21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Camera Pan / Tilt Control with Multiple Trackers

Yiming Li and Bir Bhanu
Center for Research in Intelligent Systems, University of California, Riverside, CA 92521
yimli@ee.ucr.edu, bhanu@cris.ucr.edu

Abstract

In this paper, we consider the multi-camera tracking and the camera active control (pan and tilt). Auction mechanism from economics is developed to choose the best available camera. By modeling the camera bids with prior knowledge of the camera homographies, the system can "think" ahead to perform necessary panning or tilting operations. The uncertainties of homographies are considered inherently in the metrics used for computing camera bids. Further, to have a better tracking result, we use multiple trackers simultaneously. The trackers are rectified periodically based on the previous auction results. The proposed approach is evaluated in a real-world camera network.

1. Introduction

Video surveillance in a camera network requires the collaboration and competition among cameras. The problem of efficient cooperation among multiple cameras has risen to the forefront of the video sensor networks. There are many existing works that consider how to operate multiple cameras in a video network to track objects efficiently. However, to the authors' best knowledge, there are very few works that consider a potentially available camera for tracking by panning or tilting the camera to another position. This is meaningful because there can be cases when an object is visible in some cameras, but none of these camera provide a preferable view while panning or tilting one of these cameras or some other camera may render a better view of the same object.

There is a large amount of work done in the field of multi-camera multi-person tracking [1, 2, 3, 4, 5, 6]. There are also some other works done in the field of active camera control. For example, in [7], the authors proposed an approach to do camera pan / tilt control for SLAM. However, this approach is for single camera only. In [8], the authors proposed a system by using two static cameras together with a PTZ camera to have a close-up look for face acquisition and recognition. Similar idea is used in [1]. Most of these works do not consider potentially available cameras to be involved. By potentially available camera, we mean a camera that cannot see an object at its current setting, but will see the object after being panned or tilted.

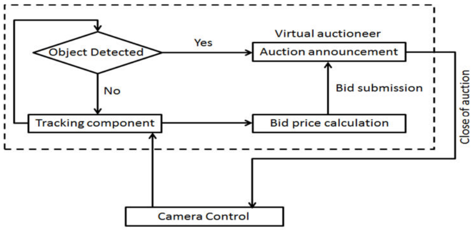


Figure 1. Overview of the proposed approach.

In this paper, we want to solve a comprehensive problem, which involves both camera active control and multi-camera multi-person tracking. We propose an approach to achieve this by using an auction-based technique. He and Ioerger develop an Auction-based mechanism for computational grids [9]. Chen et al. [10] achieve single target tracking in wireless networks by deploying auction-based coalition. However, there is very little work [1] that has ever used the auction-based technique in a camera network to select and control cameras to follow up multiple objects. The advantage of using auction-based approach lies in the fact that we can consider multiple possible settings to predict a panning or tilting operation of a camera and reach the Poreto Optimal solution.

The proposed approach in this paper aims at combining multi-tracker tracking in a camera network and the auction-based techniques together so as to solve the camera active control problem with a new perspective. We assume that homographies among the cameras with overlapping views are available as prior

978-4-9906441-1-6 ©2012 IAPR

knowledge. The uncertainties that may exist when mapping positions among different cameras are also taken into account.

The proposed work is most similar to Qureshi and different inherently:

1. In [1], a leader node is included in a single node group and it holds an auction to recruit other nodes into the group. When there are multiple groups recruiting for the same node(s), this is treated as a Constraint Satisfaction Problem. Whereas in our approach, camera groups are formed by selecting the cameras with top N bidding prices.

2. In terms of PTZ control, the approach in [1] only considers the zoom-in operation to have a close-up view when necessary. No PT controls are considered in [1]. However, in our approach, PT controls are considered by using pre-calculated homographies between camera pairs. Bidding prices are modeled as vectors to take potentially available cameras into account.

3. Synthetic data are used for the experiments in [1]. In this paper, we use real-world data to examine the proposed approach.

In the rest of this paper, we will introduce the proposed approach for active camera control in Section 2, the metrics for calculating the bid prices are described in Section 3. Experimental results are shown in Section 4 and Section 5 concludes the paper.

2. Auction-based Camera Active Control

The assumptions in this proposed framework are:

1. We assume the objects to be tracked are human beings walking on a flat planar. The feet of these persons are visible so that the position of a person in one camera can be mapped to another camera with overlapping field of views (FOVs) by using homographies.

2. Homographies are pre-calculated and the cameras' heights are known, so that we know the coordinate conversions between different camera pairs with overlapping settings.

