Gaze Directed Camera Control for Face Image Acquisition

Eric Sommerlade, Ben Benfold and Ian Reid

Abstract— **Face recognition in surveillance situations usually requires high resolution face images to be captured from remote active cameras. Since the recognition accuracy is typically a function of the face direction – with frontal faces more likely to lead to reliable recognition – we propose a system which optimises the capturing of such images by using coarse gaze estimates from a static camera. By considering the potential information gain from observing each target, our system automatically sets the pan, tilt and zoom values (i.e. the field of view) of multiple cameras observing different tracked targets in order to maximise the likelihood of correct identification. The expected gain in information is influenced by the controllable field of view, and by the false positive and negative rates of the identification process, which are in turn a function of the gaze angle. We validate the approach using a combination of simulated situations and real tracking output to demonstrate superior performance over alternative approaches, notably using no gaze information, or using gaze inferred from direction of travel (i.e. assuming each person is always looking directly ahead).We also show results from a live implementation with a static camera and two pan-tilt-zoom devices, involving real-time tracking, processing and control.**

I. Introduction

The accuracy of face recognition in high resolution images has improved steadily over recent years [18], however in common with most other methods for biometric identification, cooperation from the subject is needed to acquire the required measurement. The ability to automatically capture high resolution face images from a remote camera would allow face recognition systems to be deployed in surveillance situations where people are free to move without constraint.

When only a single person is being monitored, an active camera can simply follow them until the required image is captured, but when multiple persons are present the camera controller must make a decision about which target to follow. Since a face image can only be captured when a surveillance subject looks towards the observing camera, advance knowledge of where the people in a scene are looking can be used to guide the choice of person for an active camera to follow.

In this paper we show how to use gaze estimates from a static camera to optimise the control of one or more active cameras with the aim of maximising the likelihood that the surveillance subjects will be correctly identified. In particular we make three main contributions. First, we show how methods for tracking and control based on expected mutual information gain can be naturally extended to make use of gaze information. Second, we then demonstrate that the use of the (noisy) gaze data is indeed beneficial, yielding improved performance over no knowledge of gaze, or the first-order approximation that a person's gaze direction and their direction of motion are always the same. Finally, we demonstrate a real-time implementation with a static camera and two pan-tilt-zoom devices, involving real-time tracking, processing and control.

The authors are with the Active Vision Lab, Department of Engineering Science, University of Oxford, Parks Road, Oxford, United Kingdom. {eric,bbenfold,ian}@robots.ox.ac.uk

II. Related Work

To date a significant body of research has taken the approach of using static cameras to guide the control of active cameras in distributed visual surveillance systems. Marchesotti et al. [16] used tracking in a single static camera to guide an active camera to capture face images. A similar approach was used by Hampapur et al. [13] but with multiple active cameras. Qureshi and Terzopoulos [19] and Costello et al. [6] looked into the planning and scheduling aspect of target acquisition, drawing from a router analogy. Bagdanov et al. [2] uses reinforcement learning to explore the action space of the control problem, and Del Bimbo and Pernici[8] approached this using the kinetic travelling salesman problem. However, mainly fixed rules are used to decide which cameras should follow which targets, including those which favoured targets walking towards observing cameras to maximise the chance of detecting a face, neglecting the uncertainty in the sensing process. Whereas above papers used geometric insight, others suggested cinematographic rules [10] and analogies to mechanical forces [1].

The emphasis of this work is on the objective function, not the planning stage, as we are interested in the influence of gaze estimation, and the use of uncertainty in the sensing process. Greiffenhagen et al. [12] modelled probabilistically the uncertainty in target tracking and positioning of active cameras to obtain camera parameters guaranteeing given quality bounds on tracking the target. Tordoff and Murray [24] derived similar results for tracking targets with a Kalman filter. Denzler et al. [9] showed how to phrase this control target using entropy as an objective function. More recently, Sommerlade and Reid [22] extended this information-theoretic approach to tracking multiple targets with multiple cameras, where the goal of maximising mutual information between targets and observations results in the desired allocation and swapping of targets between cameras, a behaviour which emerges without being explicitly specified.

The approach described in this paper is unique in using coarse gaze estimates to optimise camera allocation. Previous work on coarse gaze estimation has covered only a few application areas such as attention measurement in surveillance situations [21], [14], [4], providing speaker feedback

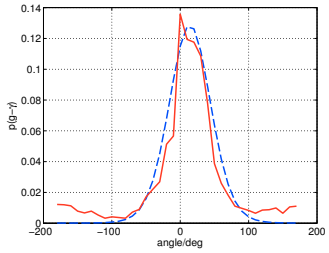


Fig. 1: red: distribution of gaze estimation errors. blue/dashed: maximum likelihood fit to a von Mises distribution

for presentations [11] and identifying the focus of attention of drivers [17] and meeting participants [20], [23].

