Tracking People with Active Cameras Using Variable Time-Step Decisions

Alparslan YILDIZ†*, Noriko TAKEMURA†, Nonmembers, Maiya HORI††, Yoshio IWAI††, and Kosuke SATO†, Members

SUMMARY In this study, we introduce a system for tracking multiple people using multiple active cameras. Our main objective is to surveille as many targets as possible, at any time, using a limited number of active cameras. In our context, an active camera is a statically located pan-tilt-zoom camera. In this research, we aim to optimize the camera configuration to achieve maximum coverage of the targets. We first devise a method for efficient tracking and estimation of target locations in the environment. Our tracking method is able to track an unknown number of targets and easily estimate multiple future time-steps, which is a requirement for active cameras. Next, we present an optimization of camera configuration with variable time-step that is optimal given the estimated object likelihoods for multiple future frames. We confirmed our results using simulation and real videos, and show that without introducing any significant computational complexities, it is possible to use active cameras to the point that we can track and observe multiple targets very effectively.

key words: active cameras, human tracking, variable time-step, PTZ camera, multiple camera geometry

1. Introduction

There is a vast amount of research on tracking multiple targets with multiple cameras [3]–[5], [7], [9], [15]. In many practical applications such as security systems, entertainment systems, surveillance and behavior analysis systems, tracking multiple targets is a vital step. Most research in the literature focuses on tracking with static cameras that are fixed in terms of position and viewing angle throughout the entire life of their use [10], [11]. However, there are many cases where fixing the camera position and viewing angle would introduce limitations such as not being able to observe all the targets. Therefore, we use a statically located pan-tilt-zoom (PTZ) camera, and the system using these cameras would be expected to observe more efficiently and obtain a higher coverage of targets.

In the current literature, there is considerable research on tracking with moving or active cameras [6], [12]–[14]. These studies focus on the detection of the human body in camera views; however, this creates some difficulties. When the number of people being tracked increases, the computational cost and expected error also increase. In this research, we adopt the ground occupancy map method used in our previous work [2]. The main advantage of using the occupancy map is that we do not need to explicitly detect the targets in camera views. All the foreground information from the camera views are fused into a common ground map, which makes it possible to predict future time-steps more smoothly. A ground occupancy map is also highly efficient as it is independent of the number of targets being tracked. This is a desirable property because with active cameras, all the processing must be done in real time to be able to make decisions on moving the camera.

In a multi-active camera multi-person tracking system, for accurate decisions on moving a camera, future time-step prediction is essential. In our previous work [1], we presented a behavior-aware human tracking and prediction method, and confirmed its effectiveness in predicting future time steps. In [1], future time-steps were predicted by efficient employment of local behavioral predictions and propagating the occupancy forward. However, multiple time-step predictions were simply achieved by consecutive application of the predictors. In this paper, we upgrade this method by reverse propagation of the occupancy together with pre-computed look-up tables, which makes it possible to compute multiple time-step predictions very efficiently.

In active camera tracking, there is a strong relationship between consecutive camera decisions. Naturally, any decision affects all future decisions as the camera movements need to be smooth. The cameras cannot jump instantly to any configuration, and must only update their configurations smoothly. This natural behavior of active cameras easily leads them to local minima (getting stuck following the same target all the time). This can be overcome by employing multiple future time-step estimates. This will enable the optimization process to decide to temporarily lose some targets, if in the long run, the total coverage of targets increases. For example, if a lost target is expected to be captured shortly by another camera, it is preferable to lose one target temporarily and to capture new targets and increase the total coverage. On the other hand, using more predictions will not improve the coverage indefinitely. As we make more predictions, the certainty will decrease, resulting in misinformed decisions. We make camera decisions based on the criterion that only the earlier and more certain $k$ number of predictions should be used, because further smoother/weaker predictions only corrupt the decisions. In our results we empirically show that this intuition works very well in real scenarios.

Manuscript received October 15, 2013.

Manuscript revised February 20, 2014.

†The authors are with Osaka University, Toyonaka-shi, 560–8531 Japan.