3. The camera's focal length is set to a fixed number such that the angle of view (the largest angle that a camera can cover without any active control) is 51.2°. Each camera has 4 overlapping pre-defined pan settings to seamlessly cover 180 degrees. We do not use one camera to cover 360 degrees because when the camera is panned over 180 degrees, it takes more time to get the image focused and because of time delay a person may be lost. There are three tilt settings, up 5°, down 5° (or -5°) and no tilt (0°). So, there are 12 settings for each cameras. We will call these 12

Terzopoulos in [1]. Although both [1] and our work use an auction process to form groups of cameras, the principles of these two approaches are

settings for a Camera C_j as $\boldsymbol{l} = \{l_j^1, l_j^2, \ldots, l_j^{12}\}$ where l_j^1 is the current location of Camera C_j.

4. There is no communication error.

Based on the above assumptions, we propose an auction protocol to select cameras automatically and dynamically to follow the objects in the network.

An auction is the process of selling an item from the auctioneer to many potential buyers, i.e., bidders. Typically, in the auction, the potential buyers first offer their prices (the price offer is also called a bid) [11, 12]. Then, the auctioneer collects the bid prices information, and decides who wins the item and how much the winner has to pay. Analogously, in the camera selection with active control, we model the cameras as bidders and the control center as the auctioneer. Quality of views (QOV) is used as the bidding prices.

An overview shown in Figure 1. Whenever an objected is detected, there is a virtual auctioneer announcing an auction for it. All available cameras calculate their bid prices for this object locally and submit it to the virtual auctioneer and a final decision will be made to close the auction. Following this model, an auction protocol inspired by [13] is described as follows:

1. ***Task announcement.*** A virtual auctioneer (program running on a central server) holds an auction for each object to be tracked. An auction message is broadcast to the whole network. The message includes information such as the location of an object and object ID. All the cameras that can "see" this object will participate in the auction. The object's location is initialized as the centroid location in the camera where the tracker has the highest confidence.

2. ***Bid price calculation***. The overall bid price from camera C_j is matrix B_j which is $N_P \times 12$, where N_P is the number of people in the network.

$$B_j = \left[\boldsymbol{b_{1j}}, \boldsymbol{b_{2j}} \cdots \boldsymbol{b_{ij}}, \cdots \boldsymbol{b_{N_P j}}\right]^T$$

where B_{ij} is the bid price from camera C_j for person P_i, and decided by a 1×12 bid vector, $\boldsymbol{b_{ij}} = \{b_{ij}^1, b_{ij}^2, \ldots, b_{ij}^k, \ldots, b_{ij}^{12}\}, k \in [1,12]$. b_{ij}^k stands for the *intermediate bid* that the camera can get by panning or tilting to the setting l_j^k (as defined in Assumption 3). If it cannot "see" an object at l_j^k, then b_{ij}^k is 0. Otherwise, b_{ij}^k is decided by the pre-defined metrics, which will be discussed in the next subsection. The order of elements in $\boldsymbol{b_{ij}}$ implies the "willingness" of the camera to follow an object or not. We prefer to use a camera without any panning or tilting to avoid unnecessary blurred images.

However, when there is no desirable camera image for an object, the bidding vector $\boldsymbol{b_{ij}}$ can show the potential capability of camera C_j to track this object at different settings. This vector representation helps to take into account the willingness of a camera, which, therefore, avoids the drawbacks of greedy algorithms.

3. ***Bid submission***. After evaluating the price for each object, all the related cameras send their bid prices for the object(s). The prices must be honest and can truly imply their willingness to follow an object.

4. ***Close of auction***. Unlike in the traditional auction, where the auctioneer will sell the good to the buyer who provides the highest bid price, the auction in our system is held for N_P objects simultaneously. So, we have to choose the solution which is globally optimal. To achieve this within a short time, we deploy the bargaining mechanism introduced in [3] so that we can make a camera selection in real-time.

3. Evaluating Bids and Integrating Multiple Trackers

For the metrics used for evaluating the bids, we mainly consider the size and the position of the person in the camera image. The score for Person P_i in camera C_j by tracker T_k is described as follows:

1. *Size of the tracked person*. Assume that γ is the threshold for the best observation, i.e. when $r = \gamma$ this criterion reaches its peak value, where $r = \frac{\#\, of\ pixels\ inside\ the\ bounding\ box}{\#\, of\ pixels\ in\ the\ image\ plane}$.

$$\mathrm{M}_{ij1}^{k} = \begin{cases} \frac{1}{\gamma} r, & when\ r < \gamma \\ \frac{1-r}{1-\lambda}, & when\ r \geq \gamma \end{cases} \tag{1}$$

2. *Position of the person*. It is measured by the Euclidean distance that a person is away from the center of the image

$$\mathrm{M}_{ij2}^{k} = \frac{\sqrt{(x-x_c)^2+(y-y_c)^2}}{\frac{1}{2}\sqrt{{x_c}^2+{y_c}^2}} \tag{2}$$

where (x, y) is the current position of the person and (x_c, y_c) is the center of the camera image plane.