III. Static Camera Tracking and Coarse Gaze Estimation

The static camera tracker uses the approach of Benfold and Reid [4] who tracked the heads of pedestrians using a combination of sparse optical flow measurements from KLT feature tracking [15] and head detections using Dalal and Trigg's HOG detection algorithm [7] trained on cropped head images. In each frame of video, the sparse optical flow from the previous frame predicts the head location and the head detections provide absolute observations which are combined with the predictions using a Kalman filter. The resulting location estimates provide stable head images which are used for gaze estimation (see figure 11, top row). The 2D image locations are converted into a 3D location estimates by assuming a mean human height of 1.7 metres using the camera calibration with a ground plane assumption.

Coarse gaze estimates are made using randomised ferns, a simplified form of randomised trees. The ferns use binary tests based on gradient directions and colour comparisons to index a leaf node containing a probability distribution over eight direction classes, of 45^o each, thus covering the full range of 360^o. To avoid having discrete direction estimates, a Gaussian Parzen window density estimate is used to interpolate the most likely angle. The resulting gaze estimates as well as target positions are mapped to a coordinate system common to all cameras.

We evaluated the accuracy of the gaze estimation. The classifier was trained on 1475 cropped images harvested from the web and manually annotated. The testing set comprises 4347 head detections resulting from automatic tracking on a video sequence. The mean absolute angular error is 38.275 degrees. Figure 1 shows the distribution of errors $p(g-\gamma)$ for detected angle g and ground truth γ, along with a maximum likelihood fit to a von Mises distribution (see section IV-B).

IV. Camera Control

We base our control method on the following observations. Targets that are known to the system are also said to be "enrolled". The simplest output of a system is a truth value d indicating whether the target is enrolled or not, and the only probabilistic feedback is the identification performance of the system (given actual enrollment value e), which yields the percentage of false positives and false negatives. Furthermore, typical identification systems have a better performance when the target's face is captured frontally. For example, a recent method [5] first looks for frontal face

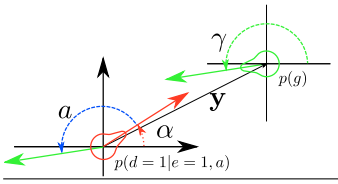


Fig. 2: Top-down view of camera in direction α (red) facing a target with mean gaze γ, yielding an angle of incidence a.

detections, and then tries to identify the captured frames. We denote this ideal identification performance as $p_0(d|e)$.

A. Information theoretic Approach

Our control objective takes into account the uncertainty of the sensing process in the identification system, as well as in the tracking part. For each camera, we want to observe the target that is expected to provide the most information about its identity. This criterion can be expressed in terms of mutual information between the target's state e and the observations d, which depend on the camera orientation α. This information is the difference between the current uncertainty about the target's identity and the remaining uncertainty after it has been observed:

$$I_\alpha = H(e) - H_\alpha(e|d) \quad (1)$$

$$= \sum_{d,e} p_\alpha(e|d)p(d)(\log p_\alpha(e|d) - \log p(e)) \quad (2)$$

where the uncertainties are expressed by (Shannon) entropy and conditional entropy, correspondingly. The best parameter is then the observation angle that provides maximal information. The resulting objective function follows unidentified targets as long as they provide more information than other targets. We now derive the required likelihoods that contribute to the objective function.

B. Observation Likelihood

Figure 2 shows a top-down schematic of a camera viewing a single target with an estimate about its gaze g at an estimated position $\mathbf{y}$, resulting in the angle of incidence a. To model the identification process, we maintain a (latent) belief state e, representing whether the target is enrolled ($e = 1$), or not ($e = 0$). We model the identification process via the conditional likelihood of successfully identifying a target, $p(d|e,a)$, which is specified in terms of the true positive and negative rates of the system. This is a function of the angle of incidence a. The indicator variable d is the binary event of a successful identification measurement from the system, and $a = \alpha - g$ is obtained from the camera's direction and the target's gaze estimate.

More specifically, in our notation, the system's true positive rate with respect to the angle of incidence of the subject's gaze is denoted as $p(d = 1|e = 1,a)$. The distributions over e, $p(e), p(e|d)$ are then interpreted as the prior and posterior belief whether the target is known to the system or not.

