the visual correlation variance σ_λ and the visual mutual attention variance σ_ω predefined as $\pi/4$.

$$\Lambda_{ijt} = \frac{1}{\sigma_\lambda \sigma_\omega 4\pi^2} e^{-\frac{|\theta_{ijt} - \theta^G_{it}|^2}{2\sigma^2_\lambda} - \frac{|\theta^G_{it} - \theta^G_{jt}|^2}{2\sigma^2_\lambda}} \quad (3)$$

Where θ_{ijt} is the direction from person i to person j at time t. We next incorporate the visual interest into our system as another feature in the existing social similarity metric. We measure the visual interest similarity between each potential socially connected individuals and include this with a weighting of 1 into the social similarity metric. Thus for any two people the features that determine grouping likelihood in the social pairing table are; proximity, temporal overlap, direction, speed, and visual interest. The total social grouping strength between person i and person j for all frames is then given by:

$$\kappa_{ijt} = \frac{1}{T} \sum_t^T IV_{ijt} I\Theta_{ijt} \Delta P_{ijt} \tau_{ijt} \Lambda_{ijt} \quad (4)$$

$\tau_{ijt}, IV_{ijt}, I\Theta_{ijt}, \Delta P_{ijt}, \lambda_{ij}$ are the temporal overlap, mutual information for speed, mutual information for direction, proximity and visual interest difference between person i and j. Specific definitions for the motion features $\tau_{ijt}, IV_{ijt}, I\Theta_{ijt}, \Delta P_{ijt}$ are given in Leach et al [2].

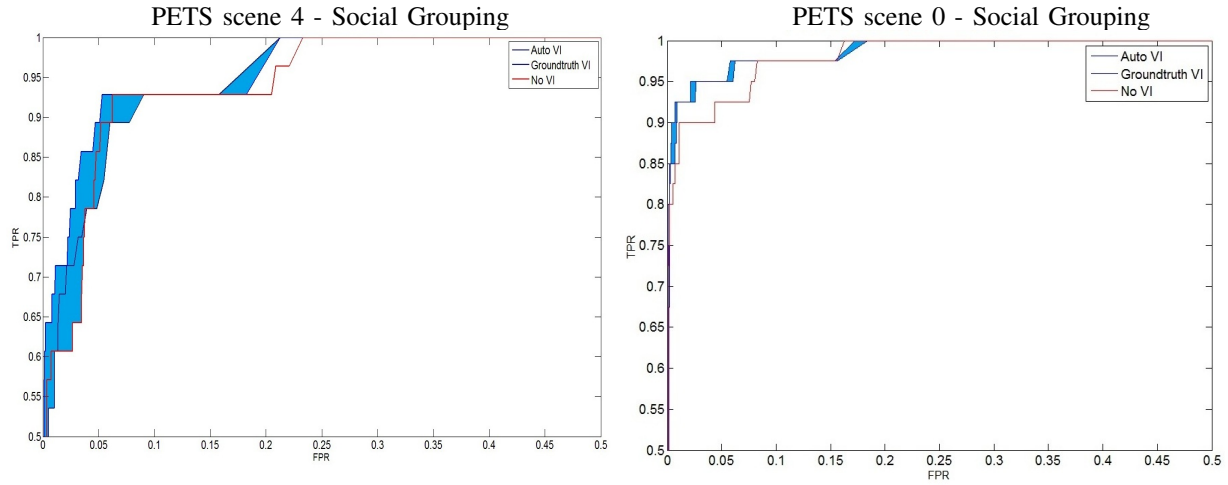
III. Experiment

We wish to evaluate whether the use of visual focus of attention is indicative of social engagement, and if these features can be used to better classify social groups in multiple surveillance datasets. We evaluate the strength of the visual interest features by a comparison of the motion based social grouping and motion plus visual interest social grouping in the following way.

We test upon the publicly available PETS 2007 dataset [4] and the publicly available Oxford town center data [10]. The PETS data offers a source of multi camera real world surveillance footage. The datasets consists of 8 sequences each captured from 4 different viewpoints. We consider the PETS 2007 data to be a crowded scene. The data we use from this dataset contains a total of 372 individuals over 8000 frames, averaging 24 people in the scene at any given frame in a space measuring 16.2 meters by 7.2 meters. Social groups in this scene are characterized by small clusters of 2 - 4 people typically moving together or waiting together. The exception to this are four individuals who are actively engaging in abnormal loitering behavior which separates them for relatively long periods of time. These individuals talk to each other at times in the scene and stand together at times, and as such are still considered to be socially connected. The Oxford data contains 430 tracked pedestrians over 4500 frames. There are an average of 15 individuals in any given frame, with a minimum of 5 and a maximum of 29. We consider this data as sparse to moderately populated. The trajectory motion in the Oxford data is far more structured; the vast majority of individuals travel at walking pace in one of two directions. In the Oxford data the trajectories of socially unconnected pedestrians are often very similar, and often close in proximity - giving the appearance of social connectivity. It is our prediction that the visual interest of pedestrians in this scene will be a relatively strong feature to detect social groups given the motion similarity of socially disconnected people. We evaluate upon 2 non-sequential videos from the PETS 2007. PETS Scene 00 consists of 4500 images, and Scene 04 is 3500 images long. both sequences are imaged at 25fps. The single scene from the Oxford dataset is captured at 25fps and 4500 frames in length. We apply the tracking procedure outlined earlier II upon the jpeg the format images with no other pre-processing.