Since there is no single tracker that can act perfectly in all scenarios, in this paper, we integrate the results from multiple trackers to have a more reliable tracking performance. The contributions of different trackers are prorated according to their tracking confidence values as shown in Equation (3). Each intermediate bid b_{ij}^{l} is decided by the above metrics and is calculated

$$b_{ij}^{l} = \sum_{k=1}^{2} \frac{conf_k}{\sum_{k=1}^{2} conf_k} \sum_{m=1}^{2} w_m M_{ijm} \tag{3}$$

where w_m is the weight for different metrics. $conf_k$ is the confidence value returned by the k^{th} tracker. In the case of using the boosting trackers, this is the confidence value of the boosted classifier. The calculation of these M_{ijm} is described in the experimental part.

The bid price for P_i from C_j, B_{ij} is computed as

$$B_{ij} = \left(\alpha_1 (b_{ij}^{1})^{\lambda} + \alpha_2 (b_{ij}^{2})^{\lambda} + \cdots + \alpha_{12} (b_{ij}^{12})^{\lambda}\right)^{\frac{1}{\lambda}} \tag{4}$$

where $\alpha_1 + \alpha_2 + \cdots + \alpha_{12} = 1$, $\lambda \in (-\infty, +\infty)$. The parameter λ in equation (4) measures the degree of easiness in substitution among different dimensions in the intermediate bid vector $\boldsymbol{b_{ij}}$ and α_k measures the camera's relative preference on b_{ij}^{p} to b_{ij}^{q} $(p \neq k \backslash q)$.

4. Experiments

We use two trackers are used in this paper: the online boosting tracker [14] and the semi-supervised online boosting tracker [15]. The reason why we choose these two trackers are: 1) The implementation for these two trackers are publicly available. 2) Although the author claimed that the semi-supervised boosting tracker should have a better performance than the online-boosting tracker, we run these two trackers on different data sets, and find that they can compensate each other in most cases. 3) These two trackers both belong to the boosting trackers, such that the confidence valued returned by the tracker are comparable with no post processing. Hence, we do not need to do any further analysis to normalize the confidence values.

Data association among different cameras is done by the pre-calculated homographies. The trackers are rectified using the information from the tracker with the highest confidence periodically [16].

Since we need to do camera active controls, there is no way to use any public datasets. For online camera controls, it is hard to define the ground truth, since the experiments have to continue based on the operations happened in previous frames. So if the selected tracker in the selected camera has 70%-150% overlap with the object, and there is no other camera has a higher QOV based on the current view, then it is considered correct. Otherwise, there is an error.

We use 3 to 4 Axis 215 PTZ cameras to collect our own data and have 3 experiment trials, which are shown in Table 1. For *real-time camera control*, we use multithreads to implement the system. One tracker for one person in one camera is implemented as one thread. In our most complicated scenario, where there are 4 persons and 4 cameras with 2 trackers, there are a total of 32 threads. The parameters in the experiments are set empirically. The threshold for the size of the person is $\gamma = \frac{1}{15}$. The weights for different metrics are selected as $w_1 = 0.6$ and $w_2 = 0.4$. To show the effectiveness of using multiple trackers for doing the camera selection and active control, we pre-define the

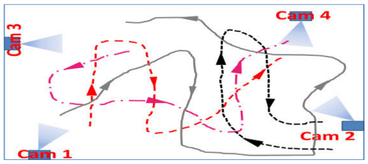
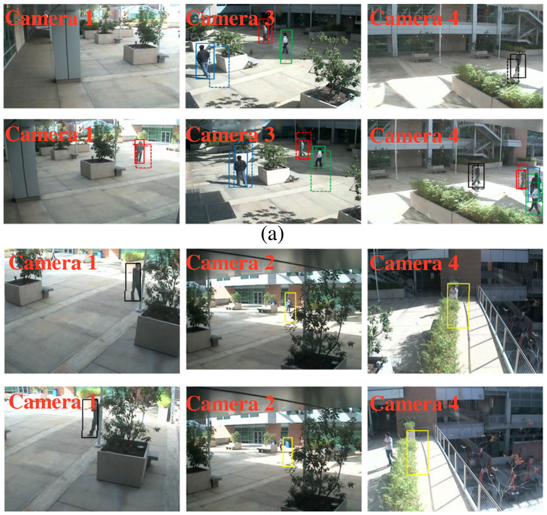


Figure 2. Pre-defined trajectories in case 3.