At any instant, our model requires a distribution $p(g)$ over the direction of the target's gaze. We obtain the maximum likelihood gaze direction from a gaze detection method [4] (also briefly summarised in section III). We model the uncertainty of the gaze direction with a von Mises distribution:

$$p(g) = f_p(g;\gamma,\kappa_g) = e^{\kappa_g cos(g-\gamma)}/Z_g \quad (3)$$

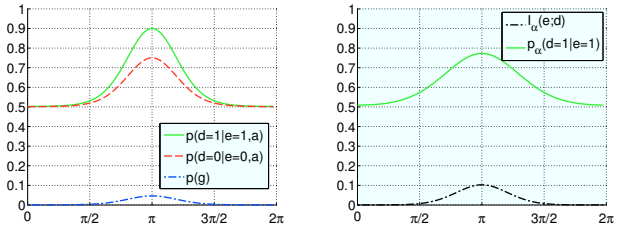


Fig. 3: Left: Sample model of detector performance over angle of incidence a. Green: true positive identification likelihood, $\kappa_1 = 8/\pi$, true positive rate: 0.90. Red/dashed: true negative identification likelihood, true negative rate: 0.75. Blue/dash-dotted: sample distribution over gaze direction g. Right: Detector performance and mutual information for targets placed in a circle around the camera, each with the same distributions over gaze directions as in the left figure. Green: true positive identification likelihood. Black/dash-dotted: resulting information gain.

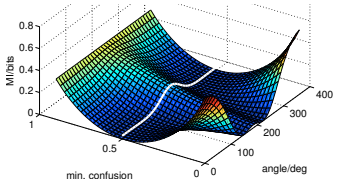


Fig. 4: Surface of mutual information for varying angles and minimum performance setting ε. The mutual information for a minimum confusion of $\varepsilon = 1/2$ is highlighted in white.

which has a mean direction γ, angular covariance κ_g and a normalisation constant Z_g. The spread κ_g is determined from the empirical performance of Benfold's algorithm by a maximum likelihood fit to the estimation errors presented in section III. The resulting mean and angular spread are 212 and 14 degrees.

With an orientation of the camera of α, the angle of incidence a is given as $a = g - \alpha$. Note that we assume perfect knowledge of the cameras' intrinsic and extrinsic parameters (which is realistic given accurate internal calibration and pan-tilt encoder feedback), otherwise the uncertainty in the angle α had to be included in the following derivation.

If a target is known ($e = 1$), we assume that the detection performance for all viewing angles is unimodal and is modelled by a similar function as in equation 3:

$$f(a) = \exp(\kappa_a \cos(a + \pi))) \tag{4}$$

The likelihood of successfully making a correct classification ($d = e$), given performance $p_0(d|e)$ (for a frontal view) and incidence angle a is

$$p(d|e,a) = (p_0(d|e) - \varepsilon) f(a) + \varepsilon. \tag{5}$$

The likelihood of an incorrect classification is then $p(d = i|e = \neg i, a) = 1 - p(d = i|e = i, a)\, i \in 0, 1$. An example for the resulting identification performance can be seen in figure 3 for $\varepsilon = 1/2$. We will return to this constant in the next section.

We obtain the resulting expected detection likelihood – depending on the camera's angle α – by marginalising out the gaze direction g:

$$p_\alpha(d|e) = \int_{0..2\pi} p(d|e,a)\, p(a)\, da \tag{6}$$

This integral has no closed form solution, but can be evaluated trivially for a discrete number of angles (we choose 256). Figure 3(left) shows the expected detection likelihood for a set of targets surrounding the camera.

Regarding the identification process, we now have all elements in place for evaluation of mutual information in equation 2. It is shown in figure 3(right) for a set of targets surrounding the camera. The constant ε in equation 5 now can be understood as the minimum confusion of the classifier. It addresses the fact that the system should be the least certain, or "blind", when the target is facing away. If the likelihood of identifying the target as belonging to the system was zero at this angle, the classifier would be confident about the target class, and provide information. The influence of this threshold on the mutual information is shown in figure 4 for a performance of the classifier of $p(d = e|e) = 0.9$. When the target faces away ($a = 0$), the identification yields a mutual information of 0 for $\varepsilon = 1/2$.

The development so far does not take into account the limited field of view of each camera, which influences the expected visibility of each target. To address visibility, we consider the position $\mathbf{y}$ of the target, which influences the chance of actually making an observation o. In our formulation, this target position does not influence the actual identification capability $p_\alpha(d|e)$, but only the chance of observing the target, i.e. how likely it is to actually capture a bounding box containing the face of the target for a given camera angle.

