Active multi-camera surveillance using mutual information

Eric Sommerlade and Ian Reid
Active Vision Group
Dept of Engineering Science
University of Oxford
Oxford, OX13PJ, UK
Email: eric,ian@robots.ox.ac.uk

I. Introduction

The active vision paradigm directly addresses the question of how limited sensing and computationally resources can be best employed to understand a visual scene. Our concern in the present work is the observation of a scene by multiple pan-tilt-zoom cameras and our goal is to control the parameters of each camera in real-time in order to arbitrate between different and potentially conflicting aims. One interest is to obtain the maximum resolution of a target to facilitate classification (e.g. facial recognition, closeups to disambiguate specific gestures, or properties such as their direction of gaze). A second interest is the accuracy of target tracking which is improved by zooming in. But this must be traded off against other interests such as that of minimising the risk of losing a target once it has been detected, or of ongoing observation of the environment in order to minimise the risk of not recording events of importance. Both of these latter interests aims tend to require a wider field of view. In the present abstract we summarise results from various prior aspects of our work [11, 10, 12].

In common with [6, 7, 5, 9], we argue that an objective function based on *expected mutual information* between sensor data and scene representation is an appropriate metric to maximise, but to our knowledge we are the first to employ this in the context of multi-active-camera, multi-target tracking and scene exploration. Before making an observation at time k, we select the best parameter $\mathbf{a}_k$ for this future time step. The parameter $\mathbf{a}_k$ contains all settings for the cameras in our system, i.e. pan, tilt and zoom settings, and is chosen to be the one which is expected maximally to increase knowledge about the state of the scene, $\mathbf{x}_k$. The resulting observations from applying this observation parameter are $\mathbf{o}_k$, which finally update the distribution $p(\mathbf{x})$.

The process of parameter selection at time $k-1$ can thus be summarised as

$$\mathbf{a}_k^* = \arg\max_{\mathbf{a}_k} I_{\mathbf{a}_k}(\mathbf{x}_k; \mathbf{o}_k) \tag{1}$$

In our application, the state vector $\mathbf{x}$ comprises two elements. One part contains all targets currently being tracked, and addresses aims related to tracking, e.g. zoom selection for a particular target and hand-off between cameras. This is explained in more detail in section III.

The other part of the state vector contains the belief about existence of targets at a discrete set of scene points. These targets are to be tracked, but have not been detected yet. How this triggers explorative behaviour of the scene is detailed in the next section.

The resulting behaviours in typical situations for multi-camera systems, such as camera hand-off, acquisition of close-ups and scene exploration, are emergent and intuitive. We quantitatively show that without the need for hand crafted rules they address the given objectives.

II. Scene exploration

In order to quantify the information gain of *searching* for a target, we consider both the prior probabilty on the existence of a target, and the performance of a generic object detector. The former is modelled with a birth-and-death process with equal rates λ, i.e. the appearance of an object is equally likely as a disappearance [4]. Some locations – doorways, for instance – are more liklely to give rise to new targets than others, and so the rate λ is set in a location-dependent manner. Presently this is done from off-line learned patterns of activity, but a sensible future extension would be for this paranmeter to be learned on-line.

Typical object detectors are trained for certain resolution targets and therefore their performance is zoom-dependent. We characterize the performancer of a detector by two functions of zoom level z, $p_z(d|e=0,1)$, (i.e. the chance of a detection given existence or not) representing the performance in terms of true and false positives. These are also learned off-line; in our work we use Histogram of Oriented Gradient whole and upper body detectors [3], and evaluate the true and false positive distributions on data acquired from our cameras and hand-labelled.

Several cameras can observe the same scene location, and so the final information gain (assuming independent observations) gain for C cameras and N locations is thus:

$$I = \sum_{i=1\ldots N} H(\{d_{i,c}\}_C) - \sum_{c=1\ldots C} \hat{H}(d_{i,c}|e_i) \tag{2}$$

where $H(\{d_{i,c}\}_C)$, which is the joint entropy of all measurements for location i. This then is a relatively simple formula

calculated from the detector performance characterizations and the birth-death process.

While the MI does indeed increase for more observations, there are diminishing returns for more and more cameras observing the same location. For better raw detector performance the effect is more pronounced (a perfect detector would have $H(d) = 0$ and no further observations would add information). This tradeoff is important for the collaborative exploration of the scene by several cameras – extensive overlap of the supervised area does not necessarily yield more information than a disparate setting, and thus the MI gain objective naturally leads to cooperative exploration.

III. Tracking with multiple targets

We represent the motion of a target in the scene in ground plane coordinates, facilitating integration of measurements from different cameras. The position of each target is estimated using a sequential Kalman filter [1]. The mutual information gain associated with each tracked target is computed as a function of the differential entropy of a Gaussian (from the Kalman filter) modulated by the overall chance of making an observation, which is governed by the field of view of the camera.