††The authors are with Tottori University, Tottori-shi, 680–8550 Japan.

a) E-mail: yildiz@sens.sys.es.osaka-u.ac.jp

DOI: 10.1587/transinf.E97.D.2124

Copyright © 2014 The Institute of Electronics, Information and Communication Engineers
Our previous work [2] presented a system for tracking humans that is capable of tracking moving targets with multiple active cameras. This research is fundamentally different from [2], with the same objective. In our previous work, we employed a detection-less tracking and estimation method that did not incorporate previous behaviors, as our current tracking method does. Furthermore, our previous work used a constant number of future time-step estimates based on the panning speed of the active camera. Fixation on the prediction length limits the system in situations where further prediction will lead to more accurate camera movement decisions. At each time-step, a different prediction length will lead to an optimal decision. Thus, in this research, we present variable time-step predictions that are computed for each frame. The presented method is optimal in terms of maximized coverage.

In Sect. 2, we review our tracking method and in Sect. 3 we present our main contribution for optimal camera decisions with variable time-step decisions. Finally we present experimental results in Sect. 4 and provide concluding remarks in Sect. 5.

2. Future Time-Step Estimation

In multi-active camera human tracking, a highly accurate prediction method is required, as active cameras move relatively slowly with respect to acquisition speed, and camera movement decisions are solely dependent on the multiple time-step predictions of human locations. Localizing and predicting a human position is a well-studied subject in the literature [3], [5], [9]; however, predicting multiple future time-steps is still an open subject. For our accuracy requirements, we employ our previous method for tracking and predicting human motion in closed environments [1]. This method trains local behavior predictors via Support Vector Machine regression and adaptively fuse them to the linear motion model depending on the availability of training data (See Eq. (1)). In the presence of sufficient data, a higher weight is assigned to the learned predictors leading to a behavior-aware predictor. In the absence of sufficient data, the system automatically reverts to the linear motion model. For details see [1].

\[
\text{PredictionModel} = \alpha \text{LearnedPredictionModel} + (1 - \alpha) \text{LinearMotionModel}
\]

In the following subsection, we show how to compute multiple time-step predictions efficiently with the method presented in [1] using reverse propagation of occupancy.

2.1 Fast Multiple Time-Step Estimation

We now revisit the method presented in [1]. Consider a particle representing the location of a person on the ground plane. In the first prediction step, this particle is guided using a prediction model. In this model, each particle, or person, is represented by an occupancy value, which the particle carries along the predicted direction. At some locations on the ground plane, the prediction model may have forks and divergences, and in some locations the model may predict linear directions. If a particle arrives at a fork, we split the particle into two with the corresponding predicted directions and propagate them independently. The occupancy value is also split among the new particles such that the total occupancy on the whole ground map is preserved. Splitting the occupancy is straightforward as it is directly proportional to the confidence of the prediction model in possible directions.

For multiple time-step estimations, we focus on a sample case in Fig. 1 (a). In this figure, the prediction model is shown by directed arrows. The location A in the figure has a 0.8 occupancy value. This occupancy is propagated into the adjacent cells according to the prediction model. First, the value 0.8 is split into 0.5 and 0.3 according to the confidence levels of the prediction model. This is the estimate of a single time-step. Next, the occupancy value 0.3 is further split and propagated into 0.1 and 0.2 occupancy values in the adjacent locations. This is the estimate of two time-steps.

For multiple predictions the direct method is to repeat the same process with the new particles as many times as required. Each particle, or person, can be processed individually and in parallel if we reverse the predictions, as shown in Fig. 1 (b). In this case, each location on the ground plane will make a reverse prediction, namely each location will put out candidate locations from which a particle may enter. This is achieved with no additional cost by reversing the prediction models during learning. Next, for each location, the occupancy values in the candidate locations are summed up to be the predicted occupancy. This process does not have any write operations into the same location and can be done in parallel very efficiently. For the prediction of the \(k\)-th time-step, this process is simply repeated \(k\) times.