A. Automatic gazing direction

We use both groundtruth and automatic gaze direction estimations in our experiment. By taking the mean angular error (MAE) between the automatic estimated field of view and the groundtruth gazing direction we found that for the Oxford data we achieved an automatic gaze estimation with MAE of 25.4 degrees compared to the groundtruth. For the more challenging PETS scene 4 data we achieved a MAE of 36.9 degrees. This represents a moderate estimated field



(a) (b)

Figure 4. The true positive rate (TPR) and false positive rate (FPR) from the pair connection likelihood matrix for the method without gaze information (red) and with gaze information (blue). The blue band illustrates the the results with groundtruth gaze direction as it degrades when automatic gaze direction is used. Image (a) shows the results for PETS scene 4 and image (b) for PETS scene 0.

of view offset from the true gazing direction. Our results are comparable to Benfolds results (MAE of 23.9) on the Oxford data [10]. Chen and Odobez achieve an angular error of 18.4 degrees [12] on the Oxford dataset. As of yet there are no published gaze direction statistics for the PETS 2007 dataset, however we consider our results to be particularly good given the far greater deviation from walking direction 2, and lower quality image data than the Oxford data.

B. Visual interest social grouping

We illustrate below the true positive rate (TPR) and false positive rate (FPR) social group classification result for the three sequences we evaluated upon. In each case we ran the motion only social grouping method, the automatic visual interest and motion social grouping, and the groundtruth visual interest and motion social grouping. We use both groundtruth and automatic gazing direction estimates to illustrate the theory under ideal conditions, and to demonstrate the impact of noisy data. The output of our social grouping is a social connection likelihood matrix, entailing the likelihood of each pair of individuals being socially connected, as detailed in the pair strength equation 4. This matrix entails multiple grouping hypothesis, each hypothesis characterized by a different grouping strength threshold. To find the true positive and false positive connections for different grouping thresholds we vary the grouping threshold from 0 to 1 in increments of 0.001; the hypothesis varies from no social connections to everyone in one social group. We find for the following optimal social grouping results by varying the connection threshold:

We find that in each dataset the inclusion of automatic gaze direction into the social grouping model improves the social grouping capabilities for the optimal threshold.

Dataset	Auto TP/FP	GT TP/FP	Motion TP/FP
Oxford	0.90/0.07	**0.93/0.05**	0.88/0.07
PETS S4	0.89/0.06	**0.93/0.05**	0.89/0.06
PETS S0	0.93/0.02	**0.95/0.02**	0.92/0.04

Social Grouping Optimal Results

Table I
WE ILLUSTRATE HERE THE OPTIMAL SOCIAL GROUPING RESULT, SELECTED FROM THE ROC CURVES 5, 4. FOR THE OXFORD DATA, WE SEE A 5.6% IMPROVEMENT IN TRUE POSITIVES AND 28.5% REDUCTION IN FALSE POSITIVES. FOR THE PETS SCENE 4 DATA WE SEE A 4.5% IMPROVEMENT IN TPR AND A 16.6% DECREASE IN FPR. THE PETS SCENE 0 DATA YIELDS A 3.3% INCREASE IN TPR AND A 50% DECREASE IN FPR. GROUND TRUTH GAZE DIRECTION SCORES HIGHEST FOR ALL THREE SEQUENCES AND HAS JOINT OR LOWEST FPR.

For all thresholds we illustrate the improvement that the inclusion of visual attention provides in the social grouping efficacy figures 5,4. The visual attention feature is a subtle and inherently noisy feature, and the motion only method achieves a result close to optimal, as such the improvements are only a small percent of the total value. For the Oxford data, we see a 5.6% improvement in true positives and 28.5% reduction in false positives. For the PETS scene 4 data we see a 4.5% improvement in TPR and a 16.6% decrease in FPR. The PETS scene 0 data yields a 3.3% increase in TPR and a 50% decrease in FPR.