The expected visibility of the target in the camera's field of view is a scalar

$$w_\alpha = \int_{\Omega(\alpha)} p(o)\, do, \tag{7}$$

where the integral is over the field of view $\Omega(\alpha)$ of the camera which is a function of the of the chosen camera viewing angle α and its zoom value. We model the position of the target using a Gaussian, as would result from tracking using a Kalman filter, and linearise the projective transform so that the observation likelihood is a Gaussian, too. The integral in equation 7 has a simple solution in the form of the error function. This formulation of visibility is inspired by [9], and further details can be found there.

The conditional entropy information term for a target, observed by a single camera, becomes

$$\begin{aligned} H_\alpha(e|\mathbf{y},d) &= -\int_{\mathbf{y}\in\Omega(\alpha)} p(\mathbf{y}) H_\alpha(e|d)\, d\mathbf{y} && (8)\\ &\quad -\int_{\mathbf{y}\notin\Omega(\alpha)} p(\mathbf{y}) H(e)\, d\mathbf{y} && (9)\\ &= w_\alpha H_\alpha(e|d) + (1 - w_\alpha) H(e) && (10) \end{aligned}$$

From this results the mutual information for a target:

$$I_\alpha(e;d,y) = H(e) - H_\alpha(e|\mathbf{y},d) = w_\alpha I_\alpha(e;d) \tag{11}$$

An example is given in figure 5. Note how the mutual information gain drops towards the boundary of the camera's

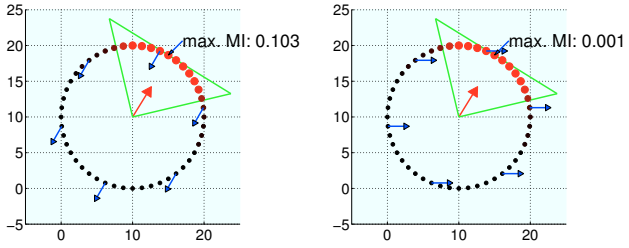


Fig. 5: Different information gains for target positions and views. A camera (centre) surrounded by hypothetical targets, with orientations towards (left), and away from the camera(right). The resulting information gain is proportional to the radius of the circle on each target. The targets facing away from the camera still yield an information gain due to the uncertainty in the gaze estimate.

field of view. The cutoff at the edge of the field of view is more pronounced with a smaller covariance of the target position. As the actual target position and its uncertainty is taken into account, even a target outside of the field of view of the camera yields an information gain, because the uncertainty in its position means there is a chance that it is in fact in the field of view and will be observed. After each report of a successful or unsuccessful identification, we update the hidden target state $p(e|d)$ using Bayes' rule and propagate this to the next observation period as the prior $p(e)$ (i.e. Markovian assumption). This results in a diminishing return for longer observation of the same target once it has been identified. Again, if the employed identification system had a perfect identification performance, the resulting probability $p(e=1|d)$ would be either 1 or 0. Such target state yields an information gain of zero, and would not be considered by the control method any more.

An example based on real data is given in figures 6, 7 and 8, where three targets are on different trajectories

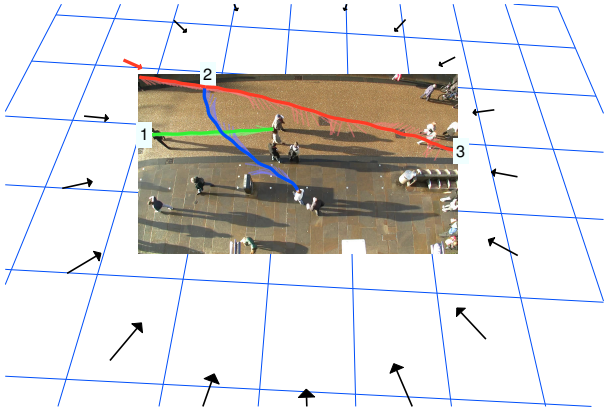
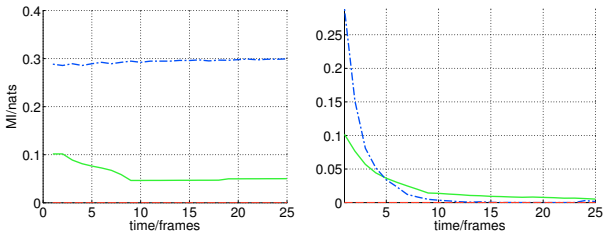


Fig. 6: Example view of the data set and camera positions used for evaluation, including three trajectories used for visualisation in the next plots (best seen in colour), and a camera located in the top left corner. The camera observes targets with varying orientations: towards the camera (blue, 2), away from the camera (red, 3), and slowly turning away from the camera (green, 1).