In practice, local predictors pass the existent occupancy among each other either directly or via splitting/joinning. This property is easily employed to track any lost target. When a target goes out of sight, the corresponding ground region is not directly erased but rather reduced in occupancy by a simple forgetting effect, and the reduced occupancy is continued to pass along. On the other hand, if a location is observed by any of the cameras, the corresponding ground region is updated accordingly.

![Fig. 1](image-url)
3. Camera Configuration Optimization

In this section, we describe our main objective and show how we can optimally solve variable time-step decisions with Dynamic Programming (DP). Our main objective is to find the optimal camera configuration that maximizes the coverage of the targets, given the estimated positions in multiple time-steps.

The active cameras we employ in this study cannot move instantly, thus at each time-step, we need to select the configurations from among those reachable in the current configuration, resulting in smooth movement of the cameras. While this fact limits the search space in a beneficial way, it also requires that we consider not only the immediate next time-step, but multiple time-steps, to make the most accurate decision. The above fact forces us to build a smooth sequence of camera configuration decisions based on multiple time-step predictions. However, the number of time-steps required is unknown. In this research, we optimally solve for the number of prediction time-steps and the configurations from among those reachable in the current configuration. For a single camera, and considering only the panning operations, we have 3 feasible configurations, namely \( \{ \text{left}, \text{right}, \text{stay} \} \), for each possible pan-tilt configuration. Here \( \text{f}^{t+1}_c \) is a suitable encoding of camera configurations for \( N_c \) cameras. Maximizing occupancy coverage on \( G^+ \) is formulated as

\[
\max_{F} \frac{1}{k} \sum_{i=1}^{k} \sum_{c=1}^{N_c} M_c(f^{t+1}_c) \otimes G^{t+1}_c,
\]  

where \( M_c(f^{t+1}_c) \) is the camera view mask of the \( c \)-th camera on the ground plane for the camera configuration \( f^{t+1}_c \) and \( \otimes \) simply sums the occupancy values under the camera view mask. Equation (2) maximizes the occupancy coverage, which is the total occupancy under all camera view masks.

Optimizing Eq. (2) is straightforward when \( k = 1 \). The total number of feasible configurations for \( F^{t+1} \) is relatively small because it should be a neighbor configuration to the current configuration. For a single camera, and considering only the panning operations, we have 3 feasible configurations, namely \( \{ \text{left}, \text{right}, \text{stay} \} \), for each possible panning movement of the camera. For \( N_c \) cameras, the size of the feasible set is \( 3^{N_c} \), which is 81 for 4 cameras. Similarly, including tilt configurations would lead to \( 9^{N_c} \) (6561 for 4 cameras) feasible configurations at each step. Because combined camera view masks can be pre-computed, we only need to solve a DP problem with a table size of 6561 \( \times \) \( K \). The majority of the computation time is needed for the multiple time-step decisions, thus both pan-tilt configurations can easily be optimized for multiple cameras on today’s computers.

For the case \( k > 1 \), Eq. (2) is a well-known 1D optimization of a smooth sequence of length \( k \), which can be solved optimally using DP, as shown in the following subsection.

3.1 Fast Computation of a Variable Time-Step Decision

Given \( K \) future predictions of the ground occupancy map \( G^+ = \{ G^{t+1}_c, G^{t+2}_c, \ldots, G^{t+k}_c \} \), Eq. (2) can be solved optimally via DP. For multiple active cameras we have the joint states as \( f = \{ f_1, f_2, \ldots, f_{N_c} \} \). If we let \( (f, N_c) \) enumerate all feasible joint camera configurations for \( N_c \) cameras, we can construct a DP table as follows.

| \( k = 1 \) | \( k = 2 \) | \( \ldots \) | \( k = K \) |
|---|---|---|---|
| \( (f, N_c)_1 \) | \( (f, N_c)_2 \) | \( (f, N_c)_3 \) | \( \ldots \) |

Each column of the above table is filled from left to right, and each cell contains the coverage value corresponding to the configuration \( f^{t+1} \) on \( G^{t+1} \). Once the table is full, back tracking can be performed to extract the optimal sequence of camera configurations.