Omitting the derivation, we obtain the mutual information gain for each target as:

$$\begin{aligned} I_{\mathbf{a}}(\mathbf{x};\mathbf{o}) &= H(\mathbf{x}) - \hat{H}_{\mathbf{a}}(\mathbf{x}|\mathbf{o}) \\ &= -n/2 \sum_{c \in C^*} \log |\mathbf{I} - w_c(\mathbf{a})\mathbf{K}_c\mathbf{H}_c|. \end{aligned} \quad (3)$$

where $w_c(\mathbf{a})$ is the probabilty that the target is in the field of view of camera c with parameters $\mathbf{a}$.

This objective leads to natural behaviours such as (i) camera hand-off; (ii) prioiritization of targets in a round-robin fashion.

IV. Combining objectives

We follow Manyika ([7], p129), and note that we can express our multi-objective optimization based on mutual information gains as a single utility which is a simple linear combination of the individual ones. We thus compose the two information gains from detection and tracking via linear blending, which yields a combined utility for both goals – exploration and investigation – of the control:

$$U = \zeta I_{T,\mathbf{a}}(\mathbf{x};\mathbf{o})/I_{T,max} + (1-\zeta)\hat{I}_{N,\mathbf{a}_t}/\hat{I}_{N,max} \quad (4)$$

The parameter ζ can be seen as the control that balances between different objectives: target tracking, versus exploration for new targets.

V. Implementation and Results

We have implemented the theory above on two systems. The first is a simulation environment comprising real high-defintion video of a surveillance scene (PETS 2001) in which we simulate pan, tilt and zoom via image scaling and cropping. This allows experiments under exactly repeated conditions and has therefore been crucial in exanining and evaluating the ideas.

In order to implement our objective function for scene exploration, we have quantized the pan and tilt values into M values (not necessarily evenly spaced) and zoom into N steps. The choice of parameters then reduces to an exhaustive search over the $(M^2N)^C$ parameters. For modest C (i.e. 2 or 3) the search space is not unreasonably large, but rapidly becomes unwieldy for four or more cameras.

The second system is a live, real-time system comprising two pan-tilt-zoom cameras. The architecture of this system, which supports a variety of heterogeneous cameras and visual processing is detailed in [2]. For our MI-based control, we perform target detection using a GPU-accelerated implementation [3] (described in [8]). Once detected for the first time, targets are tracked using repeated detection and simple nearest neighbour data association. Example videos showing target hand-off can be seen at `http://www.youtube.com/user/PTZEric`.

VI. Conculsions

We have presented a unified method using maximisation of mutual information to control multiple heterogeneous cameras observing a common environment with multiple targets. Basing our system's overall objective function on the mutual information between observations and the scene representation means that we can naturally combine apparently disparate aspects of the problem, such as detector performance, and actor appearance and disappearance rates, and disparate goals, such as exploration and tracking.

References

[1] Y. Bar-Shalom and T. E. Fortmann. *Tracking and data association*, volume 179 of *Mathematics in Science and Engineering*. Academic Press Professional, Inc., San Diego, CA, USA, 1987.

[2] N. Bellotto, E. Sommerlade, B. Benfold, C. Bibby, I. Reid, D. Roth, L. V. Gool, C. Fernández, and J. Gonzalez. A distributed camera system for multi-resolution surveillance. In *Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC 2009)*, 2009.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page 1:886893, 2005.

[4] B. Gnedenko. *Theory of Probability*. Mir Publishers, Moscow, 3rd edition, 1976.

[5] B. Grocholsky. *Information-theoretic control of multiple sensor platforms*. PhD thesis, The University of Sydney, 2002.

[6] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, e fficient algorithms and empriical studies. *Machine Learning Research*, 9, 2008.

[7] J. Manyika and H. Durrant-Whyte. *Data Fusion and Sensor Management a decentralized information-theoretic approach*. Electrical and Electronic Engineering. Ellis Horwood, Chichester, UK, 1994.

[8] V. Prisacariu and I. Reid. fasthog - a real-time gpu implementation of hog. Technical Report 2310/09, Department of Engineering Science, Oxford University, 2009.

[9] N. Roy and C. Earnest. Dynamic action spaces for information gain maximization in search and exploration. In *Proceedings of the American Control Conference (ACC 2006)*, Minneapolis, USA, 2006. IEEE.

[10] E. Sommerlade and I. Reid. Influence of zoom selection on a kalman filter. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, September 2008.

[11] E. Sommerlade and I. Reid. Information theoretic active scene exploration. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR)*, 2008.

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