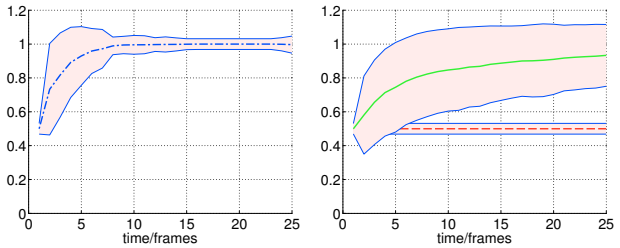


(a) $p(e)$ constant (b) $p(e)$ updated

Fig. 7: The left graph shows the mutual information gain from each target for a constant, indiscriminate belief $p(e)$. Target 1 is green, target 2 blue/dash-dotted, and target 3 red/dashed. Plot 7b visualises the mutual information gain when constantly updating the belief according to simulated detections. Target 1 yields more information gain than target 3 from frame 5 onwards.

with different gaze directions and yield different information gains for longer observations. Targets are observed from a constant camera position and view angle, and observations generated according to the performance of a given detector model (using the same parameters as for the example in figure 3). We repeat the observation process for 100 trials and report average performance. The graph 7a shows how only the target with a frontal view towards the camera is selected if the belief state is not updated. However, graph 7b shows how the second target provides more information after a number of observations, as the first target has been observed long enough. This amount of observations depends on the system's performance (e.g. for a perfect identification a single observation suffices), and the actual result of the observations. The curves 8a and 8b show the development of the successful identification of the targets, $p(e=1)$. Note how for target 2 this converges rapidly towards 1, whereas the target 1 requires more observations. The state of the third target remains uncertain, as it is never facing the camera.

Lastly, we extend our method to several cameras by greedy assignment of targets. All cameras are ordered, and target selection is performed sequentially for each camera. Targets that have been chosen by cameras before are updated by the expected observation of their corresponding camera. This discounts multiple observations of the same target.



(a) Target 2 (blue) (b) Target 1 and 3 (green, red/dashed)

Fig. 8: Development of the belief state $p(e=1)$ for each of the targets.

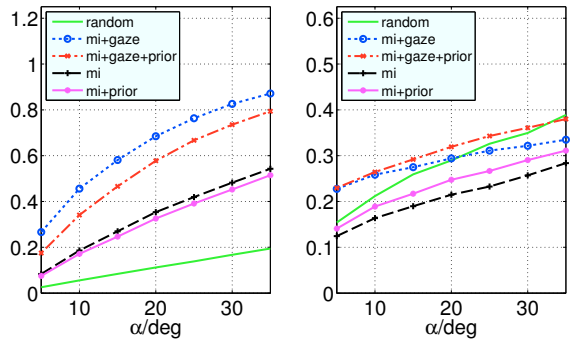


Fig. 9: Performance comparison for camera control using gaze estimates. The detection performance is imperfect, and modelled as in figure 3. Left: percentage of frames where a target has been observed at an angle less than α. Right: percentage of targets observed at least once.

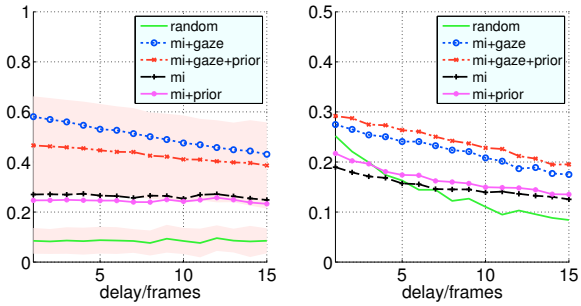


Fig. 10: Left: percentage of frames where a target has been observed at an angle less than 15°, for varying delay between target selection and evaluation. Right: percentage of targets that have been observed at least once at an angle less than 15°.

V. Results

To compare different methods in a fair manner, they should be exposed to the same input data. We choose to use simulation to obtain quantitative and comparable results for different algorithms, and implemented one of the methods in a live system to show the feasibility and qualitative performance of the control method. The recorded images are currently not fed into an actual identification system, which removes any bias from this end. Instead, we posit that frontal images provide better identification performance, and a control method that captures more frontal views should benefit any identification system.

A. Quantitative Evaluation

To obtain quantitative results, we use simulation based on ground-truth data. The data set we use comprises manually tracked and labelled gaze directions from a typical high street, where pedestrians are expected to visually explore the shopping windows. It consists of 75 individual targets in 1500 frames (50s).

A virtual camera is placed at the periphery of all trajectories, such that all targets are visible from this position. We run the simulation for 18 different placements of the virtual camera and report the average performance for the methods. The setup is shown in figure 6.