A very useful property of our decision method is that it employs the same predicted ground occupancy maps \( G^+ \) for all sub problems of different lengths. As it perfectly suits the nature of DP, we can easily extract optimal camera configuration sequences using the same table. As the above table fills from left to right, once a column is computed, back tracking from that column is performed to extract the corresponding optimal sequence. In this way, we can extract \( K \) optimal sequences with their corresponding coverage values using the same DP table. Once the optimal sequences are computed, we finally pick the sequence with the highest coverage and apply its first node as the camera configuration decision for the current time-step.

3.2 Discussion

In Sect. 3, the immediate next frame is represented by \( t + 1 \) superscript and the following predictions are represented by \( t + 2, t + 3, \ldots \) superscripts on the ground occupancy map \( G \). However, in practice these maps do not align with the actual frame intervals of the video streams from the cameras. The reason for this is the physical speed of active cameras. Consider some standard PTZ cameras, such as Sony EVI-D100 [17] with a typical panning speed of 180° in 1 s, thus, reaching one neighboring panning configuration in about 60 ms. To make the computations seamless with simulations and real videos, we need to calibrate the panning speed of the cameras and align the camera acquisition such that the frame interval is exactly matched with the camera movement speed of 60 ms, resulting in an acquisition of
16.6 fps or 8.3 fps depending on the acquisition speed.

We also need to calibrate the overhead needed to start the camera movement, which is less than 60 ms. This is required to issue the command and start the motor that moves the camera. Thus, during multiple time-step predictions, we need to make one extra prediction to cover the overhead and return the multiple prediction maps starting from the second actual prediction.

In our formulation, the pan/tilt space is discretized and the rows of our DP table are enumerated accordingly. However, it is also possible to formulate the optimization in a continuous way over the values of the pan/tilt configurations. This way, the loss function to be optimized will be the amount of occupancy missed by the current camera configuration. We would need to compute the Jacobian (or Hessian) over the pan/tilt parameters. However, the loss function for this problem is not smooth, thus the first/second derivatives do not exist everywhere. A smooth change in the loss function only happens when objects cross the boundaries of the camera viewing masks. In practice this rarely happens. With some care, e.g., taking the nonzero columns of the Jacobian matrix to update only a subset of parameters. This method cannot be extended to multiple time-steps readily because there is a time dependency between consecutive pan/tilt parameters. For this problem, our DP formulation is simpler and is guaranteed to give an optimal solution (with respect to the objective function) without any special care.

4. Experiments and Results

We performed simulation experiments and real video experiments with POM (Fig. 3) and PETS-2006 (Fig. 4) datasets to evaluate the performance of our method. In the simulations, we built a closed space where more than 10 subjects were moving with various goals. Figure 2 shows the top view of the synthetic room where camera locations are marked with circles and doors are marked with gray rectangular blocks. In this experiment, subjects in the room move from one of the doors to any other by an almost linear motion throughout the room. Simulated subjects reach their goals by visiting predefined locations on the ground and finally exiting through one of the doors. The motion history is also accumulated and shown in Fig. 2.

In our experiments, we report on accuracy as the ratio of total subjects under camera view masks to the total number of subjects based on the ground truth location information. In the accuracy graphs in Figs. 5, 7 and 8, lines marked with the Prediction-prefix represent the accuracy of our method when we use a constant \( k \) for all the videos. This means that at each time-step we compute a single sequence of camera configuration decisions and apply the first decision in that sequence. In our experiments, we show that altering the value of \( k \) dramatically affects the final accuracy. In the same graphs, the horizontal line shows our performance when we select the optimal \( k \) at each time-step as described in Sect. 3. We also report the upper-bound on the accuracy as the lines marked with GroundTruth-prefix. The upper-bound values are computed using the ground-truth locations of subjects instead of predictions. In all our experiments, we show that our method achieves results reasonably close to the upper-bound while always maintaining a greater accuracy than any constant length decision.