(a)

(b)

Figure 3. Example frames. Tracking results for the same person from different trackers are demonstrated in solid line and dashed line respectively. (a) Results with 2 trackers. (b) Results with a single tracker.

TABLE 1 EXPERIMENTAL CASES

Cases	# of Camera	# of Persons	Error rate	
			1 tracker	2 trackers
Case 1	3	2	5.8%	3.2%
Case 2	3	4	9.7%	4.2%
Case 3	4	4	10.2%	4.8%

trajectories of the persons. There are grids on the experimental ground, making it easy to repeat the trajectories, as shown in Figure 2. Also it helps in making comparisons between using multiple trackers and a single tracker. On a quad core 3GHz computer, the processing speed is 15-22 fps.

We show example frames only for the most complex scenario, case 3. In Figure 3 (a), we show the results by using boosting tracker and semi-supervised online boosting tracker simultaneously while in Figure 3 (b) the results by using a single tracker. We can see that when using a single tracker, the system can make correct camera controls under simple conditions where there is no tracking ambiguity, for example in Figure 3 (b) camera 1. However, when the tracker is lost (in camera 2 and camera 4), the system makes a wrong decision based on the lost tracker. While similarly in Figure 3(a) camera 3 and camera 4, where one of the trackers is lost and the other one works well, the system still makes the correct selection because the working tracker has a higher confidence and thus has a higher weight in this case. The overall results in all the three cases are shown in Table 1.

5. Conclusions

In this paper, we model the camera selection problem with PT control as an auction process. By submitting a bidding vector, the camera can express its “willingness” to be panned or tilted to track a particular target. Multiple trackers are deployed simultaneously to get a more accurate tracking result. The score of different trackers are weighted by the trackers’ confidence accordingly. The experimental results show that the proposed approach can be used in real-time surveillance systems. The monitored area can be enlarged with camera Pan / Tilt controls with keeping the number of cameras unchanged.

Acknowledgement

This work was partially supported by NSF grants 0551741, 0622176, 0905671 and ONR grants (DoD Instrumentation and the Aware Building).

References

[1] F.Z. Qureshi and D. Terzopoulos. Distributed coalition formation in visual sensor networks: a virtual vision approach. *ICDCS 2007*.

[2] L. Tessens *et al*. Principal view determination for camera selection in distributed smart camera networks. *ICDSC* 2008.

[3] Y. Li and B. Bhanu. Utility-based camera assignment in a video network: A game theoretic framework. In *IEEE Sensors Journal*, Issue 3, 2011.

[4] K. Chen *et al.* An adaptive learning method for target tracking across multiple cameras. *CVPR* 2008.

[5] E. Monari and K. Kroschel. A knowledge-based camera selection approach for object tracking in large sensor networks. *ICDSC* 2009.

[6] Shen *et al.* A Multi-Camera Surveillance System that Estimates Quality-of-View Measurement. *ICIP* 2007.

[7] T. Vidal-Challeja *et al.* Active control for single camera SLAM. *ICRA* 2006.

[8] H.-C. Choi et al. PTZ camera assisted face acquisition, tracking & recognition. *IJCB* 2010.

[9] L. He and T.R. Ioerger. Task-oriented computational economic-based distributed resource allocation mechanisms for computational grids. *ICAI*, 2004.

[10] Chen et al. Auction-based dynamic coalition for single target tracking in wireless sensor networks. *WCICA* 2006.

[11] E. Wolfstetter. Auctions: an Introduction. *Journal of economic surveys*, vol. 10(4), pages 367-420, 1996.

[12] R. P. McAfee and J. McMillan. Auctions and bidding. *Journal of economic literature*, Vol. 25, No. 2, 1987.

[13] R. G. Smith. The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Transctions on Computers* C-29(12), 1980.

[14] H. Grabner, and H. Bischof. Online-boosting and vision. *CVPR* 2006.

[15] H. Grabner, C. Leistner, and H. Bischof. Semi-supervised on-line boosting for robust tracking. *ECCV* 2008.

[16] Y. Li and B. Bhanu. Fusion of multiple trackers in video networks. *ICDSC* 2011.