For each frame and each target visible, we evaluate the expected information gain in equation 11 for a camera setting that centres this target in its own field of view. The camera is then directed to fixate the target which maximises the (expected) information gain.

In order to evaluate the efficacy of our approach using information gain together with gaze, we compare against a variety of other possibilities: (i) the full system uses gaze estimates, together with an update of the latent belief state using Bayes' rule at every frame (denoted mi+gaze+prior); (ii) without a recursive update of the latent belief state, but still using gaze estimates (mi+gaze); (iii) recursive update of the belief state but using target motion direction as an approximation of gaze direction (mi+prior); (iv) no recursive update of the belief state, and using target motion direction as an approximation of gaze direction only; and (v) as a baseline method, we also choose a target at random.

As a metric of the performance we count how often the chosen target looks into the camera in the next frame. We determine this by thresholding the difference between the target's actual gaze and the camera's view direction. We also count the number of targets that have been looking into the camera at least once. This would be the metric for a perfect identification system, as each target could be identified from this single observation.

We performed two experiments. The first experiment measures how much the control methods influence the number of frontal faces captured. For this, we vary the threshold for the observation angle and keep the other parameters fixed ($\kappa_2 = 40/\pi$, detector parameters as in figure 3, delay 1 frame). Figure 9 shows the results. Whereas the random selection method observes roughly as many unique targets as the proposed method, the overall number of frontal observations is far smaller. Note that updating the identification belief state recursively over time increases the number of uniquely observed targets. This is because the mutual information gain decreases with the number of observations of the same target, thus favouring a change of target even though the current target might still face the camera frontally. The performance of the random method is only better if all of the captured images are sufficient to identify the targets – i.e. for a perfect identification system. For such a system, the mutual information gain would be 0 after the first observation, and our proposed method would trigger a faster change of targets, gathering more unique observations.

The second experiment shows the dependency on the temporal delay in the control. After a target has been selected by one of the objective functions, we assume that the positioning of the camera takes a number of frames, fixed for every chosen movement, and during this time no target can be sensed. We vary the delay between one and fifteen frames, which corresponds to $1/30s$ to $1/2s$. Figure 10 shows the average number of persons observed over all viewing angles. The standard deviation for the mutual information based approaches is about $\sigma = 0.16$, and for the random

selection $\sigma = 0.05$, which shows that there is a statistically significant performance improvement when using the gaze based performance, even for a delay of up to half a second in the control method.

The right graph in figure 10 shows the relative number of unique observations for each of the methods. The shaded area depicts the standard deviation of the random selection and the method of mutual information gain from gaze estimates plus updated prior (red/dash-dotted).

As the random selection method does not prioritise targets, it is more likely that targets are selected that do not face the camera or disappear before they can be observed. However, all mutual information based methods perform worse for a longer delay, as it is more likely that the target turns away.

B. Live System

To test the feasibility of our approach, we implemented a live system based on the architecture presented in Bellotto et al. [3], which consists of a network of active cameras and static cameras sharing information via an SQL database. A static camera runs the gaze tracker, and two active cameras are controlled according to the information gained without update of priors. The parameters of the active cameras are pan, tilt and zoom; we exploit the zoom capabilities by allowing the camera to look at several targets at the same time. Instead of selecting a single target with maximum information gain as in the previous section, we vary the parameters for each camera, and sum the information gain from each target observed. A full search of a discretised parameter space runs at two frames per second per camera.

Sample frames from the live system are shown in figure 11. The full video is part of the supplemental material. The cameras follow the targets, keeping those targets centred that look into the camera.

VI. Conclusions

This paper presents an extension of mutual information based tracking of targets by integrating a target's gaze direction. Evaluation based on real data shows that the inclusion of gaze estimates results in more targets captured frontally, which will benefit identification systems.

We choose the von Mises distribution as it is a closed approximation to the wrapped normal distribution and computationally tractable. A further extension of the approach presented here could take into account the varying performance of face detections and classifications resulting from a height difference between the supervised target and the camera. If the camera is located above, yet close to the target, the resulting observations will likely show the top of the head only, and not necessarily support identification. This can be addressed by fitting an appropriate distribution to the performance of an identification method, e.g. a von Mises-Fisher distribution, which is able to model non-isotropic covariances. The integral in equation 6 is then extended to spherical angles, and the remaining derivations follow correspondingly.