In Fig. 6, we show the selected \( k \) values using our
method for the simulation videos. In most of the frames a shorter decision sequence leads to higher accuracy. This is because further predictions become smoother and the occupancy values more scattered, while in the immediate next time-step prediction $G^t+1$, the occupancy values are condensed in smaller areas.

In Figs. 7 and 8 we show the performance of our method on “terrace” videos of POM set, and PETS-2006 datasets, respectively. In these video sets, we have synthesized active cameras by extracting narrow field-of-view subsections out of full frames. We evaluated the accuracy using manually marked human locations in all videos.

On the “terrace” videos, the crowd density is relatively high, resulting in relatively lower coverage value. However, our variable time-step decision method performs better than any constant length decision on these videos. PETS-2006 videos on the other hand have a relatively low resolution because crowd density in them is relatively low. For different lengths of decisions, we usually obtain the same decision sequence on these videos. Our variable time-step decision method again outperforms the constant length decision scenario because it is able to alter the length of the decision sequence when necessary.

Finally, we would like to note that the initial value of the blue plots in Figs. 5, 7 and 8 represents the accuracy of the $k = 1$ case, which is identical to our previous work in [2]. We have shown that in all experiments, variable time-step decision making gives a higher rate of coverage than constant length or immediate decisions ($k = 1$).

5. Conclusions and Discussion

We presented a variable time-step human tracking system using active cameras. Our system consists of two main parts, namely, tracking and estimating human occupancy on the ground plane, and optimizing camera configuration decisions at each time-step. Our tracking and estimating algorithm is capable of producing multiple time-step estimates into the future. Our camera configuration optimization algorithm considers all the estimated occupancy maps simultaneously and optimally selects the best time-step length with the highest occupancy coverage criteria.

The major advantage of our method is that both the estimation and optimization parts have natural recurrence relationships, which let us produce multiple time-step estimations and optimizations without introducing additional complexities. Our tracking method learns local motion patterns
of the environment and uses them to estimate multiple time-steps into the future. Furthermore, our decision optimization method is optimal in terms of the coverage of ground occupancy given the estimated occupancy maps. We exploit the recurrence relationship among the decision sequences and employ DP to quickly solve variable time-step optimization.

In our formulation, we only used the pan/tilt feature of the active cameras. Most active cameras, such as ours, have a zoom feature as well, which can be used to keep the targets in the camera view at a predefined scale. Our future work includes in depth investigation of the zoom feature for active camera tracking as it requires additional care. The zoom option would alter the field of view, and hence the relative speed and shape of the targets between different configurations, making it a challenging optimization problem.

Our method can also be applied to any type of object if the ground occupancy map can be computed. For objects abiding by the ground plane assumptions, such as non-flying objects, the ground occupancy map can easily be computed using simple background subtraction and planar homographs. Tracking cars or other moving vehicles using active cameras would be a very useful application of our method.

Acknowledgments

This work was partially supported by JST CREST (Core Research of Evolutional Science and Technology) research promotion program “Creation of Human-Harmonized Information Technology for Convivial Society” Research Area.