IV. DISCUSSION

Our results provide a strong indication that the inclusion of visual attention improves the capability of the motion based social grouping in crowded human surveillance. We tested upon three video sequences; two PETS sequences considered challenging due to motion complexity, occlusion

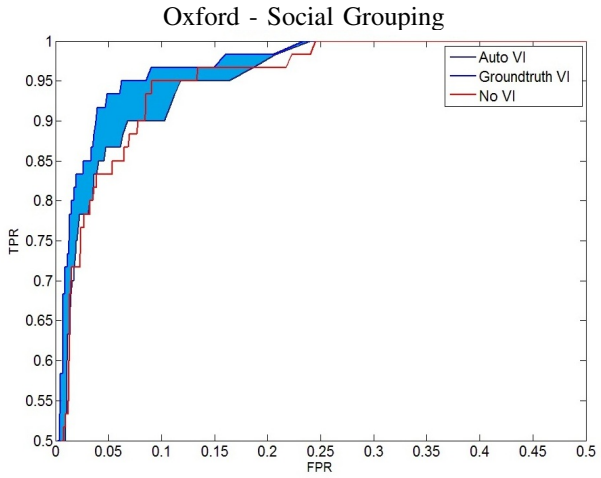


Figure 5. The true positive rate (TPR) and false positive rate (FPR) from the pair connection likelihood matrix for the method without gaze information (red) and with gaze information (blue). The blue band illustrates the the results with groundtruth gaze direction as it degrades when automatic gaze direction is used.

and crowding, and the oxford data which is challenging due to a highly structured scene masking salient social motion. We note that our system shows a susceptibility to gaze direction feature noise. An angular error of average 25 degrees can reduce the efficacy of visual attention and motion social grouping to below that of motion alone in the worst cases 5. However, the predominant result is an improvement when using automatic gaze direction above motion alone, and an even greater improvement when using ground truth gaze direction.

The power of the visual attention feature is that it is independent from the motion influences the environment presents, such as channeling people, queuing areas. The use of the visual attention feature is clearly additive to motion based social grouping. There is however a computational cost to extracting gaze direction features from data. We computed gaze direction estimates as a batch process taking between 8 to 10 hours. However, Benfold [10] has demonstrated this process can be achieved at video rate when the feature space is sub-sampled, and still achieving good accuracy.

Our visual attention social grouping demonstrates, for the first time, the use of visual information in a generalized social grouping task, rather than used to detect specific or anecdotal social events. Our work demonstrates the applicability of visual information upon real world surveillance tasks, using a fully automated system. Our approach is most applicable to scenarios in which there is high motion similarity between social grouped people and un-grouped people, such as airports, stadiums, train stations, and busy town or city surveillance, particularly for use with automated human behavior analysis.

ACKNOWLEDGMENT

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/J015180/1 and the MOD University Defence Research Collaboration in Signal Processing, and in conjunction with the Engineering Doctorate center at Heriot Watt University.

REFERENCES

[1] K. S. L. V. G. S. Pellegrini, A. Ess, "You'll never walk alone: Modeling social behavior for multi-target tracking," *IEEE 12th International Conference*, 2009. 1

[2] N. R. M. Leach, E. Sparks, "Contextual anomaly detection in crowded surveillance scenes," *Pattern Recognition Letters*, 2013. 1, 3, 5

[3] S. G. T. Hospedales and T. Xiang, "Identifying rare and subtle behaviours: A weakly supervised joint topic model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011. 1

[4] 2, 5

[5] R. T. C. W. Ge and B. Ruback, "Automatically detecting the small group structure of a crowd," *IEEE Workshop on Applications of Computer Vision*, 2009. 1

[6] K. P. T. Yu, S. lim and N. Krahnstoever, "Monitoring, recognising and discovering social networks," *IEEE Computer Vision and Pattern Recognition*, 2009. 1

[7] B. R. N. Oliver and A. Pentland, "Statistical modelling of human interactions," *CVPR Workshop on Interpretation of Visual Motion*, 1998. 1

[8] I. D. R. N. M. Robertson, "Automatic reasoning about causal events in surveillance video," *EURASIP Journal on Image and Video Processing*, 2011. 2

[9] L. B. D. T. M. Farenzena, A. Tavano, "Social interactions by visual focus of attention in a three-dimensional environment," *Expert Systems*, 2011. 2

[10] I. R. B. Benfold, "Stable multi-target tracking in real-time surveillance video," 2011. 2, 4, 5, 6, 7

[11] C. Dataset, "http://homepages.inf.ed.ac.uk/rbf/caviar/," 2005, accessed 24/09/2012. 2

[12] J. O. C. Chen, "We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video," *IEEE CVPR conference, Providence*, 2012. 6