In the current implementation of the live system, there is no incentive in the objective function to increase the zoom of the camera. Not only does this reduce the likelihood of making an observation (the field of view in equation 7 is smaller), but as mutual information is never negative, the inclusion of any target in the field of view will result in a higher mutual information gain, unless other targets become less likely to be observed successfully. This, of course, neglects the intuition that higher resolution faces should yield greater identification success. Within the present framework this could be modelled naturally by introducing a dependence of the identification process success rate on the zoom value. As the zoom increases, the false positive and negative rates would drop, resulting greater potential information gain, thereby providing pressure for a camera to zoom in. This is an obvious direction for future work.

References

[1] A. E. Abdel-Hakim and A. A. Farag. Robust virtual forces-based camera positioning using a fusion of image content and intrinsic parameters. In *Proc. International Conference on Information Fusion*, volume 1, pages 522–529, 2005.

[2] A. D. Bagdanov, A. Del Bimbo, and F. Pernici. Acquisition of high-resolution images through on-line saccade sequence planning. In *Proc. ACM International workshop on Video surveillance & sensor networks (VSSN)*, pages 121–130, New York, NY, USA, 2005.

[3] N. Bellotto, E. Sommerlade, B. Benfold, C. Bibby, I. Reid, D. Roth, C. Fernández, L. V. Gool, and J. Gonzàlez. A distributed camera system for multi-resolution surveillance. In *Proc. of the 3rd ACM/IEEE Int. Conf. on Distributed Smart Cameras (ICDSC)*, Como, Italy, 2009.

[4] B. Benfold and I. Reid. Guiding visual surveillance by tracking human attention. In *Proc. British Machine Vision Conference (BMVC)*, September 2009.

[5] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[6] C. J. Costello, C. P. Diehl, A. Banerjee, and H. Fisher. Scheduling an active camera to observe people. In *Proc ACM International Workshop on Video Surveillance & Sensor Networks (VSSN)*, pages 39–45, New York, NY, USA, 2004.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 886–893, June 2005.

[8] A. Del Bimbo and F. Pernici. Towards on-line saccade planning for high-resolution image sensing. *Pattern Recognition Letters*, 27:1826–1834, November 2006.

[9] J. Denzler, M. Zobel, and H. Niemann. Information theoretic focal length selection for real-time active 3-d object tracking. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 400–407. IEEE Computer Society, 2003.

[10] P. Doubek, I. Geys, T. Svoboda, and L. Van Gool. Cinematographic rules applied to a camera network. In *Omnivis2004: The fifth Workshop on Omnidirectional Vision, Camera Networks and Non-Classical Cameras*, pages 17–29, 2004.

[11] T. Gao, C. Wu, and H. Aghajan. User-centric speaker report: Ranking-based effectiveness evaluation and feedback. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1004–1011, 2009.

[12] M. Greiffenhagen, V. Ramesh, D. Comaniciu, and H. Niemann. Statistical modeling and performance characterization of a real-time dual camera surveillance system. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2335–2342, 2000.

[13] A. Hampapur, S. Pankanti, A. Senior, Y.-L. Tian, L. Brown, and R. Bolle. Face cataloger: multi-scale imaging for relating identity to location. In *Proc. IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 13–20, 2003.

[14] X. Liu, N. Krahnstoever, T. Yu, and P. Tu. What are customers looking at? In *Proc. IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 405–410, 2007.

[15] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, pages 674–679, 1981.

[16] L. Marchesotti, L. Marcenaro, and C. Regazzoni. Dual camera system for face detection in unconstrained environments. In *Proc.*