References

[1] A. Yildiz, N. Takemura, Y. Iwai, and K. Sato, “Regression based trajectory learning & prediction for human motion,” Workshop on Geometric Computation for Computer Vision, 2013.
[2] A. Yildiz, N. Takemura, Y. Iwai, and K. Sato, “Tracking people with active cameras,” International Conference on Human Computer Interaction, Nevada, 2013.
[3] A. Mittal and S. Larry, “M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene,” Int’l J. Comput. Vis., vol.51, issue 3, pp.189–203, 2002.
[4] S.M. Khan and M. Shah, “Tracking multiple occluding people by localizing on multiple scene planes,” IEEE Trans. Pattern Anal. Mach. Intell., vol.31, no.3, pp.505–519, 2008.
[5] A. Yilmaz, O. Javed, and M. Shah, “Object tracking: A survey,” ACM J. Computing Surveys, 2006.
[6] N. Takemura, Y. Nakamura, Y. Matsumoto, and H. Ishiguro, “A path-planning method for human-tracking agents based on long-term prediction,” IEEE Trans. Syst. Man Cybern., C, vol.42, no.6, pp.1543–1554, 2012.
[7] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank, “Principal axis-based correspondence between multiple cameras for people tracking,” IEEE Trans. Pattern Anal. Mach. Intell., vol.28, issue 4, pp.663–671, 2006.
[8] F. Fleuret, J. Bercicz, R. Lengagne, and P. Fua, “Multicamera people tracking with a probabilistic occupancy map,” IEEE Trans. Pattern Anal. Mach. Intell., vol.30, issue 2, pp.267–282, 2008.
[9] J. Bercicz, F. Fleuret, and P. Fua, “Robust people tracking with global trajectory optimization,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2006.
[10] K. Kim and L. Davis, “Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering,” Proc. ECCV, 2006.
[11] S.M. Khan and M. Shah, “A multiview approach to tracking people in crowded scenes using a planar homography constraint,” Proc. ECCV, 2006.
[12] Y. Xie, L. Lin, and Y. Jia, “Tracking objects with adaptive feature patches for PTZ camera visual surveillance,” International Conference on Pattern Recognition, 2010.
[13] A.D. Bimbo and F. Pernici, “Uncalibrated 3D human tracking with a PTZ-camera viewing a plane,” 3DTV Conference: The True Vision — Capture, Transmission and Display of 3D Video, 2008.
[14] C. Huang and L. Fu, “Multitarget visual tracking based effective surveillance with cooperation of multiple active cameras,” IEEE Trans. Syst. Man Cybern., B, Cybern., vol.41, no.1, pp.234–237, 2011.
[15] A. Yildiz and Y. Akgul, “A fast method for tracking people with multiple cameras,” Third Workshop on HUMAN MOTION Understanding, Modeling, Capture and Animation, Greece, 2010.
[16] PETS 2006 Benchmark Set, http://www.cvg.rdg.ac.uk/ PETS2006/data.html
[17] Sony EVI-D100 PTZ Camera, http://pro.sony.com/bbsc/ssr/ cat-industrialcameras/cat-robotic-product-EVID100/

Alparslan Yildiz completed his B.Sc. and M.Sc. in the Gebze Institute of Technology, in 2007 and 2009, respectively. In 2011, he enrolled in Osaka University for the Doctoral Program in Computer Science. His research interests include multiple-camera geometry, human tracking, pattern recognition and machine learning methods in computer vision.

Noriko Takemura received her B.Eng., M.Eng., and Ph.D. (Eng.) degrees from Osaka University, Osaka, Japan, in 2006, 2007, and 2010, respectively. In 2010, she joined the Osaka University as a Research Fellow. Her research interests include multiple-camera geometry, human tracking, pattern recognition and distributed vision systems.

Maiya Hori received her B.Eng. degree in systems science from Osaka University in 2005. He received his M.Eng. and Ph.D. degrees in information science from the Nara Institute of Science and Technology in 2007 and 2011, respectively. He is currently a Research Fellow at Tottori University. His research interests include computer vision and pattern recognition. He is a member of IEEE.
Yoshio Iwai graduated from Osaka University in 1992 and completed his M.Sc. and doctoral programs in 1994 and 1997, respectively. He was then appointed a Research Associate at the university, subsequently becoming an Associate Professor. From 2004 to 2005, he was a Visiting Researcher at Cambridge University. He is currently a Professor in the Graduate School of Engineering at Tottori University. He is engaged in studies related to computer vision. He is a member of IEEE, the Information Processing Society, and Japanese Robotics Society. He holds a D.Eng. degree.

Kosuke Sato received his B.Sc., M.Sc., and Ph.D. degrees from Osaka University, Japan, in 1983, 1985, and 1988, respectively. He was a Visiting Scientist at the Robotics Institute, Carnegie Mellon University, from 1988 to 1990. He is currently a Vice Dean of the Graduate School of Engineering Science, Osaka University. His research interests include image sensing, 3D image processing, digital archiving, and virtual reality. He is a member of ACM and IEEE.