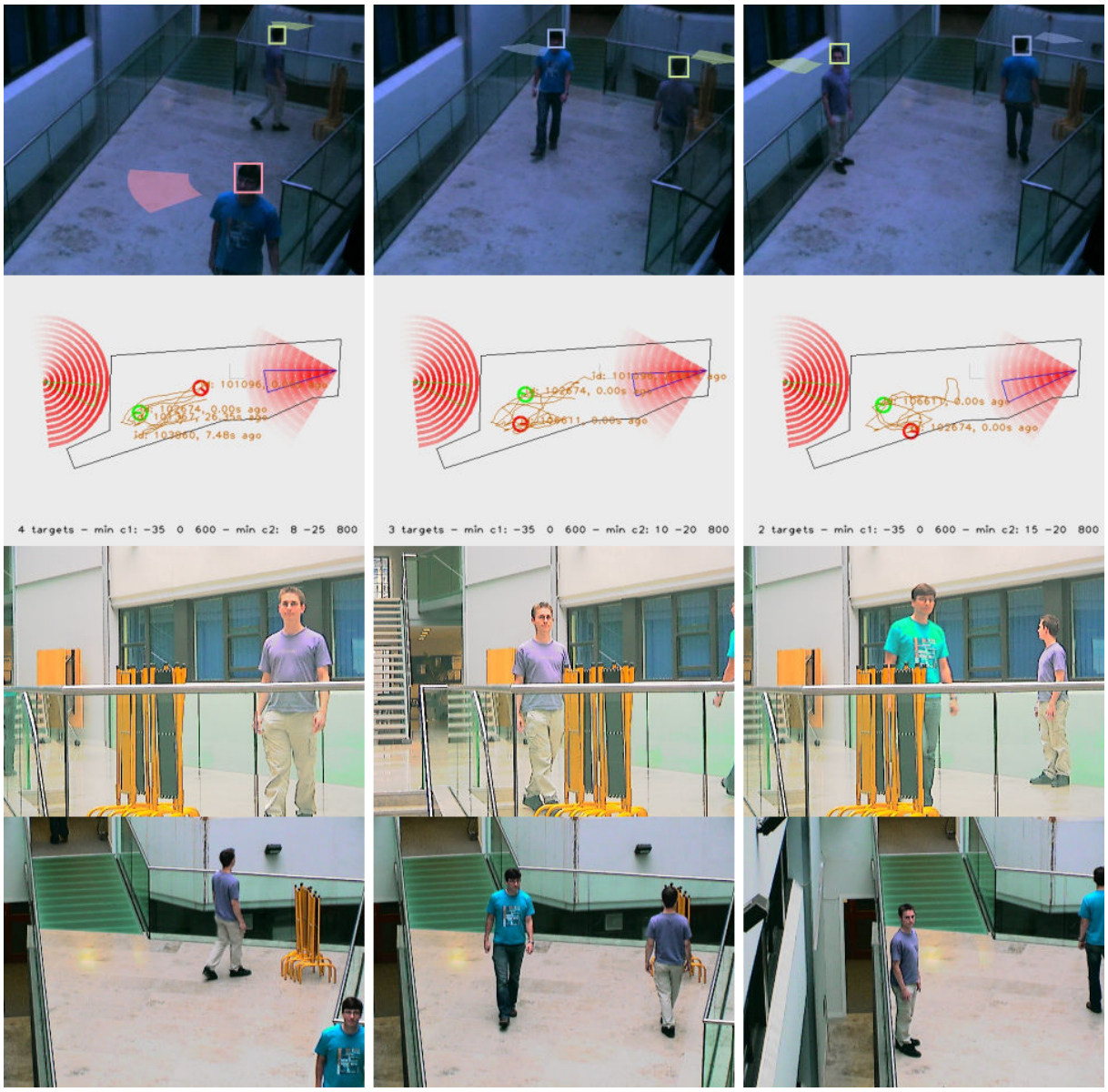


Fig. 11: Sample images showing the operation of the live system at three different times (left to right). The top row shows images from the static camera. The tracked heads are annotated with boxes and their gaze direction represented by a circular section. The second row shows a schematic of the camera control method. The active cameras are depicted as cones. The objective function for pan, tilt and zoom settings for each camera is marked in red. Active targets are circled, and the target with the most information gain for the left camera is green. The trajectories of all targets from the last 30 seconds are drawn. The third row shows the images recorded by the left active camera, and the last row contains images from the right camera.

International Conference on Image Processing (ICIP), volume 1, pages 681–684, September 2003.

[17] R. Pappu and P. Beardsley. A qualitative approach to classifying gaze direction. In *Proc. IEEE Conference on Automatic Face and Gesture Recognition*, pages 160–165, 1998.

[18] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale experimental results. *IEEE TPAMI*, 32(5):831–846, May 2010.

[19] F. Z. Qureshi and D. Terzopoulos. Surveillance in virtual reality: System design and multi-camera control. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[20] J. Sherrah, S. Gong, A. J. Howell, and H. Buxton. Interpretation of group behaviour in visually mediated interaction. In *Proc. International Conference on Pattern Recognition (ICPR)*, pages 266–269, 2000.

[21] K. Smith, S. O. Ba, D. Gatica-Perez, and J.-M. Odobez. Tracking the multi person wandering visual focus of attention. In *ICMI '06: Proc. International Conference on Multimodal interfaces*, pages 265–272, New York, NY, USA, 2006. ACM Press.

[22] E. Sommerlade and I. Reid. Probabilistic surveillance with multiple active cameras. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 440–445, May 2010.

[23] R. Stiefelhagen. Tracking focus of attention in meetings. In *Proc. IEEE International Conference on Multimodal Interfaces (ICMI)*, pages 273–280, 2002.

[24] B. Tordoff and D. Murray. A method of reactive zoom control from uncertainty in tracking. *Computer Vision and Image Understanding*, 105:131–144, 2007.